Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives

---

Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües

Edited by
Gloria Corpas Pastor

Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives

Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües

2016. Editions Tradulex, Geneva

ISBN: 978-2-9700736-5-9

ORGANISING COMMITTEE/COMITÉ ORGANIZADOR

Chair/Presidencia

- Gloria Corpas Pastor

Organisers/Organizadores

- Rosario Bautista Zambrana
- Cristina Castillo Rodríguez
- Hernani Costa
- Isabel Durán Muñoz
- Jorge Leiva Rojo
- Gema Lobillo Mora
- Pablo Pérez Pérez
- Míriam Seghiri Domínguez
- M.ª Cristina Toledo Báez
- Míriam Urbano Mendaña
- Anna Zaretskaya

Secretary/Secretaría

- Míriam Buendía Castro
- Rut Gutiérrez Florido

COMITÉ CIENTÍFICO/ PROGRAMME COMMITTEE

- Margarita Alonso Ramos (Universidad da Coruña, Spain)
- Ignacio Bosque (Universidad Complutense de Madrid, Spain)
- Jenny Brumme (Universitat Pompeu Fabra, Spain)
- František Čermák (Univerzita Karlova v Praze, Czech Republic)
- Jean-Pierre Colson (Université Catholique de Louvain, Belgium)
- Dmitrij Dobrovol'skij (Russian Academy of Sciences, Russia)
- Peter Ďurčo (Univerzita sv. Cyrila a Metoda v Trnave, Slovakia)
- Xesús Ferro Ruibal (Centro Ramón Piñeiro para a Investigación en Humanidades, Spain)
- Sabine Fiedler (Universität Leipzig, Germany)
- Natalia Filatkina (Universität Trier, Germany)
- Thierry Fontenelle (Translation Centre for the bodies of the European Union (CdT), Belgium)
- Miguel Á. García Peinado (Universidad de Córdoba, Spain)
- Jose Enrique Gargallo Gil (Universitat de Barcelona, Spain)
- Maribel González Rey (Universidade de Santiago de Compostela, Spain)
- Annelies Häcki Buhofer (Universität Basel, Switzerland)
- Patrick Hanks (University of Wolverhampton, United Kingdom)
- Ulrich Heid (Universität Hildesheim, Germany)
- Adam Kilgarriff (University of Brighton, United Kingdom)
- Ramesh Krishnamurthy (Aston University, United Kingdom)
- Elvira Manero Richard (Universidad de Murcia, Spain)
- Josep Marco (Universitat Jaume I, Spain)
- Manuel Martí Sánchez (Universidad de Alcalá, Spain)
- Carmen Mellado Blanco (Universidade de Santiago de Compostela, Spain)
- Salah Mejri (Université Paris 13, France)
- Ruslan Mitkov (University of Wolverhampton, United Kingdom)
- Pedro Mogorrón (Universitat d'Alacant, Spain)

- Johanna Monti (Università degli Studi di Sassari, Italy)
- Esteban T. Montoro del Arco (Universidad de Granada, Spain)
- Rosamund Moon (University of Birmingham, United Kingdom)
- Carmen Navarro (Universita' degli studi di Verona, Italy)
- Michael Oakes (University of Wolverhampton, United Kingdom)
- Inés Olza (Universidad de Navarra, Spain)
- Antonio Pamies (Universidad de Granada, Spain)
- Inmaculada Penadés Martínez (Universidad de Alcalá, Spain)
- Rosa Piñel (Universidad Complutense de Madrid, Spain)
- Carlos Ramisch (Aix Marseille Université, France)
- Leonor Ruiz Gurillo (Universitat d'Alacant, Spain)
- Agata Savary (Université Francois Rabelais, France)
- Elmar Schafroth (Universität Düsseldorf, Germany)
- Violeta Seretan (Université de Genève, Switzerland)
- Julia Sevilla Muñoz (Universidad Complutense de Madrid, Spain)
- Inès Sfar (Université Paris 13, France)
- Kathrin Steyer (Institut für Deutsche Sprache, Germany)
- Joanna Szerszunowicz (Uniwersytet w Białymstoku, Poland)
- Aina Torrent (Fachhochschule Köln, Germany)
- Eric Wehrli (Université de Genève, Switzerland)
- Gerd Wotjak (Universität Leipzig, Germany)
- Stefanie Wulff (University of Florida, United States)
- Pablo Zamora (Universidad de Murcia, Spain)

INVITED SPEAKERS/CONFERENCIANTES PLENARIOS

Jean-Pierre Colson

Professor of Translation Studies and Linguistics at the Université Catholique de Louvain, and President of the Louvain School of Translation and Interpreting (LSTI).

"The contribution of corpus-based phraseology to translation studies: from experiments to theory"

29 June 2015/29 de junio de 2015

The notion of phraseology is now used across a wide range of linguistic disciplines: Phraseology (proper), Corpus Linguistics, Discourse Analysis, Pragmatics, Cognitive Linguistics, Computational Linguistics. It is, however, conspicuously absent from most studies in the area of Translation Studies (e.g. Delisle 2003, Baker & Saldanha 2011). The paradox is that many practical difficulties encountered by translators and interpreters are directly related to phraseology in the broad sense (Colson 2008, 2013), and this can most clearly be seen in the failure of SMT-models (statistical machine translation) to deal efficiently with the translation of set phrases (used here as a generic term for all categories of phraseological constructions, from collocations to proverbs).

Although corpus-based and computational phraseology still need to be clearly delineated from other concurrent disciplines, a possible way of narrowing the gap between phraseology and translation studies is proposed here: the recourse to experiments involving on the one hand set phrases and, on the other, evidence from parallel translation corpora or SMT-machines such as Google Translate. We will argue that both phraseology and translation studies have much to gain from this cross fertilisation, because both disciplines are regularly criticised for their lack of coherent terminological description and for the insufficient number of reproducible experiments they involve. The aim of this paper is not to draw up an exhaustive list of the possible experiments showing the interweaving of phraseology and translation studies, but to propose directions for future research involving a number of key issues that are posed by phraseology and are illustrated by translation practice.

A first series of experiments relating to this subject matter concerns the problems posed by phraseology to human translation. Decoding phraseology in the source text is far from easy for translators and interpreters, all the more so as they are usually not native speakers

of the source language. Also, finding a natural formulation in the target language and avoiding translationese requires an excellent mastery of the phraseology of the target language. I will argue that experiments with translation corpora may precisely shed some light on some crucial notions of phraseology and of translation studies. Experiments have shown that translation errors due to phraseology are legion in many translation corpora, even in the official translations of the European Union. A contribution of corpus-based phraseology would therefore consist in making human translators aware of the pitfalls of phraseology in the source text. Even experienced professionals sometimes fail to detect the fixed or semi-fixed character of a source text construction. Experiments along these lines should therefore also include the creation of large, multilingual phraseological databases, which brings us back to two serious shortcomings of computational phraseology:

1. There is no universally accepted algorithm for the automatic extraction of phraseology, especially not for ngrams larger than bigrams.
2. There is no consensus as to the proportion of set phrases in relation with the rest of the vocabulary: according to Jackendoff (1995), there are about as many fixed expressions as there are single words in the dictionary, but others (such as Mel'čuk 1995) hold the view that fixed expressions far outnumber single words.

I will argue in that respect that algorithms derived from text mining and information retrieval techniques (Baeza-Yates, R. & B. Ribeiro-Neto 1999) can be efficient and (computationally) cost-effective in order to build up unfiltered collections of recurrent fixed or semi-fixed phrases, from which translators could gain information about the number of set phrases in the source text. Such an algorithm has been proposed in Colson (2014), and a provisional database of about 700,000 English set phrases (tokens) has been assembled, which seems to confirm that Jackendoff's view about the total number of fixed expressions was not correct.

A second series of experiments that would turn out to be profitable to a better theoretical understanding of both phraseology and translation studies, has to do with the specific problems posed by phraseology to automatic translation. Phraseology has only recently been identified as one of the main sources of errors in automatic translation systems, including the most recent SMT-systems (Monti, Mitkov, Corpas Pastor & Seretan 2013). I will however point out that the theoretical underpinnings of phraseology are at stake in order to provide a coherent explanation for the serious shortcomings in the automatic translation of sentences containing phraseology. The crux of the matter seems to be the complex interplay between association and frequency in fixed expressions. Recent

evidence shows that, contrary to what is assumed by most statistical scores, there should be no relationship between the statistical association of the grams constituting a set phrase, and its frequency in a huge corpus. The countless examples of wrong translations of phraseologically rich sentences by Google Translate, for instance, all point to the fundamentally wrong way in which ngrams were traced down, namely by giving the highest priority to frequency.

Further experimentation should also shed some light on the overall statistical distribution of set phrases in large corpora. The well-know zipfian distribution of words in a corpus poses theoretical problems as far as phraseology is concerned. Corpus-based studies (Baroni 2008) indicate that the distribution of ngrams themselves may display a Zipf-Mandelbrot curve. This is an important theoretical challenge to the theory of phraseology and also to semantics, having therefore consequences on the way meaning may be expressed in different languages and be adequately translated from one language into another. I will point out that a general theory of phraseology, as outlined by Mejri (2006), may offer a new insight into the statistical underpinnings of both morpheme associations (in words) and of word association (in set phrases).

References

BAEZA-YATES, R. & B. RIBEIRO-NETO (1999). *Modern Information Retrieval*. New York: ACM Press, Addison Wesley.

BAKER, M. & G. SALDANHA (EDS.) (2011). *Routledge Encyclopedia of Translation Studies*. New York: Routledge.

BARONI, M. (2008). Distributions in text. In: A. Lüdeling & M. Kytö, (eds.), *Corpus linguistics. An international handbook*. Berlin, New York: Walter de Gruyter, p. 803-821.

BARONI, M., BERNARDINI, S., FERRARESI, A. & E. ZANCHETTA. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation*, 43, p. 209-226.

COLSON, J.-P. (2008). Cross-linguistic phraseological studies: An overview. In: Granger, S. & F. Meunier (eds.), *Phraseology. An interdisciplinary perspective*. John Benjamins, Amsterdam / Philadelphia, p. 191-206.

COLSON, J.-P. (2010a). The Contribution of Web-based Corpus Linguistics to a Global Theory of Phraseology. In: Ptashnyk, S., Hallsteindóttir, E. & N. Bubenhofer (eds.), *Corpora, Web and Databases. Computer-Based Methods in Modern Phraselogy and Lexicography*. Hohengehren, Schneider Verlag, p. 23-35.

COLSON, J.-P. (2010b). Automatic extraction of collocations: a new Web-based method. In: S. Bolasco, S., Chiari, I. & L. Giuliano, *Proceedings of JADT 2010,Statistical Analysis of Textual Data*, Sapienza University of Rome, 9-11 June 2010. Milan, LED Edizioni, p. 397-408.

COLSON, J.-P. (2013). Pratique traduisante et idiomaticité : l'importance des structures semi-figées. In: Mogorrón Huerta, P., Gallego Hernández, D., Masseau, P. & Tolosa Igualada, M. (eds.), *Fraseología, Opacidad y Traduccíon. Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation* (Herausgegeben von Gerd Wotjak).

Frankfurt am Main, Peter Lang, p. 207-218.

COLSON, J.-P. (2014). Set phrases around *globalization* : an experiment in corpus-based computational phraseology. Paper presented at *CILC 2014, 6th International Conference on Corpus Linguistics*. University of Las Palmas de Gran Canaria, 22-24 May 2014.

CORPAS PASTOR, G. (2013). Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In Olza, I. & R. Elvira Manero (eds.) *Fraseopragmática*. Berlin: Frank & Timme, p. 335-373.

DELISLE, J. (2003). La traduction raisonnée. Ottawa: Presses de l'Université d'Ottawa.

JACKENDOFF, R. (1995). The boundaries of the lexicon. In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schroeder (eds.), *Idioms: Structural and psychological perspectives*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, p. 133-165.

MEJRI, S. (2006). Polylexicalité, monolexicalité et double articulation. *Cahiers de Lexicologie*, 2 :209-221.

MEL'CUK, I. 1995. Phrasemes in language and phraseology in linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schroeder (eds.), *Idioms: Structural and psychological perspectives*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, p. 167-232.

MONTI, J., MITKOV, R., CORPAS PASTOR, G. & V. SERETAN (EDS) (2013). Workshop Proceedings: *Multi-word units in machine translation and translation technologies*, Nice 14th Machine Translation Summit.

Ulrich Heid

Professor of computational linguistics and language technology at University of Hildesheim

"Extracting linguistic knowledge about collocations from corpora"

29 June 2015/29 de junio de 2015

We start from the assumption that (lexical) collocations and verbal idioms are a type of multiword expressions which deserve a detailed linguistic and lexicographic description (Gouws/Heid 2006: treatment units of their own in dictionaries). If this is so, then there is a need for corpus-based tools which allow us to find out about the contextual (= syntagmatic) and paradigmatic properties of collocations. We intend to show how much of these data can be identified with acceptable quality in large enough corpora.

Syntagmatic properties have to do with the distributional behavior of collocations: preferences in number (have high hopes, pl.), determination (article use) or modifiability are well-known examples; some collocations and many verbal idioms have their own syntactic valency constructions (cf. be in a position [+to+INF]) or they co-occur preferentially with certain lexical items, e.g. as modifiers (cf. DE Kritik üben ("criticize") which prefers adjectives that typically collocate with Kritik: harsche, scharfe Kritik üben ("criticize severely"), cf. Häcki-Buhofer et al. 2014).

Examples of paradigmatic properties include the exchangeability of lexical elements of the collocation against synonyms, or the availability of nominalizations (submit a proposal – submission of a proposal) or of compounds in Germanic languages (DE Antrag einreichen ("submit a proposal") – Einreichung eines Antrags – Antragseinreichung). Another example are pragmatic marks and preferences with respect to domain-specific languages.

We will give examples of such data from German, English, French and Italian and we will assess to what extent such linguistic knowledge may be needed in translation and in (mother tongue or foreign language) text production. Thereafter, we intend to show which types of data of the above kind can be extracted with acceptable quality from corpus texts, and with which language processing techniques; we claim that state of the art dependency parsing provides a fair amount of such data thus facilitating the description work of terminologists and lexicographers.

References

GOUWS, RUFUS H. AND HEID, ULRICH. 2006. "A model for a multifunctional dictionary of collocations", in Corino, Elisa et al. (Eds.): Proceedings of the XIIth EURALEX International Congress, (Alessandria: Edizioni dell'Orso), 979 – 988.

HÄCKI-BUHOFER, ANNELIES ET AL. 2014. Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag, (Tübingen: Francke).

Patrick Hanks

Professor in Lexicography at the Research Institute of Information and Language Processing in the University of Wolverhampton.

"Meaning and Phraseology: a Corpus-Driven Approach"

30 June 2015/30 de junio de 2015

Interesting aspects of the meaning are revealed by corpus-driven lexical analysis. The typical function of nouns is to create referring expressions—terms that either refer to objects in the world or denote abstract concepts. The typical function of verbs, on the other hand, is to create propositions, in which noun phrases play roles that are mediated by a verb. According to the Theory of Norms and Exploitations (Hanks, 1994, 2004, 2013), a verb has only meaning potential (not meaning as such) until it is put in context. There is no 'semantic invariable' that is common to all normal uses of a verb. Consider the verb blow. 'A gale was blowing', 'They blew up the bridge', 'He blew his nose', and 'She blew the whistle on government malpractice' have little in common semantically, but all four sentences represent realizations of conventional lexico-syntactic patterns of English. The meanings lie in collocation and phraseology, not in the words themselves.

Different questions must be asked about nouns and verbs, and different apparatuses are required for corpus analysis of these two categories. When shower is used as a noun, we can ask how many different kinds of shower there are—rain showers, snow showers, spring showers, etc., as opposed to bathroom showers and power-driven showers. What distinctive properties or common features does each category have? On the other hand, if shower is used as a verb, relevant question are prompted by the collocates and syntagmatics: for example, 'Is it normal to say in English, "It showered yesterday"?' Patterns with prepositions such as with and on prompt questions such as 'Who showers what on whom?' 'Who showers whom with what?' 'What is the relationship between such patterns?' In this way, we can start to compile an inventory of patterns of word use that seem to be already available to the unconscious minds of users of a language.

Ruslan Mitkov

Director of the Research Institute of Information and Language Processing (RIILP) and Professor of Computational Linguistics and Language Processing at the University of Wolverhampton

"Knowledge- and resource-poor identification and translation of multiword expressions across language pairs"

30 June 2015/30 de junio de 2015

The correct identification, interpretation and translation of multiword expressions in both general and specialised languages, is vital for the successful operation of most Natural Language Processing applications and computer-aided tools supporting various users including phraseologists, translators, interpreters, terminologists, language learners and teachers. As parallel corpora are scarce and in order to benefit from the wider availability of comparable corpora (which can be also compiled with a specific task in mind), there is a pressing need to develop approaches which can extract and translate multiword terms from comparable corpora. In this presentation the speaker will propose a novel methodology based on computing semantic similarity to extract and translate multiword units from comparable corpora termed as 'knowledge- and resource-poor' as it is not dependent on the use of any dictionaries or linguistic knowledge. While this particular study focuses on English and Spanish, the methodology is not restricted to any particular pair of languages.

TABLE OF CONTENTS/ÍNDICE DE CONTENIDOS

CORPUS-BASED PHRASEOLOGY
FRASEOLOGÍA BASADA EN CORPUS

Xabier Altzibar Aretxabaleta, Xabier Bilbao López **23**
Locuciones en euskera: necesidad y pautas para su recopilación y ordenación a partir de los corpus textuales existentes

Gloria Corpas Pastor **31**
Fraseología ¿sexista?: lo que el corpus esconde

Laura Giacomini **38**
Systematic vs. non-systematic collocational patterns in LSP: paradigmatic variation in the technical domain

María Araceli Losey León **48**
Computerisation of phrase-to-phrase matching from a Standard Marine Communication Phrases corpus: A preliminary empirical study

Agnès Tutin, Emmanuelle Esperança-Rodier, Manolo Iborra, Justine Reverdy **60**
Annotation of multiword expressions in French

Petr Zemánek, Jiří Milička **68**
Restricted Collocability and its Use in Arabic Corpus Linguistics

Zuriñe Sanz Villar **80**
German-into-Basque/Spanish translation analysis of binomials in a parallel and multilingual corpus

NLP AND/OR CORPUS-BASED IDENTIFICATION AND CLASSIFICATION OF PHRASEOLOGICAL UNITS
IDENTIFICACIÓN Y CLASIFICACIÓN DE UNIDADES FRASEOLÓGICAS BASADA EN CORPUS O MEDIANTE TÉCNICAS DE PLN

Mariangela Albano **94**
Traduire des idiomes françaises en langue étrangère (italien, allemand, espagnol): traitement cognitif et stratégies d'interprétation

Jorge Baptista, Graça Fernandes, Rui Talhadas, Francisco Dias, Nuno Mamede **102**
Implementing european portuguese verbal idioms in a natural language processing system

Sara Castagnoli, Gianluca E. Lebani, Alessandro Lenci, Francesca Masini, Malvina Nissim, Lucia C. Passaro **116**
POS-patterns or Syntax? Comparing methods for extracting Word Combinations

Svitlana Chornobay, Jorge Baptista **129**
Semantic Structuring of Verbal Idioms from the Conceptual Domain {Death} A Russian-Portuguese-English Contrastive Approach

Tetyana Fukova, Svitlana Chornobay, Jorge Baptista **139**
Lexicon-Grammar of Russian Verbal Idioms

Vincenzo Lambertini **154**
Paremiología basada en corpus WaCky: enfoque (intra- e inter) lingüístico y conceptual

Marta Morer Murcia **163**
Contrastive Analysis of Phraseological Units with specific animal constituents in English, Spanish and German

Suzanne Mpouli, Jean Gabriel Ganascia **179**
"Pale as death" or "pâle comme la mort": Frozen similes used as literary clichés

Lucia C. Passaro, Alessandro Lenci **188**
Extracting terms with EXTra

Diana Peppoloni **197**
Statistical automatic extraction of V-N italian collocations from an academic spoken corpus

Sónia Reis, Jorge Baptista **208**
Portuguese Proverbs: Types and Variants

Dorota Sikora **218**
Identification d'unités phraséologiques et équivalence sémantique dans la traduction

Shiva Taslimipoor **232**
Cross-lingual extraction of multiword expressions

Amalia Todirascu, Mirabela Navlea **238**
Integrating Verb+Noun Collocations into a French - Romanian Lexical Alignment System for Law Domain

COMPUTER-AIDED AND/OR CORPUS-BASED ANALYSIS OF PHRASEOLOGICAL UNITS
ANÁLISIS DE UNIDADES FRASEOLÓGICAS BASADO EN CORPUS O ASISTIDO POR ORDENADOR

František Čermák, Marie Kopřivová **250**
Idioms in Spoken Corpus. A Sample of Czech Data

Abdellatif Chekir **262**
Phraséologie et traduction : perspective contrastive à base d'un corpus bilingue français-arabe tunisien

Itsuko Fujimura, Shigenobu Aoki **271**
A new score to characterise collocations: Log-R in comparison to mutual information

Daniel Gallego Hernández **283**
Variation diatopique en phraséologie spécialisée dans le domain financier. Étude comparative basée sur corpus

Henry Hernández Bayter **293**
Dime cómo hablas y te diré quién eres: Las unidades discursivas con carácter fraseológico en el discurso político

Zita Hollós **302**
Korpusbasierte intra- und interlinguale Kollokationen

Herbert J. Holzinger **316**
Mit Bedacht: Korpuslinguistische Untersuchungen zu Strukturen [Präposition + Substantiv] mit adverbialer Funktion

Anastasia Kovaleva **330**
Fraseologuismos de color en español y ruso: estudio de fraseologismos sin análogos en la otra lengua

Belén López Meirama **340**
A tiros y a balazos: análisis construccional

John Anthony McKenny **349**
An exploration of the phraseology of a large corpus of Academic English: a new more teachable taxonomy of multiword expressions

Nikoleta Olexová **360**
Verbale Kollokationen: Jeder kennt seinen Platz. Jeder weiß, wo sein Platz ist

Gabriela Orsolya **370**
Fühlen oder empfinden? Ein Vergleich der Kookkurrenzprofile der partiellen Synonyme

Sunock Shin, Pierre-André Buvet **383**
Contrastive analysis of verb-noun collocations of 'utterance' in French and Korean

Wojciech Sosnowski **396**
The parallel Polish-Bulgarian-Russian corpus: problems and solutions

María Rosario Bautista Zambrana **407**
Aprender fraseología mediante corpus: un caso aplicado a la enseñanza del alemán

Javier Martín Salcedo **417**
¿Dar o echar un piropo? Me quedo loco, nunca acierto. Colocaciones verbales en español y portugués

María Eugenia Olímpio de Oliverira Silva, Inmaculada Penadés Martínez 425
Linguee como herramienta de enseñanza-aprendizaje de las unidades fraseológicas

PHRASEOLOGY IN E-LEXICOGRAPHY AND E-TERMINOLOGY
LA INFORMACIÓN FRASEOLÓGICA EN LA LEXICOGRAFÍA Y LA TERMINOLOGÍA ELECTRÓNICAS

Tsiuri Akhvlediani, George Kuparadze 436
Phraseology - Cultural Code of Ethnicity (On the material of French, English, and Georgian languages)

Miriam Buendía Castro, Pamela Faber 449
Phraseological correspondence in English and Spanish specialized texts

Heloisa Fonseca 457
La web como corpus y base de investigación científica

Maciej Paweł Jaskot 469
Do we need equivalence-based e-tools?

Jorge Leiva Rojo 477
Fraseografía y lingüística de corpus: sobre el tratamiento de locuciones verbales en la nueva edición del Diccionario de la lengua española

Adriane Orenha-Ottaiano 486
The compilation of an online corpus-based bilingual collocations dictionary

Marie-Sophie Pausé 494
Pour un continuum des phrasèmes non-compositionnels

Irena Srdanovic 506
Pragmatic information and unpredictability in learner's Dictionaries

Simon Clematide, Johannes Graën, Martin Volk 513
Multilingwis - A Multilingual Search Tool for Multi-word Units in Multiparallel Corpora

Kristina Kocijan, Sara Librenjak 523
Comparative Structures in Croatian: MWU Approach

Christine Konecny, Andrea Abel, Erica Autelli, Lorenzo Zanasi 533
Identification and Classification of Phrasemes in an L2 Learner Corpus of Italian

Joanna Szerszunowicz **543**
Corpora, the World Wide Web and questionnaires as sources of information on recent phraseological borrowings: The case study of the Polish unit 'wyglądać jak milion dolarów'

Arsenio Andrades **553**
Estudio fraseológico basado en el corpus CORBICON

Ivo Fabijanić, Lidija Štrmelj **564**
The Adaptation of Anglicisms - Phraseological Units in Croatian Economic Terminology

Valentina Piunno **572**
Italian Multiword Adverbs: distributional features and functional properties. A corpus based analysis

Antonio Rico Sulayes **587**
Contribution of multi-element features in automatic text classification for authorship attribution

Ana María Ruiz Martínez **597**
La marcación de las unidades fraseológicas a partir del examen de corpus

Abdelghani Yahiaoui, Joseph Dichy **605**
Une approche mixte de l'alignement sous-phrastique de corpus parallèles arabes-anglais

FOREWORD

The European Society of Phraseology (EUROPHRAS) organised its 2015 Conference in Malaga, Spain, from 29 June to 1 July. Under the rather suggestive title "Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives", this edition focused on phraseology from various computational and technology-orientated perspectives. More specifically, the Conference intended to create a platform for the presentation and discussion of the most recent and advanced computational and corpus-based methods applied in Phraseology, and consequently to promote further development and innovations in the field, including not only monolingual, but also multilingual phraseology and translation.

The EUROPHRAS2015 Conference intended to build solid bridges between Phraseology and Technology, namely, Corpus Linguistics, Natural Language Processing (NLP), Computational Linguistics and Machine Translation. For the first time, the EUROPHRAS Conference had a co-located event: the 2nd Workshop on "Multi-word Units in Machine Translation and Translation Technology" (MUMTTT2015), whose first edition was successfully held at the Machine Translation Summit (Nice, 2013). The invited speaker of the MUMTTT2015 was Dr. Kathrin Steyer, President of EUROPHRAS, who contributed to the "bridge" by delivering an insightful paper on corpus-driven approaches to the description of multiword patterns and networks in language use.

This edition was also one of the most international ones so far, with over 150 speakers and attendants from all over the world: Europe (Austria, Belgium, Bulgaria, Czech Republic, Crimea, Croatia, France, Germany, Hungary, Italy, Poland, Portugal, Russia, Switzerland, Slovakia, Slovenia, Spain, Ukraine and United Kingdom), Asia (North Korea, South Korea, Japan), Africa (Tunisia, Emirates) and Canada, United States and Latin America (Mexico and Brazil).

The academic programme featured four invited speakers and four conference panels (around 100 papers out of 167 submissions) covering (i) corpus-based Phraseology, (ii) NLP and/or corpus-based identification and classification of phraseological units, (iii) computer-aided and/or corpus-based analysis of phraseological units, and (iv) Phraseology in e-lexicography and e-terminology. Papers addressed a variety of research aspects and applications in a monolingual, multilingual or cross-language fashion for a good number of languages (Arabic, Basque, Bulgarian, Croatian, Czech, English, French, German, Italian, Japanese, Korean, Romanian, Russian, Polish, Portuguese and Spanish). The plenary talks revolved around hot and central topics in NLP and/or corpus-based Phraseology, with special emphasis on collocation and cross-language issues. Jean-Pierre Colson discussed the contribution of corpus-based Phraseology to Translation Studies; Patrick Hanks presented a corpus-driven approach to meaning and phraseology within TNT (*Theory of Norms and Exploitations*); Ulrich Heid examined linguistic knowledge extraction about collocations from corpora with a view to translation and text production; and Ruslan Mitkov proposed a novel 'knowledge- and resource-poor' methodology for the identification and translation of multiword expressions from comparable corpora across language pairs.

Special emphasis was laid on research, particularly on early career researchers and women's researchers with two Best Paper Awards that were announced during the main conference. The Best Paper by a Young Researcher Award was granted to Shiva Taslimipoor for her work "Cross-Lingual Extraction of Multiword Expressions", whereas Zuriñe Sanz Villar got the Best Paper Award by a Woman for her work "German-into-Basque/Spanish Translation of Binomials in a Parallel and Multilingual Corpus".

Special thanks go to the conference organisers and the volunteer students, the programme committee, the invited speakers and the conference attendants, as well as to EUROPHRAS, SIGLEX and PARSEME. And a very special thought for late Adam Kilgarriff, dear colleague and member of the programme committee, who would surely have enjoyed EUROPHRAS2015, the fruits of our joint labour and the challenges that lie ahead.

Gloria Corpas Pastor

Chair of EUROPHRAS2015

PREFACIO

La Sociedad Europea de Fraseología (EUROPHRAS) organizó su congreso de 2015 en Málaga (España) del 29 de junio al 1 de julio. Bajo el sugestivo título de "Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives/ Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües", esta edición se ha centrado en diferentes aspectos de la fraseología relacionados con las tecnologías lingüísticas. En concreto, se pretendía crear una plataforma para presentar y debatir los métodos computacionales y basados en corpus más recientes e innovadores aplicados a la fraseología con el objetivo de fomentar el desarrollo y avance de la disciplina en sus vertientes monolingüe y multilingüe, así como en el campo de la traducción.

Con EUROPHRAS2015 se ha pretendido tender un puente sólido entre la Fraseología y la Tecnología, esto es, la Lingüística de Corpus, el Procesamiento de Lenguaje Natural, la Lingüística Computacional y la Traducción Automática. De hecho, por primera vez en la historia se ha contado con un evento asociado al congreso: la segunda edición del "Workshop on Multi-word Units in Machine Translation and Translation Technology" (MUMTT2015), cuya primera edición tuvo lugar junto al Machine Translation Summit celebrado en Niza (2013). En esta ocasión, la conferenciante invitada ha sido la actual presidenta de EUROPHRAS, la Dra. Kathrin Steyer, quien contribuyó de forma notable a la construcción del "puente" con un magnífico trabajo sobre enfoques de corpus aplicados a la descripción de redes y patrones multiverbales en el uso de la lengua.

Esta edición ha sido, además, una de las de mayor participación internacional, con más de 150 participantes y asistentes provenientes de países de todo el mundo: Europa (Alemania, Austria, Bélgica, Bulgaria, Crimea, Croacia, Eslovaquia, Eslovenia, España, Francia, Hungría, Italia, Polonia, Portugal, República Checa, Reino Unido, Rusia, Suiza y Ucrania), Asia (Corea del Norte, Corea del Sur y Japón), África (Túnez y Emiratos Árabes), Canadá, Estados Unidos y Latinoamérica (México y Brasil).

El programa académico comprendió cuatro conferenciantes invitados y cuatro paneles (unas 100 comunicaciones de un total de 167 propuestas recibidas) sobre i) Fraseología basada en corpus, ii) Identificación y clasificación de unidades fraseológicas basadas en corpus o mediante técnicas de PLN, iii) Análisis de unidades fraseológicas basado en corpus o asistido por ordenador, y iv) La información fraseológica en la lexicografía y la terminología electrónicas. Los distintos trabajos abordaron diversos aspectos y aplicaciones de la investigación fraseológica desde una perspectiva monolingüe, multilingüe o contrastiva, y en relación a un amplio número de lenguas (alemán, árabe, búlgaro, checo, coreano, croata, español, francés, inglés, italiano, japonés, polaco, portugués, rumano, ruso y vasco). Las conferencias plenarias versaron sobre aspectos centrales y muy actuales de la Fraseología basada en corpus o mediante técnicas de PLN, con especial referencia al fenómeno colocacional y el contraste entre lenguas. Jean-Pierre Colson abordó la contribución de la Fraseología basada en corpus para los Estudios de Traducción; Patrick Hanks presentó un enfoque de corpus para el análisis semántico de base fraseológica en el marco de la TNE (*Theory of Norms and Exploitations*); Ulrich Heid se ocupó de la extracción de conocimiento lingüístico sobre colocaciones a partir de corpus con vistas a la traducción

y a la producción textual; y Ruslan Mitkov propuso un método novedoso para la identificación y traducción de expresiones multiverbales que utiliza como único recurso corpus comparables en varias lenguas.

En esta edición se ha hecho especial hincapié en la investigación fraseológica realizada por jóvenes y mujeres mediante la concesión de dos premios durante el congreso. El Premio al Mejor Trabajo de Jóvenes Investigadores recayó en Shiva Taslimipoor, por su trabajo titulado "Cross-Lingual Extraction of Multiword Expressions" [Extracción de expresiones multiverbales en varios idiomas]; mientras que el Premio al Mejor Trabajo Realizado por una Mujer le fue concedido a Zuriñe Sanz Villar por su trabajo "German-into-Basque/Spanish Translation Analysis of Binomials in a Parallel and Multilingual Corpus" [Análisis de las traducciones de binomios del alemán al vasco y al español en un corpus paralelo multilingüe].

Mi más sincero agradecimiento al comité organizador, a los voluntarios, al comité científico, a los conferenciantes invitados, a los ponentes y asistentes, así como a EUROPHRAS, SIGLEX y PARSEME. Un recuerdo muy especial también a la memoria de Adam Kilgarriff, querido colega y miembro del comité científico, quien hubiera disfrutado a buen seguro de EUROPHRAS2015, de los frutos del esfuerzo conjunto y de los muchos retos que aún nos quedan por superar.

Gloria Corpas Pastor

Presidenta de EUROPHRAS2015

CORPUS-BASED PHRASEOLOGY

FRASEOLOGÍA BASADA EN CORPUS

LOCUCIONES EN EUSKERA: NECESIDAD Y PAUTAS PARA SU RECOPILACIÓN A PARTIR DE LOS CORPUS TEXTUALES EXISTENTES

Xabier Altzibar
Universidad del País Vasco
josejavier.alcibar@ehu.eus

Xabier Bilbao
Universidad del País Vasco
xabier.bilbao@ehu.eus

Resumen

El artículo plantea la utilidad que tendría para los vascoparlantes un diccionario de locuciones actual y moderno, como un primer paso para futuros diccionarios bilingües o multilingües. Valora la utilidad de OEH [Diccionario General Vasco] y los corpus actuales como depósitos para la extracción de fraseologismos y analiza algunos fenómenos lingüísticos de pérdida de uso, recurso al calco, etc. que hacen más patente la necesidad de un diccionario de este tipo. Plantea también algunos problemas y particularidades que presentan las locuciones vascas de cara a una recopilación, como, por ejemplo, la cuestión de si una sola palabra puede conformar una locución. Propone aceptar las locuciones de uso actual y atestiguado, aunque sean calcos, e informar suficientemente sobre la forma, categoría y acepciones, ilustrando la información con ejemplos modernos.

1. INTRODUCCIÓN: EL CONTEXTO BILINGÜE

Las locuciones cobran especial importancia en un contexto bilingüe. En la Comunidad Autónoma Vasca del Estado español la enseñanza en euskera (lengua vasca) y español es general y se va implantando el modelo de enseñanza trilingüe (euskera-español-inglés). En ese contexto, el uso de las locuciones ayuda a mejorar la competencia comunicativa y la expresividad pero, al mismo tiempo, esas piezas lexicales constituyen uno de los obstáculos más difíciles de superar en el aprendizaje de la segunda o tercera lengua. Por otra parte, los hablantes vascoparlantes traducimos constantemente expresiones fijas del español, incluso en la conversación diaria.

Es cierto que con la unificación el euskera se ha convertido en una lengua más culta, en parte gracias a la traducción de obras modernas y clásicas de las principales lenguas del entorno. Pero, realmente, ¿cuenta el euskera con diccionarios o instrumentos que den una buena información de la forma y significado de las locuciones y demás fraseologismos? ¿La

expansión del euskera unificado ha impulsado también el uso de esas unidades? ¿Qué particularidades o problemas específicos pueden plantear las locuciones euskéricas a un lexicógrafo que pretenda recogerlas? ¿Qué criterios debería seguir éste y qué información debería proporcionar?

En este artículo pretendemos contestar a esas preguntas, en dos bloques. En el primero explicaremos por qué es necesaria para los vascoparlantes una recopilación o un diccionario de locuciones, y para ello analizaremos concisamente el valor y la utilidad de los principales instrumentos de consulta existentes (1), y describiremos brevemente tanto fenómenos que suponen pérdidas de uso como también, probablemente, compensaciones (2). En el segundo bloque ofreceremos unas pautas para una recopilación de las locuciones, planteando previamente algunos problemas y particularidades (3), y proponiendo a continuación criterios generales para la confección y ordenación de la información (4). Concluiremos recalcando la urgencia y utilidad de una recopilación o diccionario actual que informe suficientemente de las locuciones en uso (5).

En este estudio nos basamos en una muestra de somatismos importantes, como son los formados por *aho* 'boca' y *esku* 'mano'. Como se sabe, los nombres de las partes del cuerpo son imágenes o metáforas y éstas representan conceptos que, gracias a la imagen, se hacen más comprensibles.

En cuanto al concepto de locución que manejaremos, denominamos locuciones (siguiendo a Corpas: 1996: 88) a un determinado tipo de unidades fraseológicas, propias del sistema de la lengua, que tienen fijación interna, unidad de significado y fijación externa pasemática. La unidad de significado es su principal característica, más que la no composicionalidad semántica. Su significado figurado varía: muchas tienen significado composicional, es decir, se pueden entender hasta casi literalmente (*ahoz aho*, esp. *de boca en boca*); otras son no composicionales y semánticamente menos transparentes o más opacas (*ahoa neurri* esp. *a pedir de boca*, lit. 'la boca como medida').

2. VALOR Y UTILIDAD DE LAS RECOPILACIONES, CORPUS Y DICCIONARIOS NORMATIVOS

La primera recopilación importante de locuciones es el *Manual del vascófilo* de P. Zamarripa (Zamarripa 1913). El autor llama modismos a locuciones, fórmulas y otras expresiones peculiares del euskera y traduce las mismas al español.

R.M. Azkue recopiló casi 3000 refranes y más de 2000 locuciones y modismos en su obra *Euskalerriaren Yakintza–Literatura Popular del País Vasco* entre 1935 y 1947 (Azkue 1935-1947). La mayoría de ellos fueron recogidos directamente de los hablantes. Contienen ejemplos con traducción en español, sin especificar muchas veces la forma de la locución, y con información del dialecto en que se usan, o de la fuente escrita en el caso de modismos recogidos de los libros.

Guilsou publicó parcialmente su recopilación de fraseologismos (*Errantegia*), que ordenó según una palabra en francés y tradujo literalmente a muchos otros idiomas (Guilsou 1950). D. Intza recogió refranes y locuciones en Navarra, con breve explicación, en euskera, de su significado (Intza 1974).

K. Izagirre publicó unas 7.000 locuciones y otras expresiones que denominó expresamente, y por vez primera en euskera, locuciones (Izagirre: 1981). Las entradas especifican la forma de la locución, contienen al menos un ejemplo, expresan el significado por medio de una palabra-clave no figurativa en euskera, y en muchos casos con

locuciones equivalentes en francés y español. Aparecen clasificadas por categorías gramaticales, lo que dificulta su búsqueda. Por otra parte, no faltan expresiones literarias ingeniosas o locuciones espúreas. Hoy día es consultable en la red (Intza Proiektua), con la ventaja de que se pueden hacer búsquedas por conceptos, palabras integrantes de la locución y palabras relacionadas con su significado.

La recopilación más amplia anterior a la era informática es la de J. M. Mokoroa (Mokoroa 1990). A pesar de su título (*Repertorio de locuciones del habla popular vasca*), la mayor parte de los más de 92.000 ejemplos son de fuente literaria o escrita, y no todas son locuciones. Está ordenada por conceptos. En realidad, se trata de frases donde aparece marcada en negrita la expresión peculiar y traducida en español; sin embargo, no se especifica la forma de la locución. Ofrece también información del dialecto y la fuente. Se puede consultar en CD o internet.

La recopilación actual más completa y valiosa es el Diccionario General Vasco– Orotariko Euskal Hiztegia (DGV-OEH) de 16 volúmenes, obra de Euskaltzaindia– Academia de la Lengua Vasca, realizada por Mitxelena y Sarasola entre 1987-2005. Recoge información sobre el léxico de los libros escritos hasta la unificación del euskera aproximadamente. Es un corpus no etiquetado de 6 millones de palabras, que se puede consultar en formato electrónico. Junto a las palabras individuales, están dispuestas con entrada propia las unidades de dos o más palabras, entre las que se encuentran tanto palabras compuestas como fraseologismos, los cuales deben ser identificados por el lexicógrafo.

Además, son dignos de mención bastantes lexicones y estudios lexicográficos de hablas de determinadas localidades, que incluyen locuciones y fraseologismos, no pocos de los cuales faltan en DGV-OEH.

Además de OEH, Euskaltzaindia ha publicado dos diccionarios normativos. El primero, *Hiztegi Batua* [Diccionario Unificado], abreviadamente HB, recoge la forma de las palabras del euskera batua y es consultable en formato electrónico. El segundo, *Euskaltzaindiaren Hiztegia: Adierak eta Adibideak* (2012) [Diccionario de Euskaltzaindia: Acepciones y ejemplos], abreviadamente EHAA, tiene como objeto definir las palabras y diferenciar las acepciones, y no está aún disponible en formato electrónico. Ambos diccionarios recogen algunas locuciones, pero no la mayoría de ellas ni sus variantes, ni dan información suficiente de la forma de la locución o acepciones figuradas.

Aparte de recopilaciones y diccionarios, poseemos una serie de corpus de textos escritos del euskera actual. Algunos de los más importantes son los siguientes: *Ereduzko Prosa Gaur* (EPG) [Prosa de referencia actual], de 25 millones de palabras (2001-2008), y su continuación, *Ereduzko Prosa Dinamikoa* [Corpus dinámico de prosa de referencia] (2009-2013), constituidos ambos por textos periodísticos y literarios y elaborados por el Instituto de Euskara de la Universidad del País Vasco. Euskaltzaindia ha creado *XX. mendeko Euskararen Corpusa* (XXMECE) [Corpus del euskera del s. XX], muestra variada de textos, de 4,6 millones de palabras, y *Lexikoaren behatokiaren corpusa* (LBC) [Corpus del observatorio del léxico], de 11 millones de palabras, constituido por textos periodísticos del s. XXI. La Fundación Elhuyar ha creado *Web-corpusen Ataria* (WCA) [Portal de corpuses de la web], corpus de textos extraídos de la web, de 125 millones de palabras, aproximadamente. Todos ellos son corpus etiquetados lingüísticamente, al menos a nivel morfológico. Para una información más completa sobre los corpus véase Salaburu 2015: 559-576.

En contraposición a los diccionarios que recogen el léxico tradicional (OEH) y normativos (HB, EHAA), estos corpus de textos escritos son –junto a los orales, que no mencionaremos aquí por falta de espacio– útiles y necesarios para conocer el uso real actual del euskera.

Sin embargo, estos corpus no están específicamente orientados a la extracción de fraseologismos (tan solo se pueden buscar pares frecuentes de palabras). En ellos no aparecen locuciones coloquiales. Una de las razones para ello estriba en que en los corpus priman los textos periodísticos informativos y las traducciones de novelas de otros idiomas; y apenas se recogen obras teatrales ni otras que reflejen el léxico popular.

Recientemente, la Fundación Labayru ha publicado en la red un diccionario fraseológico, *Labayru Hiztegi Fraseologikoa* [Diccionario Fraseológico Labayru], que, de momento, ofrece la traducción de unas 3.500 locuciones o frases hechas del español al euskera, sumando unas 18.000 en esta lengua (incluidas las variantes), pero desprovistas de entrada e información sobre su uso; tan solo algunas de ellas van acompañadas de algun ejemplo o sinónimos en español. Es un proyecto interesante para la traducción, que pretende completarse con la versión euskera-español.

En resumen:

1. Reconociendo el mérito de las recopilaciones individuales anteriores, a partir de OEH contamos con instrumentos tecnológicamente más modernos que proporcionan cada vez más información. Otro de los méritos de las recopilaciones analizadas consiste en que ofrecen equivalencias, expresiones correspondientes o traducciones en español o francés. Sin embargo, las recopilaciones no distinguen generalmente entre fraseologismos y no fraseologismos, ni entre las locuciones y otros fraseologismos.

2. Los repertorios, diccionarios y corpus que poseemos son instrumentos relativamente válidos y complementarios entre sí para extraer locuciones. Pero ninguno de ellos describe sistemáticamente el uso actual (estructuras morfosintácticas y significación en el contexto) ni proporciona la información suficiente necesaria.

3. PÉRDIDAS Y POSIBLES COMPENSACIONES

Partimos del hecho de que en el País Vasco y Navarra el bilingüismo es asimétrico (el 60% de la población es monolingüe española) y los individuos bilingües dominan el español. Se podría pensar que la expansión del euskera unificado impulsaría o traería emparejado un uso mayor de las locuciones. Nuestra impresión es que se están produciendo pérdidas pero posiblemente también compensaciones. Mencionaremos primeramente tres fenómenos actuales que atestiguan la pérdida.

1º. En situaciones de alternancia de código (*code switching*) en una conversación entre vascoparlantes, muchos de ellos recurren directamente a locuciones del español. La introducción de la locución española se hace frecuentemente con una justificación en euskera: "erderaz esaten den moduan" [como se dice en español].

> *Horregaz, erderaz esaten den moduan... "hay que tener mil ojos". Horrek beti "arrima el ascua a su sardina".* [con ese, como se dice en español, hay que tener mil ojos. Siempre arrima el ascua a su sardina].

2º. En la lengua hablada y escrita existe una tendencia a sustituir las locuciones por palabras abstractas. Por ejemplo, en la tradición se han usado tanto la palabra abstracta *aipatu* 'mencionar' como *aho(t)an erabili/hartu/eduki* o *aho(ta)ra ekarri* ('mencionar', lit. 'llevar/tomar/traer a la boca'), pero actualmente se usa casi exclusivamente *aipatu*.

3º. Algunas locuciones polisémicas (la polisemia es la regla y no la excepción) van perdiendo acepciones figurativas, reduciéndose sus significados. Por ejemplo, *ahoan kurutze* [lit. 'cruz en la boca'] con los verbos *egon, gelditu, utzi* 'estar, quedar, dejar' significa 'sin comer', 'sin nada', *in albis*, 'burlar(se) (de uno)', 'con un palmo de narices'. Hoy tan solo algunos escritores la usan en esta última acepción.

Sin embargo, existen otros fenómenos ambivalentes que, gestionados convenientemente, pueden resultar beneficiosos: por ejemplo, el calco y la concurrencia entre variantes.

En el euskera hablado o escrito de gente cultivada son normales muchos calcos del español, aunque otros son rechazados. Actualmente se usan en el periodismo calcos que no figuran en OEH ni en el Diccionario Fraseológico Labayru, pero están bien atestiguados en algunos corpus actuales como LBC. Un ejemplo de lo dicho es *aho txikiarekin* (o *aho txikiz*). [*con la boca chica* (o *chiquita*, o *pequeña*). adv (col.). Por mero cumplido o sin convicción. Normalmente con el verbo decir, DFDEA]. La locución española citada proviene del registro coloquial, lo que deja aún más patente la influencia del calco en el euskera normativo.

La concurrencia entre variantes acarreará pérdidas al irse imponiendo tan solo una de ellas (según inferimos de los datos del corpus EPD), pero puede redundar en beneficio de una mayor unificación:

> *ahotik ahora, ahotik hortzera, hitzetik hortzera* ['repentinamente', 'de improviso'] → *hitzetik hortzera.*

Generalmente, se va imponiendo la más breve:

> *Aho(a) bete hortzekin, aho bete hortz* (con los verbos *utzi, gelditu* 'dejar (a), 'quedar (una pers.) con un palmo de narices' –> *aho bete hortz.*

O la equivalente a la locución española:

> *Esku abil/trebe/egoki/on* [*buena mano* 'habilidad para tratar a alguien o para algo'] → *esku on.*

Estos hechos nos inducen a pensar que las pérdidas y las probables compensaciones, sobre todo en la lengua hablada, están relacionadas con fenómenos como las interferencias del español, la normalización y normativización del euskera, el consiguiente postergamiento del euskera popular y dialectal (ámbito tradicional y natural de las locuciones) y la ley de la economía del lenguaje. Por todo ello, poseer un buen instrumento de consulta de los fraseologismos ayudaría a frenar o equilibrar el proceso de pérdida, disminución o sustitución de las locuciones vascas.

4. PROBLEMAS Y PARTICULARIDADES

Una recopilación de las locuciones vascas tendría que plantearse el siguiente problema: siendo el euskera una lengua aglutinante, con sufijos pospositivos, al contrario que las lenguas flexivas que tienen preposiciones ¿podría considerarse locución una sola palabra con sufijo? Por ejemplo:

> *behingoan* [*a la vez, a un tiempo*; *de repente*; 'rápidamente'; *de una vez* 'en una sola acción', 'por fin'; 'total y completamente'; 'una primera vez'; 'inmediatamente'; lit. 'vez-de-la-en']
>
> *(-en) eskutik* [*por medio de, por obra de; por cuenta de; de la mano* 'juntamente' con los verbos *venir, ir;* lit. '(-de) mano-de'].

En ambos casos se trata de unidades sintagmáticas formadas por una palabra que es el núcleo; cambiando un elemento se cambia el significado (por ejemplo, *(-en) eskura* [*a cargo, a disposición,* 'sometido a'; *a mano* 'cerca', *al alcance*].

Tradicionalmente, algunos recopiladores como Zamarripa y Azkue han incluido entre los modismos a unidades de una sola palabra, y, actualmente, algunos investigadores consideran que en lenguas no flexivas como el euskera se podrían considerar como locuciones palabras que constituyen una unidad tipográfica (Urizar: 2012: 59, 79), si bien en la práctica, a efectos de tratamiento computacional, el grupo IXA no toma en cuenta las locuciones de una única palabra en su base de datos EDBL [Base de Datos lexical del euskera].

Una particularidad de las locuciones euskéricas es la abundancia y complejidad de variantes lexicales y morfosintácticas. He aquí sendos ejemplos:

> *eskua estutu (*o *eskua eman/luzatu,* o *eskua(k) tinkatu)* [*estrechar la mano,* 'saludar'].
>
> *eskurik esku, eskutik eskura* [*de mano en mano* 'de una persona a otra'].

La razón de esta abundancia y complejidad es la fragmentación dialectal. Una prueba clara de esta fragmentación es la existencia de locuciones de ámbito local. Por ejemplo, Azkue recoge en su diccionario (1905) bastantes locuciones de localidades vizcaínas, muchas de las cuales no parecen estar actualmente en vigor.

5. PAUTAS PARA LA RECOPILACIÓN

En un trabajo de recogida o extracción de locuciones, el lexicógrafo debería, a nuestro juicio, aceptar únicamente las locuciones cuyo uso actual esté atestiguado en corpus, lexicones o estudios, incluso si fueran calcos recientes. También se debería considerar que los fraseologismos tienen un contexto cultural y un uso pragmático, y se deberían extraer los significados a partir de contextos reales o prototípicos. Asimismo, se deberían diferenciar las categorías gramaticales y describir brevemente las estructuras morfosintácticas diferentes de cada locución.

En cuanto al orden, se podrían agrupar las locuciones en orden alfabético bajo una palabra ordenadora, que sería la palabra base. Cada locución debería tener su entrada correspondiente, que proporcionaría la siguiente información, en el siguiente orden:

1. La forma de la locución (y, en su caso, de las variantes usadas).

2. La categoría gramatical de la locución.

3. (En caso necesario) El nivel de uso o registro, la actitud del hablante (humorístico, despectivo, irónico etc.), la frecuencia o la extensión geográfica o dialectal.

4. El significado o las diversas acepciones figuradas, cada una al menos con un ejemplo actual en su contexto (o, a veces, con un ejemplo por cada estructura morfosintáctica diferente).

Un posible modelo podría ser DFDEA, que ordena las locuciones bajo la palabra base.

6. ALGUNAS CONCLUSIONES GENERALES

Necesitamos un diccionario de locuciones que informe suficientemente de su uso, preferiblemente en formato electrónico por razones de utilidad. Otra opción sería que los diccionarios nuevos marquen las locuciones y contengan entradas de las mismas. En todo caso, una u otra opción deberían diferenciarlas de otras unidades fraseológicas (colocaciones, fórmulas, paremias, etc.).

Por supuesto, también un diccionario de locuciones trilingüe o mejor aún cuatrilingüe (vasco-español-francés-inglés) con equivalencias o al menos explicación de las locuciones contribuiría a mejorar la comunicación en general.

Una última reflexión. Euskaltzaindia–Academia de la Lengua Vasca y las instituciones encargadas de fomentar y enseñar el euskera deben favorecer el uso —sobre todo en la lengua hablada— de las locuciones más extendidas. Para ello, la Academia, tras unificar las locuciones equivalentes de los diferentes dialectos, debería incorporar en sus diccionarios las de uso actual extenso o incluso más restringido, atestiguadas en corpus u otros repertorios modernos, y especificar sus acepciones figuradas. El peligro de desaparición de las locuciones puede aumentar si el euskera unificado no asimila y unifica la diversidad de las mismas.

Bibliografía

AZKUE, R.M. 1935-1947. *Euskalerriaren Yakintza–Literatura Popular del País Vasco*, III.–2ª edición: Espasa Calpe, 1969. 233-351.

CORPAS PASTOR, G., 1996. *Manual de fraseología española*. Madrid: Gredos.

DFDEA: Seco, M.–Andrés, O. & Ramos, G. 2004. *Diccionario fraseológico documentado del español. Locuciones y modismos españoles actual*. Aguilar.

EHAA: EUSKALTZAINDIA – ACADEMIA DE LA LENGUA VASCA. 2012. *Euskaltzaindiaren Hiztegia. Adierak eta adibideak* [Diccionario de Euskaltzaindia – Acepciones y Ejemplos].

GUILSOU, P. "Erran zahar eta erran-airak" [Dichos antiguos]. (Publicado parcialmente, sin nombre de autor en el semanario *Herria* en la década de 1950).

INTZA, D. 1974. *Naparroa-ko euskal-esaera zarrak*. [Antiguos dichos de Navarra] Pamplona: Diputación Foral de Navarra - Institución Príncipe de Viana.

IZAGIRRE, K. 1981. *Euskal lokuzioak espainolezko eta frantsesezko gida-zerrendarekin*. [Locuciones vascas con listado en español y francés] Hordago.

MAKAZAGA, J.M. 2009. *Elgoibarko hizkera*. [El habla de Elgoibar] Tesis de doctorado. UPV-EHU.

MOKOROA, J. M. 1990. *Ortik eta emendik. Repertorio de locuciones del habla popular vasca*. Labayru-Eusko Jaurlaritza-Etor eds. URL: http://www.hiru.com/hirupedia.

SALABURU, P. 2015. "Hizkuntza corpusak eta euskara", *Eridenen du zerzaz kontenta: Sailkideen Omenaldia Henrike Knörr irakasleari (1947-2008*). Ezeizabarrena, M.J.–Gomez, R. (ed.). 559-576.

URIZAR, R. 2012. *Euskal lokuzioen tratamendu konputazionala*. [Tratamiento computacional de las locuciones vascas - Tesis de doctorado] Universidad del País Vasco, UPV-EHU.

ZAMARRIPA, P. *Manual del vascófilo. Libro de modismos, onomatopeyas, elipsis, uso distinto de la s y la z, y otras cosas que conviene saber para hablar y escribir bien en vascuence vizcaíno*. Bilbao: J. A. de Lerchundi, 1913.

Córpuses y diccionarios en la web

DGV-OEH: MITXELENA, L., SARASOLA, I., 1987-2005. *Diccionario General Vasco–Orotariko Euskal Hiztegia*. 16 vls. Bilbao: Academia de la Lengua Vasca – Euskaltzaindia. [3a. edic., 2013]. URL: http://www.euskaltzaindia.eus/oeh .

EDBL: GRUPO IXA Base de Datos lexical del euskera. URL: http://ixa2.si.ehu.es/edbl/.

EPD: INSTITUTO DE EUSKERA DE LA UNIVERSIDAD DEL PAÍS VASCO, EPD: *Ereduzko Prosa Dinamikoa* [Corpus Dinámico de Prosa de referencia]. URL: http://ehu.eus/ehg/epd/.

EPG: INSTITUTO DE EUSKERA DE LA UNIVERSIDAD DEL PAÍS VASCO, EPG: *Ereduzko Prosa Gaur* [Prosa de referencia actual]. URL: http://ehu.eus/euskara-orria/euskara/ereduzkoa/.

HB: EUSKALTZAINDIA – ACADEMIA DE LA LENGUA VASCA. *Hiztegi Batua*. [Diccionario Unificado]. URL: www.euskaltzaindia.eus.

INTZA PROIEKTUA. *Euskal lokuzioak sarean* [proyecto Intza. Locuciones vascas en la red], http://www.intza.armiarma.com.

LABAYRU IKASTEGIA. *Labayru Hiztegi fraseologikoa* [Diccionario Fraseológico Labayru]. URL: www.hiztegia.labayru.eus.

LBC: EUSKALTZAINDIA – ACADEMIA DE LA LENGUA VASCA. *Lexikoaren behatokiaren corpusa*. [Corpus del observatorio del léxico]. URL: http://lexikoarenbehatokia.euskaltzaindia.eus.

XXMECE: EUSKALTZAINDIA – ACADEMIA DE LA LENGUA VASCA. *XX. mendeko Euskararen Corpusa*. [Corpus del euskera del s. XX]. URL: http://www.euskaracorpusa.net/Xxmendea.

WCA: ELHUYAR. *Web-corpusen Ataria*. [Portal de corpuses de la web]. URL: http://webcorpusak.elhuyar.org.

FRASEOLOGÍA ¿SEXISTA?: LO QUE EL CORPUS ESCONDE

Gloria Corpas Pastor
Universidad de Málaga
gcorpas@uma.es

Resumen

En este trabajo abordamos los usos sexistas del lenguaje, con especial referencia a la fraseología como uno de los recursos más efectivos, pero menos estudiados desde la perspectiva de género. En cierto sentido, las guías de lenguaje no sexista han servido para denunciar la discriminación e invisibilidad social de la mujer, y reivindicar, al mismo tiempo, sus derechos fundamentales. La fraseología sexista ha sido objeto de estudio con respecto a los valores e ideas contenidos en refranes y algunas locuciones que denigran la imagen femenina, expresan misoginia, o fomentan actitudes de dominación hacia la mujer. Pero frente a la fraseología explícitamente sexista, existen otras unidades aparentemente inocuas, pero de contenido extraordinariamente machista. De ello nos ocuparemos en este trabajo de carácter introductorio.

1. INTRODUCCIÓN

El lenguaje es un reflejo del androcentrismo típico de la sociedad. Por androcentrismo entenderemos el estudio, análisis o investigación realizado desde una perspectiva eminentemente masculina, que se presenta como central a la experiencia humana. Desde esta visión del mundo, el varón es la referencia por defecto, y la mujer, "lo otro" (Gilman, 1911). El sexismo lingüístico es una de las consecuencias del androcentrismo, como también lo es la discriminación y la (in)visibilidad de "lo otro". En palabras de Celia Amorós (1995): "Es sabido que quien tiene el poder es quien da nombres a las cosas (y a las personas)". Los estereotipos, los roles preestablecidos, la publicidad o los medios de comunicación, entre otros, contribuyen sin duda a alguna a perpetuar actitudes discriminatorias y de exclusión social.

Los movimientos feministas vienen llamando la atención desde hace varias décadas sobre estos aspectos y su reflejo en el lenguaje, lo cual se ha intentado paliar parcialmente mediante la publicación de diversas de guías de lenguaje no sexista. Pero frente a los ejemplos típicos de sexismo lingüístico, de los cuales forma parte también la fraseología explícitamente sexista, existen otras unidades aparentemente inocuas, pero de contenido extraordinariamente machista. De ellas nos ocuparemos en este trabajo, de naturaleza introductoria, en el cual vamos a realizar un estudio de diversas unidades fraseológicas con metodología de corpus (cf. Colson, 2011; Corpas Pastor, 2013 y 2014). En primer lugar, trataremos los fenómenos de sexismo lingüístico analizados en las guías y las propuestas formuladas en estas para evitarlos. Y, en segundo lugar, abordaremos el sexismo lingüístico

en relación a la fraseología, con especial referencia a las valoraciones y usos implícitos que se desvelan mediante el análisis de corpus. Para ello, haremos uso de WebCorp,[1] un sistema de gestión de hipertextos que permite utilizar la red como repositorio textual o corpus. Finalmente, haremos un breve balance, a modo de resumen y líneas de trabajo futuro.

2. LAS GUÍAS DE LENGUAJE NO SEXISTA

El sexismo lingüístico constituye un aspecto controvertido en el que convergen, junto al Feminismo, la Sociología, la Lingüística, la Filosofía y la Política, entre otras disciplinas. Ya desde la base no parece haber acuerdo sobre si el lenguaje es el origen-causa de la discriminación social hacia la mujer, o si simplemente es el reflejo de tal situación, cuya responsabilidad habría que buscarla en otros agentes implicados. En otras palabras, la dificultad radica en dilucidar si el lenguaje es sexista *per se* o si más bien se detectan usos sexistas del lenguaje.

En cualquier caso, y sin profundizar en tales cuestiones, lo cierto es que desde los albores del nuevo siglo proliferan guías de lenguaje no sexista que aspiran a aumentar la visibilidad de la mujer y luchan para combatir la discriminación real de este colectivo en la sociedad actual. Con idéntica finalidad que los tradicionales libros de estilo, las guías contienen una serie de criterios preceptuados para la redacción de documentos a modo de normas lingüísticas y de estilo. En definitiva, dichas guías son producto de una política lingüística concreta que aspira a normalizar el discurso (especialmente el administrativo y el de las instituciones públicas) a fin de eliminar cualquier sesgo discriminatorio contra la mujer. La mayoría de las guías de lenguaje no sexista recogidas en el informe redactado por Ignacio Bosque para la RAE (Bosque, 2012) se centran en señalar (y denunciar) el abuso del masculino genérico como manifestación típica del lenguaje sexista. En su lugar, se ofrecen pautas para evitar el uso del masculino que abarca ambos sexos mediante el empleo de otras alternativas posibles dentro del sistema de la lengua española. Por lo general se trata de formas colectivas, metonímicas, perifrásticas o impersonales, sin marcas de género o que visibilizan lo femenino mediante desdoblamientos, barras y aposiciones, como ilustramos en la Tabla 1 a partir de ejemplos tomados de las citadas guías.

MASCULINO GENÉRICO	ALTERNATIVAS	
La presión social obliga a **los jóvenes** a buscar pareja a una edad muy temprana.	***APOSICIONES EXPLICATIVAS***	La presión social obliga **a los jóvenes, de uno y otro sexo,** a buscar pareja a una edad muy temprana.
El **alumno** deberá asistir puntualmente a clase.	***COLECTIVOS***	El **alumnado** deberá asistir puntualmente a clase.
Los **directivos** acudirán a la cena con sus *mujeres.*	***DESDOBLAMIENTOS***	Los **directivos y las directivas** acudirán a la cena con sus *parejas.*
Unos expresan opiniones que sinceramente no entiendo.	***GENÉRICOS***	**Unas personas** expresan opiniones que sinceramente no entiendo.
El secretario está facultado para decidir sobre la adquisición de mobiliario nuevo.	***METONIMIAS***	**La secretaría** está facultada para decidir sobre la adquisición de mobiliario nuevo.
Sus investigaciones sobre **el hombre** le han valido la concesión de diversos premios.	***PERÍFRASIS***	Sus investigaciones sobre **el género humano** le han valido la concesión de diversos premios.
Los trabajadores deben exigir sus derechos.	***ARROBA***	**L@s trabajador@s** deben exigir sus derechos.
El solicitante deberá aportar el justificante de pago.	***BARRAS***	**El/la solicitante** deberá aportar el justificante de pago.
Todos los miembros recibirán la acreditación correspondiente.	***DETERMINANTES SIN MARCA DE GÉNERO***	**Cada miembro** recibirá la acreditación correspondiente.
El que haya visto ya la película, que levante	***RELATIVOS SIN***	**Quien** haya visto ya la película, que levante la

[1] URL: http://www.webcorp.org.uk/live/.

la mano.	***MARCA DE GÉNERO***	mano.
Podrán presentarse **los estudiantes** desempleados sin experiencia en el sector.	***OMISIÓN DE DETERMINANTES***	Podrán presentarse **estudiantes** desempleados sin experiencia en el sector.
El juez **dictará** sentencia en el plazo de un mes.	***ESTRUCTURAS CON SE***	La sentencia **se dictará** en el plazo de un mes.
Cuando el usuario reciba la notificación, *debe* formalizar la solicitud.	***FORMAS VERBALES NO PERSONALES***	**Al recibir** la notificación, *se debe* formalizar la solicitud.

Tabla 1. *Principales formas alternativas al uso del masculino genérico en español.*

Otros aspectos que se identifican como discriminatorios y sexistas en las guías hacen referencia a la asimetría en las formas de tratamiento (por ej., *Sr. / Sra. de*), a los duales y a las denominaciones de profesiones tradicionalmente realizadas por hombres. En los duales aparentes, la forma masculina presenta el significado primario del lexema con connotaciones neutras, mientras que la forma femenina siempre, de connotaciones negativas, siempre implica menosprecio a la mujer. Es el caso de *lagarto* ('reptil') / *largarta* ('mujer taimada, prostituta') o *sargento* ('militar') y *sargenta* ('mujer autoritaria, hombruna'). En la mayoría de los casos la forma femenina del dual se utiliza con el significado de 'prostituta', como sucede con *perra*, *zorra*, *golfa*, *fulana*, etc.; usos que recoge el DRAE y que son despectivos hacia el género femenino. En cuanto a la utilización del género en profesiones tradicionalmente realizadas por varones, pero que actualmente ejercen también mujeres, las guías realizan diversas propuestas que ilustran la variabilidad existente: forma femenina o forma masculina genérica con artículo femenino (por ej., *juez / jueza*, *la juez; edil / edila*, *la edil*; *cabo / la cabo*). Sin embargo, no sucede lo mismo con profesiones de las que hasta hace poco sólo se hacían cargo las mujeres y a las que se han incorporado los varones, donde la única opción posible es la forma masculina marcada: por ej., *enfermera / enfermero* (*el enfermera) o *matrona / matrón* (*el matrona).

3. EL SEXISMO A TRAVÉS DE LA FRASEOLOGÍA

Las guías de lenguaje no sexista presentan una postura beligerante en contra del uso del masculino genérico y de la invisibilidad de la mujer, sin entrar en mayores profundidades. Ahora bien, es bien sabido que el lenguaje sexista se perpetúa de muy diversas maneras, y una de ellas es a través de los estereotipos e imágenes de la denominada sabiduría popular. El máximo exponente sería el refranero español, que se caracteriza, precisamente, por su anticlericalismo, su antisemitismo y su misoginia (Calero Fernández, 1999: 130). Este aspecto, sin embargo, rara vez tiene cabida en las guías al uso.

Abundan las paremias que desprestigian a la mujer, fomentando una actitud negativa y agresiva hacia ella, que pasa por la imposición de roles sociales, como, por ejemplo, *La mujer en la casa y con la pata 'quebrá'* (Corpas Pastor, 1996: 169). La imagen femenina se presenta desde la perspectiva del varón, que rechaza todo lo que no se corresponda con el ideal masculino de mujer tradicional (femenina, dócil, sumisa, fiel, honesta, etc.). Algunos ejemplos ilustrativos de misoginia son *Dos hijas y una madre, la perdición de un padre*, *Palabra de mujer no vale un alfiler,* y *La mujer y la mentira nacieron el mismo día muestran misoginia*; otros implican dominación (*La oveja y la mujer, puesto el sol en casa estén, Gallinas y mujeres entre cuatro paredes*), denigración (*Más tiran dos tetas que dos carretas*) e invisibilidad (*Las señoritas no tienen espalda*).

Pero no solo esas 'perlas' de sabiduría popular vehiculan el sexismo sociolingüístico y cultural de generación en generación. Otros tipos de unidades fraseológicas, de apariencia inocua, van cargadas ideológicamente con valoraciones e implicaturas negativas hacia la mujer. Aquí entrarían locuciones como *quedarse para vestir santos* (**1. loc. verb**. Quedarse soltera., DRAE) o *quedarse alguien compuesto y sin novio/novia.* (**1. loc. verb. coloq.** No lograr

lo que deseaba o esperaba, después de haber hecho gastos o preparativos, creyéndolo indefectible., DRAE). Aunque el DRAE solo lo refleja en el primer caso,[2] ambas locuciones están restringidas a la mujer y resaltan un determinado rol reservado para ella: casarse y formar una familia como única vía de realización personal. Y no son valoraciones neutras precisamente, como muestran las líneas de concordancia analizadas mediante WebCorp: *quedarse compuesto y sin novio* se aplica mayoritariamente a mujeres, pero, en las raras ocasiones cuando se aplica a un hombre, se le ridiculiza a través de la resonancia de connotaciones negativas propias del sujeto femenino, que pasan a implicar fallo o fracaso del varón:

La Kim Basinger, de Cher (otra vez	re**compuesta**	**y sin novio**	, hasta Rossy de Palma, de Ute Lemper
dijo en su día: Me han dejado una hija	**compuesta**	**y sin novio**	y no me dejan a otra". La frase fue pronunciada
niña bien. "Entonces, ¿yo me quedo	**compuesta**	**y sin novio**	" –replicaba Juanita. Y ahí intervenía yo
an un Gobierno nacionalista no monocolor y dejen	**compuesto**	**y sin novia**	al pretendiente Benegas. Sea cual sea
Años, eso sí que no. El jodido Robledo se va a ir	**compuesto**	**y sin novia**	. A lo mejor, si hay cielo, encubierto
Resulta que no tenía copiloto. Vamos, que se quedó	**compuesto**	**y sin novia**	. Eso le pasa por no llevar un

Tabla 2. *Concordancias para el nodo [compuesto/a y sin novio/a]*

El trasvase de connotaciones negativas del femenino al masculino es una estrategia de manipulación creativa propia del discurso sexista y que se observa con frecuencia en los duales. En el siguiente ejemplo, extraído de WebCorp, *verdulero* ('hombre que vende verduras', DRAE) refleja un proceso de cambio semántico por influencia de *verdulera* ('mujer ordinaria y grosera', DRAE): *"[…] jajaja cada loco con su tema, estos del G20 son unos verduleros señora, ni tienen educación ni nada. Si les viera su madre el rapapolvo sería de órdago verdad?"*.[3]

Otras locuciones como *de pelo en pecho* (**1. loc. adj. coloq.** Dicho de una persona: Vigorosa, robusta y denodada., DRAE) y *de rompe y rasga* (**1. loc. adj. coloq.** De ánimo resuelto y gran desenfado., DRAE) son definidas en el diccionario de forma neutra. Sin embargo, el análisis de corpus revela que ambas encierran una valoración subjetiva (Tablas 3-4). Mientras la primera resalta la virilidad del hombre y su conformidad con el estereotipo, la segunda solo se aplica al sexo femenino y denota una mujer ruda, con carácter y trasgresora, que se sale del estereotipo o ideal femenino tradicional.

que no sabrá mucho de economía pero es hombre	**de pelo en pecho**	y bravuconerías, decidió resolver el problema en
mujer, no aman a ¡¡ESPAÑA!!! El español, hombre	**de pelo en pecho**	sólo llora ante la bandera y al escuchar el himno
Para dejar de pagar la deuda hacen falta hombres	**de pelo en pecho**	dispuestos a todo. A la Troika no hay que
Era el Bertín Osborne de Hollywood, un machote	**de pelo en pecho**	como muestra en esta imagen de los ochenta, la
tu voto!" es la consigna de este político	**de pelo en pecho**	que se presenta por Cambio Democrático, el
hubiera una componente muy andrógina en épocas	**de pelo en pecho**	resultaba notable, sobre todo porque se unía

Tabla 3. *Concordancias para el nodo [de pelo en pecho]*

la conocieran y la consideraran una artista	**de rompe y rasga**	. Mujer polifacética, ha sido presentadora de
vista aparece «como una mujer de arriba abajo y	**de rompe y rasga**	», «extraordinaria y valiente», y «siempre en
una evolución vocal, también, desde la cantante	**de rompe y rasga**	hasta el falsete delicado, acariciante. Desde el
hay varias mujeres que antes se denominarían	**de rompe y rasga**	y que ahora serían mujeres que saben lo que
Hemingway, Man Ray o Max Ernst) es un mito	**de rompe y rasga**	y toda una heroína del pop de los ochenta. Los
autoridad de la Región de Murcia, doce mujeres	**de rompe y rasga**	que han roto el techo de cristal que les

Tabla 4. *Concordancias para el nodo [de rompe y rasga]*

[2] Sobre la evolución de lo femenino y el sexismo en el diccionario de la Academia, véanse los trabajos recogidos en Calero Fernández, Forgás Berdet y Lledó Cunill (2004).

[3] La referencia a lo femenino como estrategia de insulto y vituperio del varón es una constante en la lengua española (cf. *ser una nenaza*, *llorar como una niña/mujer*, etc.). Y lo es desde tiempos inmemoriales, como ilustra la unidad que denigra la figura Boabdil el Chico tras la caída de Granada en 1492: *Llora como una mujer lo que no supiste defender como un hombre* (y sus variantes).

Una situación parecida ocurre en relación a la locución *(ser una) mosca/mosquita muerta*[4]: "**1**. **f. coloq.** Persona, al parecer, de ánimo o genio apagado, pero que no pierde la ocasión de su provecho." (DRAE). Si bien se aplica tanto a hombres como a mujeres, el análisis de las concordancias nos permite establecer que esta unidad aparece más frecuentemente asociada a mujeres (cf. *loba, experta, ésa, Xisca, esta, falsa, la*), con una valoración negativa implícita por engaño o artimaña (Tabla 5).

trabajando y demostrar lo que vales. -Loba no. ¿	**mosquita muerta**	? -Tampoco, tampoco. Puede que en algún papel
Kirkman. Pero poco o nada tiene que ver aquella	**mosquita muerta**	con la experta en ingeniería mecánica que se
: lo haga, pero ésa no soy yo. No me gusta ir de	**mosquita muerta**	. -¿Qué ha encontrado en Alberto? -Llegó justo e
las groupies de Rafa que pensaban que eras una	**mosquita muerta**	" Querida Xisca: Se veía venir... Has tardado un
de esta», «que es una falsa», «parece una	**mosquita muerta**	, pero ya, ya...», etc., etc. Al día siguiente,
También la calificó en el mismo debate de	**mosquita muerta**	» y «mala persona» tras lo cual textualmente

Tabla 5. *Concordancias para el nodo [mosquita muerta]*

Se da, además, una diferencia curiosa de uso según la empleen hombres o mujeres. Ellos la utilizan para describir a una mujer que, aunque al principio no lo parezca, terminará dando problemas: "*La mosquita muerta es esa chica apocada de la fila de atrás, la vecina distraída, esa compañera del curro –un poco– sosa. Nada las identifica, no hay indicios ni luces en la pista de aterrizaje. Solo una cosa las une: convertirá tu vida en un infierno.*" En el caso de ellas se utiliza frecuentemente para hacer referencia a una "roba novios": *"[…] la mujer mosquita muerta es aquella amiga nuestra, divina, generosa y hasta bonita físicamente hablando… Siempre y cuando no aparezca un hombre apetecible a cien kilómetros a la redonda*". De esta manera, la misma unidad fraseológica, dependiendo del sexo del hablante, se utiliza para denigrar a la mujer en dirección de aquello a lo que cada uno más teme. Esta peculiaridad no se recoge en el diccionario.

Finalmente, los usos creativos del lenguaje pueden suponer un uso sexista de unidades que, *a priori*, no son discriminatorias en sí mismas. Valga como ilustración el caso de la unidad *(ser) la excepción que confirma la regla* en el siguiente ejemplo, tomado de una línea de concordancia de WebCorp: "*El ser humano es inteligente, la excepción es quien tiene la regla*".

4. COROLARIO

En un primer momento, las guías de lenguaje no sexista han servido para denunciar la invisibilidad social de la mujer, y reivindicar, al mismo tiempo, su derecho fundamental a la igualdad y a alcanzar el protagonismo que se le ha venido negado sistemáticamente. No obstante, dichas guías se han ocupado casi exclusivamente del masculino genérico, proponiendo otras formas y posibilidades que ofrece la propia lengua para expresar el plural. También se han tratado los duales aparentes, las formas de tratamiento y las denominaciones de profesiones tradicionalmente reservadas a los varones. Estos aspectos, aunque importantes, se quedan, sin embargo, en la superficie. Precisamente, una de las maneras más eficaces de perpetuar los roles y estereotipos socio-culturales que subyacen a la discriminación de la mujer es a través de la fraseología, la parte más idiosincrásica de una lengua y que rara vez se contempla en las citadas guías.

La fraseología sexista ha sido objeto de estudio con respecto a los valores e ideas contenidos en refranes y algunas locuciones que denigran la imagen de la mujer, expresan misoginia, o fomentan actitudes misóginas o de dominación hacia la mujer. Algunos ejemplos ya mencionados son *La mujer en casa y con la pata quebrada*, *Palabra de mujer no vale un*

[4] El DRAE solo recoge la unidad *mosca muerta* y no su otra variante (*mosquita muerta*), a pesar de que esta última es más frecuente que la primera (6 y 34 apariciones respectivamente en el CREA).

alfiler, *Más tiran dos tetas que dos carretas* y similares. Pero frente a los ejemplos típicos de sexismo en el lenguaje, de los cuales forma parte también la fraseología explícitamente sexista, existen otras unidades aparentemente inocuas, pero de contenido extraordinariamente machista. Estos sesgos sutiles del lenguaje no pueden ser desmontados con el seguimiento de simples guías de lenguaje no sexista, por cuanto conllevan implicaturas y valoraciones sancionadas por la comunidad hablante. Tales matices no son fácilmente detectables y, por tanto, requieren ser descubiertos mediante el análisis de corpus. La metodología de corpus va más allá, por cuanto permite detectar usos sexistas en unidades fraseológicas desprovistas en principio de tal carga, pero que adquieren ese matiz mediante manipulaciones creativas. Los aspectos de género trascienden el mero uso del masculino o femenino, o la explicitación sistemática de la relación entre género y sexo.

Desde el punto de vista de la lingüística de corpus, el reto está en descubrir el sexismo implícito que se esconde en el uso del lenguaje. A día de hoy, no queda claro que solo los lectores masculinos se sientan aludidos por unidades como *Tonto el que lo lea* o *El hombre es un lobo para el hombre*; tampoco se ha comprobado el grado de adopción real de las distintas propuestas contenidas en las guías de lenguaje no sexista (por ejemplo, en Google *jóvenes de uno y otro sexo* presenta 838.000 resultados frente a los más de 47 millones de resultados que arroja el masculino genérico *jóvenes*); ni se ha estudiado la frecuencia de uso real ni la distribución diatópica de las unidades fraseológicas explícitamente sexistas, como tampoco se ha analizado el uso sexista de unidades fraseológicas aparentemente neutras. Son muchos los frentes que tenemos abiertos y para los que la metodología de corpus resulta especialmente adecuada. Sería conveniente, pues, seguir explorando estas nuevas vías de investigación antes de que *se nos pase el arroz*.

Bibliografía

AMORÓS, C. (DIR). 1995. *10 palabras clave sobre mujer*. Estella: Verbo Divino.

BOSQUE, I. 2012. Sexismo lingüístico y visibilidad de la mujer. Informe presentado en el Pleno de la RAE. Disponible en: http://www.rae.es/noticias/el-pleno-de-la-rae-suscribe-un-informe-del-academico-ignacio-bosque-sobre-sexismo.

CALERO FERNÁNDEZ, Mª A. 1999. *Sexismo lingüístico, análisis y propuestas ante la discriminación sexual en el lenguaje*. Madrid: Narcea.

CALERO FERNÁNDEZ, M.ª A.; FORGAS BERDET, E.; LLEDÓ CUNILL, E. (Coords.) 2004. *De mujeres y diccionarios: evolución de lo femenino en la 22ª edición del DRAE*. Madrid: Ministerio de Trabajo y Asuntos Sociales, Instituto de la Mujer.

COLSON, J.-P. 2010. The Contribution of Web-Based Corpus Linguistics to a Global Theory of Phraseology. En Ptashnyk, S., Hallsteindóttir, E., Bubenhofer, N. 2010. (Eds.). *Corpora, Web and Databases. Computer-Based Methods in Modern Phraseology and Lexicography*. Hohengehren: Schneider Verlag. 23-35.

CORPAS PASTOR, G. 1996. *Manual de fraseología española*. Madrid: Gredos.

CORPAS PASTOR, G. 2013. Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. En: I. Olza y E. Manero (eds.): *Fraseopragmática*. Berlín: Frank & Timme, 335-373.

CORPAS PASTOR, G. 2014. El fraseólogo internauta: cómo pasarlo pipa en la red. En: V. Durante (ed.). *Fraseología y paremiología: enfoques y aplicaciones*. Madrid: Instituto Cervantes, n.º 5 Serie «Monografías» (Biblioteca fraseológica y paremiológica). 133-152. Disponible en: http://cvc.cervantes.es/Lengua/biblioteca_fraseologica/default.htm

GILMAN, C. P. 1911. *The Man-made World, or Our Androcentric Culture*. Nueva York: Charlton Company.

REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual*. Disponible en: http://www.rae.es.

REAL ACADEMIA ESPAÑOLA. 2001: *Diccionario de la lengua española*. 22.ª edición. Madrid: Espasa Calpe. [En línea]. Disponbible en: http://www.rae.es. [DRAE].

SYSTEMATIC VS. NON-SYSTEMATIC COLLOCATIONAL PATTERNS IN LSP: PARADIGMATIC VARIATION IN THE TECHNICAL DOMAIN

Laura Giacomini
Department of Translation and Interpreting
University of Heidelberg
laura.giacomini@iued.uni-heidelberg.de

Abstract

This paper presents detailed empirical observations on the topic of collocational variation in LSP texts of the technical domain and on its relevance for text production and translation. Despite the great heterogeneity of variation examples, encoding often takes place along systematic and reproducible combinatory models, which this study aims to capture with the help of samples of domain-specific textual genres. The study deals with the technical subdomain of the automotive sector and concentrates on paradigmatic, i.e. lexical variation in collocation, a subtype of formal variation in which a set of two or more collocations refer to the same conceptual item. Given a cluster of collocations with the same syntactic structure, this phenomenon may involve any of their constituents; for instance, the Italian collocations meaning "fog light" and displaying the structure A + N allow for variation on the level of the noun/base (*faro fendinebbia, proiettore fendinebbia, luce fendinebbia*) as well as on the level of the adjective/collocate (*faro fendinebbia, faro antinebbia*). The findings from the study on paradigmatic variation of collocation in domain-specific contexts have been summarised in a variational description model, which could be applied in termbases and lexicographic microstructures.

1. INTRODUCTION

The background to this contribution is rooted in a study on variation of LSP collocations that is presently being carried out with the purpose of discovering systematic (i.e. context-specific) patterns in variation, to develop a functional typology of collocation variants and to model a termbase which is suitable for describing this phenomenon. A broad definition of collocations will be employed here to cover not only n-grams consisting of a base and one or more collocates, but also multiword expressions and compounds (Roth 2014). This perspective contributes towards a coherent, morphology-independent understanding of terminological variation and takes into account language-specific preferences. For this reason, it is extremely beneficial for text production, translation and lexicography, which are central to this study (Kerremans 2010, Schmitt 2006). The topic of

variation in collocational patterns is analysed with reference to the automotive sector. In comparison with other technical subfields, the automotive sector is often considered a relatively easy field in translation, because "in all major languages there are very good dictionaries for automotive subjects, mechanical and electric engineering, and all other subjects related to cars" (Sofer 2006: 102). Despite this encouraging situation, variation remains a weak point in terminological description, even in domains with above-average lexicographic coverage.

The study of variation of LSP collocations concentrates on three languages - Italian, German and English - and has been performed on comparable corpora, including different textual genres as well as supplementary online texts. After presenting a typology of variation that remodels and integrates Daille's proposal (Daille 2005), this contribution will focus on paradigmatic variation and introduce the idea of systematic variation as linked to specific context-dependent criteria. The phenomenon of LSP variation and its possibly systematic aspects will then be exemplified primarily by means of Italian collocations, which will be accompanied by English and German equivalents. The results of variation analysis will lead to a proposal for a description model that aims to account for all relevant information concerning collocational variants. Finally some conclusions will be drawn, in particular regarding critical issues related to the topic, and suggestions for future research will be made.

2. A TYPOLOGY OF VARIATION

A revised version of Daille's variation model (Daille 2003 and 2005) that takes into consideration translation and lexicographic needs is based on a tripartite typology including graphical, syntagmatic and paradigmatic variants, each including further specifications (Table 1):

GRAPHICAL VARIATION	
- Orthographic v.	four-wheel drive (Subaru) / four wheel drive (Mitsubishi)
- Abbreviations, Acronyms	four-wheel drive / 4WD
SYNTAGMATIC VARIATION	
- Morphological v.	
-- Grammatical morpheme v.	power-transmitting device / power-transmission device
-- Lexical morpheme v.	vehicle directional control / directional control of the v.
- Syntactic v.	brake wheel / hand brake wheel
PARADIGMATIC VARIATION	
- Base v.	forward gear / forward speed
- Collocate v.	compression-ignition engine / diesel engine
- Total v.	gasoline motor (AE) / petrol engine (BE)

Table 1. Typology of collocational variation

This paper concentrates on paradigmatic variation, which will be defined as follows:

Paradigmatic variation of an LSP collocation is the lexical variation of ≥1 of its constituents, whereby available (i.e. norm- or usage-based) variants form sets (paradigms) of mutually exclusive lexical items with a specific syntactic role, so that different combinations of variants produce clusters of synonymous terminological collocations.

This definition accounts for collocations of two or more constituents and only for synonymous variants, i.e. those variants which carry the same meaning, even though their textual distribution may differ according to specific contextual characteristics.

3. SYSTEMATIC VARIATION AND PROPOSAL FOR A DESCRIPTION MODEL

The method that has been employed in this case study consists of four main steps:

- building comparable corpora focussing on the topics Lighting, Braking System and Transmission System and including typical textual genres: rules and regulations, data sheets, specialised magazines, user's guides, and marketing texts. Rules and regulations include, for instance, UN regulations such as the World Forum for Harmonization of Vehicle Regulations WP.29, regulations of the Society of Automotive Engineers (SAE) and the Federal MotorVehicle Safety Standard (FMVSS/CMVSS 108), as well as national traffic codes (e.g. Italy: Codice della strada, Germany: STVO). Automotive terminology can also be partly found in ISO norms (cf. section 3.1.1);
- performing corpus analysis on collocation variants recorded in a technical dictionary, G. Marolli's *Grande Dizionario Tecnico Inglese* (Hoepli 2014), attempting to validate them and possibly detect new variants;
- taking into consideration collocations with varying frequencies and also allowing for low frequencies in the case of gaps due to the features of the specific corpus, for instance the under-representation of certain subtopics;
- integrating results with further online materials.

From a purely syntactic perspective, the most common structures that appear in the comparable corpora are A+N (e.g. *proiettore anabbagliante/ low beam*) and N+N (e.g. *sistema di frenatura/ brake system*), whereas collocations of the kind V+N (e.g. *sbloccare la leva del cambio/ to unblock the gear lever*) appear less frequently. The examples which will be mentioned later in this chapter belong to the first two categories.

The topic of systematicity is closely related to the idea that the distribution of synonymic variants is mostly not arbitrary, rather it depends on certain contextual features. Systematic patterns, in other words, are patterns recurring and expected in a contextual environment. This environment is a set of features that combine in different constellations to build a specific text and that may allow for preferences in the choice of a collocation variant or impose a more restricted choice.

a) communicative situation: domain-internal vs. domain-external

b) text-internal features: topic, genre, syntax

c) text-external features: e.g. diatopic level, diastratic level

A few observations need to be made about these features. The type of communicative situation along the vertical dimension of the LSP has been quite broadly distinguished as domain-internal (with experts addressing experts) and domain-external (with experts addressing non-experts). A more fine-grained distinction has been avoided for the moment in order to allow for easier categorization. Text-internal features, namely the topic (intended as the overall specialised theme the text deals with, such as Lighting or Braking System), the textual genre and the specific syntactic structure (Roelcke 2010) may also influence variant distribution. However, text-external features can also play an important role in this perspective, for instance diatopic variation (the availability of regional variants, e.g. *passing beam* AmE vs. *low beam* BrE) and diastratic variation (the use of company-internal terminologies) (Sofer 2006).

A rather straightforward and yet important consideration that has to be made is that the presence of collocation variants often depends on the language-specific tendency to variation displayed by single word terms, with their collocations inheriting the same tendency: for instance, It. *luce/ faro/ proiettore*, En. *light/ lamp* and De. *Licht/ Leuchte/ Scheinwerfer* represent highly productive clusters of synonyms and, as some of the examples will show, this peculiarity is transferred to their collocations. Other automotive terms, on the contrary, seem to have a more fixed nature, which, again, is reflected by their collocations (e.g. It. *frizione*, En. *clutch*, De. *Kupplung* or It. *freno*, En. *brake*, De. *Bremse*). Can the degree of variation of an LSP collocation be predicted by taking into consideration only the variation of its constituents? The answer is probably no. Of course, the context-based examples will show that variation of the single constituents gives a collocation a certain predisposition to variation, but collocations are likely to follow specific context-dependent patterns and behave as autonomous terminological items.

3.1. Context-based examples

This section introduces three interesting examples that aim to show how systematicity in variation can be tested by analysing corpus data according to the parameters a) *communicative situation*, b) *text-internal features* and c) *text-external features*, which have been previously described. Each example includes a set of variants of the same collocation as they have been found in the technical dictionary, selected corpus data referring to the occurrence of these variants in specific genres and, eventually, a table summarising context-based information concerning each variant.

3.1.1 Topic: Lighting / Fog light

Variants in the technical dictionary		
proiettore fendinebbia	fog light	Nebelleuchte
faro fendinebbia	fog lamp	Nebelscheinwerfer
luce fendinebbia		
fendinebbia		
faro antinebbia		
antinebbia (*rare in the corpus*)		

Fog lights, or fog lamps, are part of the automobile lighting system and, as such, have symbols that have been defined by the ISO 2575:2010 norm: a green symbol for the front fog light and a yellow one for the rear fog light.

In Italian, a large number of variants with the structure A+N can be found in the technical dictionary. They may vary both on the base level (*proiettore/faro fendinebbia*) and the collocate level (*faro fendinebbia/antinebbia*). They also allow for syntagmatic variation, whenever the adjective undergoes a process of substantivation and stands for the whole concept (*il faro fendinebbia* > *il fendinebbia*). A close look at the corpus occurrences unveils interesting aspects regarding these variants' behaviour. Here are some particularly relevant excerpts:

1) **proiettore fendinebbia** anteriore: il dispositivo che serve a migliorare l'illuminazione della strada in caso di nebbia, caduta di neve, pioggia o nubi di polvere. (RULES & REGULATIONS: CDS Art. 151)

2) La spia si accende attivando le **luci fendinebbia** anteriori… **proiettori fendinebbia**… **fendinebbia** (USER'S GUIDE: Fiat 500 Libretto di uso e manutenzione della vettura online)

3) Le **luci retronebbia** si possono inserire solo con i **fendinebbia anteriori** o i fari anabbaglianti accesi. ... **fari fendinebbia** (USER'S GUIDE: Ford Focus Manuale di istruzioni)

The Codice della strada (CDS), the Italian traffic code, only contains the variant *proiettore fendinebbia*, whereas user's guides show the highest degree of variability and interchangeability, which can often be explained in terms of syntactic and/or pragmatic choices (cf. example 3: *fendinebbia anteriori*, front fog lights, has been preferred to the longer *fari/luci fendinebbia anteriori*, in particular due to its proximity to the paired term *luci retronebbia*, rear fog lights, and to the subsequent term *fari anabbaglianti*, low beams, in which both *luci* and *fari* are used).

The results of the corpus-based analysis of contexts in which the Italian equivalents of *fog light* occur are summarised in Table 2 below:

communicative situation	**genre**	**terms**
domain-internal	rules and regulations*	proiettore fendinebbia
	data-sheet*	fendinebbia
	marketing text	proiettore/faro fendinebbia
domain-external	specialised magazine*	fendinebbia
	user's guide**	luce/faro/proiettore fendinebbia
	marketing text	faro fendinebbia/ (antinebbia)

Table 2. Equivalents of *fog light* in the Italian corpus

The summarising table shows the distribution of the variants in the Italian corpus according to genre and communicative situation. Genres with high variability and interchangeability (**) can be distinguished from those which show a clear preference for a single variant (*). Marketing texts published, for instance, by replacement of equipment stores may address either domain experts or non-expert end-users, and for this reason have been classified both as a domain-internal and a domain-external genre. Moreover, analysis of corpus data reveals a rare occurrence of the term *antinebbia*, both as an adjective and as a noun.

3.1.2. Topic: Lighting, Braking System / Brake light

Variants in the technical dictionary		
luce di arresto	brake light	Bremslicht
luce di stop	stop light	Bremsleuchte
luce stop	stop lamp	
stop		

The variants proposed by the technical dictionary are also present in the corpus. Once again, the Codice della strada (CDS) contains a single option, whereas user's guides are characterised by a greater flexibility that, at times, depends on the syntactic structure of the text. Among the following corpus excerpts, example 2 illustrates how synthetic wordings (e.g. inventories, enumerations, summaries) may allow for syntagmatic variation, in this case the ellipsis of the function word: *interruttore luci stop* instead of *interruttore delle luci di stop*, and *nodo quadro strumenti* instead of *nodo del quadro degli strumenti*. Elimination of the preposition is not compulsory (cf. example 4) and does not necessarily involve the presence of a lexicalised variant: *luci stop* is a functional syntagmatic variant of *luci di stop* which serves stylistic purposes and, from a synchronic perspective, cannot be seen as an independent term. However, this is a quite frequent phenomenon which deserves mention in terminological representation (see section 3.2).

1) **luce di arresto**: il dispositivo che serve ad indicare agli altri utenti che il conducente aziona il freno di servizio (RULES & REGULATIONS: CDS Art. 151)

2) Interruttore **luci stop**, nodo quadro strumenti (USER'S GUIDE: Fiat 500 Libretto di uso e manutenzione della vettura online)

3) **Stop** (**luci di arresto**) (USER'S GUIDE: Lancia Ypsilon Libretto di uso)

4) Prodotti > Lampadine per auto > **Luce di stop**/posteriore/di posizione (MARKETING TEXT: Bosch Auto Parts)

5) Quando si solleva il piede destro dall'acceleratore, infatti, la i3 non si limita a rallentare ma frena proprio, tanto che si accendono gli **stop** posteriori... In rilascio si accendono le **luci di stop**. (SPECIALISED MAGAZINE: Quattroruote)

6) Lampadine e Fanaleria: Lampadine per auto classiche e speciali Fanali per automezzi Luci di posizione: anteriori, posteriori e laterali Proiettori abbaglianti ed anabbaglianti Proiettori di svolta e di retromarcia **Luci di stop** Indicatori di direzione Fendinebbia: anteriori e posteriori Luci di marcia diurna (MARKETING TEXT: http://www.marautoricambinapoli.it/prodotti)

The results of the corpus-based analysis of contexts in which the Italian equivalents of *brake light* occur are summarised in Table 3 below:

communicative situation	genre	terms
domain-internal	rules and regulations*	luci di arresto
	data-sheets*	stop
	marketing texts*	luci di stop
domain-external	specialised magazine	luci di stop, stop
	user's guide**	luci di arresto, luci di stop, stop
	marketing texts*	luci di stop

Table 3. Equivalents of *brake light* in the Italian corpus

3.1.3. Topic: Transmission System / Clutch release bearing

Variants in the technical dictionary		
cuscinetto distacco frizione cuscinetto di disinnesto frizione cuscinetto reggispinta della frizione reggispinta distacco frizione manicotto distacco frizione (*not found in the corpus*)	clutch release bearing	Kupplungsausrück-lager

Different to the examples in the previous paragraphs, the concept which is presented here refers to a small component of the transmission system. As such, it is not mentioned in rules and regulations or in user's guides. The clutch release bearing appears in a limited number of textual genres, mostly in marketing texts for an expert audience (e.g. manufacturers, service technicians, etc.) that needs to buy components in order to build or repair a vehicle. *Cuscinetto reggispinta* is the most frequent Italian term, but many other terms can be found which are made up of 2-3 elements and display variations of the base or of the collocate. Syntagmatic variation on the level of the preposition is also relevant in terms of frequency.

1) La frizione si compone di una coppia di dischi (disco e spingi disco) che si trovano tra il volano del motore e il **cuscinetto reggispinta**. (SPECIALISED MAGAZINE: Giornale Motori)

2) Il **reggispinta distacco frizione** è un ricambio che permette la separazione del disco e del meccanismo. Il reggispinta distacco frizione prende la forma di un cuscinetto o di un anello. (MARKETING TEXT: http://ricambi-auto.mister-auto.it/)

3) **Cuscinetto reggispinta** (MARKETING TEXT: http://www.ricambihondaoriginali.it)

4) **Manicotto disinnesto frizione (**MARKETING TEXT: www.ricambinuovi.com**)**

Only one of the terms entered in the technical dictionary, *manicotto distacco frizione*, could not be found in the corpus, whereas a new variant, *manicotto disinnesto frizione*, joined the initial set because of its presence in the corpus and in online comparable texts. The corpus-based analysis of contexts in which the Italian equivalents of *clutch release bearing* has led to the results summarised in Table 4 below:

communicative situation	genre	terms
domain-internal	marketing text**	cuscinetto reggispinta, reggispinta distacco frizione, cuscinetto distacco frizione, cuscinetto di disinnesto della frizione, manicotto disinnesto frizione
domain-external	specialised magazine*	cuscinetto reggispinta

Table 4. Equivalents of *clutch release bearing* in the Italian corpus

3.2. Proposal for a description model

Regularities found in the distribution of variants in specific contextual environments should be part of their terminological representation. The description model proposed in this section has not been structured according to a specific purpose and is only intended to mention relevant descriptive elements in a coherent way. These elements will now be introduced with the help of the example made in section 3.1.2, the three Italian variants indicating the concept of brake light, namely *luce di arresto*, *luce di stop* and *stop*. Each descriptive item is signalled by a superscript and explained right after the model.

topic: lighting[3] **luce di arresto**[1] **(luce di stop, stop)**[2]

domain-internal[4]: rules and regulations[5] (CDS[6])

domain external: user's guides (*cf. company TL*[7])

topic: lighting **luce di stop (luce di arresto, stop)**

■ **luce stop**[8] (*rare; synthetic texts*[9])

domain-internal: marketing texts

domain-external: specialised magazines
user's guides (*cf. company TL*)

marketing texts

topic: lighting **stop** (*pl.*[10]) **(luce di arresto, luce di stop)**

domain-internal: data-sheets

domain-external: specialised magazines

user's guides (*cf. company TL*)

[1] reference term

[2] variants

[3] topic

[4] vertical dimension

[5] genre

[6] specific genre

[7] diastratic information (here: reference to the existence of company-internal terminology)

[8] syntagmatic variant

[9] pragmatic information

[10] grammatical information

Possible applications of this model could be, for instance, termbases and e-lexicographic microstructures. Each item (i.e. topic, genre, term, etc.) could be potentially employed as an access route to variational information in order to ensure coherent representation of any cluster of variants. For instance, in an electronic dictionary a query based on the input 'genre: specialised magazine' would lead to the terms *luce (di) stop* and *stop*, whereas the input 'domain-external communication' would lead to all variants, without a genre-specific distinction.

Generally speaking, the same genre can belong to different communicative situations: marketing texts can address technicians or end users, data-sheets can be found in user's guides as well as in regulations, and specialised magazines can also have different audiences. For this reason, the two descriptive items play an equally meaningful role in variation description. The ultimate goal of the model is, in fact, to collect and make available all relevant data that enable the user to embed each variant in the most suitable context.

4. CONCLUSIONS AND OUTLOOK

This paper has introduced a classification of collocational variation and a definition of context-based systematicity, referring to a set of examples from the automotive sector and proposing a description model which attempts to cover variational information in a coherent way. A few observations will now be made on critical issues that have emerged during analysis and that should be considered in future work:

- Corpus building: Size and composition should be discussed in detail to ensure a sufficient level of representativeness in terms of variation coverage. Moreover, a distinction should be made between original and translated texts, especially in the case of marketing texts. A context-dependent LSP corpus annotation including metadata (e.g. vertical dimension, genre) is a key premise in this argument, independent of the final application of the proposed description model.
- Typology: The need for an extensive typological study of variation in the technical domain and of overlapping types has become apparent, especially from a multilingual, contrastive perspective.
- Diachronic perspective: The diachronic motivation behind the emergence of some variants is an aspect which could possibly be taken into account in order to explain certain usage preferences.
- Company-specific terminology: It would be useful to analyse company-specific terminology on a large scale, in order to establish if there are always systematic choices behind it.
- Syntax-driven variation: Systematicity in the way in which syntactic structures influence variant distribution is a topic of great interest that deserves deeper analysis; it should not be subordinated to lexical and textual aspects.

Two other substantial aspects have not been mentioned in this paper but will be investigated in subsequent steps of the study. The first one is the role of borrowings in variation: both lexical and semantic borrowings, in fact, seem to play an important part in increasing variation (*luce di stop*, for instance, is a clear semi-calque of the English term *stop light*) and, possibly, in determining contextual preferences (Giacomini 2012). The second

aspect regards conceptual categories behind variation: identifying conceptual relations among terminological items, for instance of the type 'kind of' (*faro retronebbia/ rear fog light* is a kind of *faro antinebbia/ fog light*, which is a kind of *faro/ light*) and 'part of' (*pignone/ pinion* is part of the *sistema di trasmissione/ transmission system*), can help better capture and describe synonymous terms.

References

DAILLE, B., 2003. Conceptual Structuring Through Term Variations. In: F. Bond, A. Korhonen, D. MacCarthy and A. Villacicencio (eds.), *Proceedings ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, 9-16, 2003*.

DAILLE, B., 2005. Variants and application-oriented terminology engineering. *Terminology*, volume 1.

DELPECH, E., 2011. Un protocole d'évaluation applicative des terminologies bilingues destinées à la traduction spécialisée. *RNTI - Revue des Nouvelles Technologies de l'information* 2011, pp. 23–48.

GIACOMINI, L., 2012. Lexical borrowings in German and Italian IT terminology: At the crossroads between language interference and translation procedures. In: *Proceedings of the BDÜ Conference "Übersetzen in die Zukunft 2012"*, Berlin 28-30.09.12.

KERREMANS, K., 2010. A Comparative Study of Terminological Variation in Specialised Translation. In: C. Heine / J. Engberg, eds. *Reconceptualizing LSP. Online proceedings of the XVII European LSP Symposium 2009*, Aarhus.

MAROLLI, G., 2014. *Grande Dizionario Tecnico Inglese*, Milano: Hoepli.

ROELCKE, T., 2010. *Fachsprachen*, Berlin: Erich Schmidt.

ROTH, T., 2014. *Wortverbindungen und Verbindungen von Wörtern. Lexikografische und distributionelle Aspekte kombinatorischer Begriffsbildung zwischen Syntax und Morphologie*. Tübingen: Francke Verlag.

SCHMITT, P. A., 2006. *Translation und Technik*. Tübingen: Stauffenburg.

SOFER, M., 2006. *The Translator's Handbook*. 6th edition. Rockville USA: Schreiber Publishing.

COMPUTERISATION OF PHRASE-TO-PHRASE MATCHING FROM A STANDARD MARINE COMMUNICATION PHRASES CORPUS: A PRELIMINARY EMPIRICAL STUDY

María Araceli Losey León
Departamento de Filología
Francesa e Inglesa
Universidad de Cádiz
araceli.losey@uca.es

Abstract

Standard Marine Communication Phrases (IMO, 2002), formerly Standard Marine Navigational Vocabulary (IMO, 1985), are a collection of phrases conceived by the International Maritime Organization as a restricted language in the specialized setting of maritime communications to enhance maritime safety by avoiding language misunderstandings at sea. These standardized phrases were created for use over VHF radio in bridge external communications and land-based station exchanges or face to face on board communications. Its prominent characteristics include the lexical and grammatical restrictions (controlled vocabulary, modality limitations) and the controlled discourse use (sender and receiver's identification, pattern repetition, message markers, distress, urgency and safety procedural patterns, broadcast entries) they are submitted to. The standardized language restrictions in the vocabulary and phrase structure selection were designed taking into account predictable areas of language confusion and error. Availability of training material suited to the peculiar pedagogy that a controlled language involves is of utmost importance. In this context, several studies have been made (Strevens & Johnson, 1983; Losey, 2000; Pritchard & Kalogjera, 2000; Cole, Pritchard & Trenkner, 2007; CAPTAINS, 2012). However, there is still a lack of computerised corpus-based material oriented towards the specific gradual practice of the SMCP through its phraseology in semantic scenarios that may train and prepare the learner (or user) for building up predictable questions-to-answers about situations that may emerge at sea.

Within the NLP framework, the present corpus-based study attempts to develop an application for automatic generation of SMCP message phrase replies based on the Question Answering (QG) system for restricted domains (RDQA) (Mollá, Vicedo, 2007). Findings revealed during this preliminary empirical study shall be presented and examined. Finally, it is expected that this approach, which to this author's knowledge has not been explored so far, may contribute to SMCP training by e-learning, blended, face-to-face courses or Intelligent Tutoring Systems (ITS).

1. INTRODUCTION

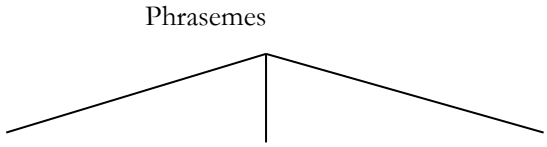
Phraseology represents one of the most outstanding phenomena in specialized communication and largely contributes to build up the conceptual tissue that shapes the oral and written discourse of a specialized domain. Broadly speaking, it has traditionally been associated with fixed expressions, set phrases, idioms, collocations, multi-word units, fixed lexical chunks and multi-word expressions. Scholars in various fields have coined their own phraseology-terms and provided descriptive studies about what should be conceived as 'unidades fraseológicas' or phraseological units (Corpas, 1996), phraseological expressions (Schmidlin, 2007), terminological phrasemes (Meyer & Mackintosh, 1996), phraseologism (Gries, 2008), polilexical terminological units (Cabré, Lorente and Estopà[5], 1996; Bevilacqua, 2001), or set phrases (Colson, 2008), to name but a few. This situation reveals that there is a wide range of overlapping or contrasting definitions of what phraseology encompasses in different fields of enquiry and that there is no consensus among scholars as to the precise contents or inclusions of phraseology that may help to draw a strict line. Despite differing terminology and lack of a clear-cut definition, it undeniably is a central notion and asset in our field of research, corpus linguistics. In this context, Corpas (1996) provides a fine-grained concept of phraseology setting up the criteria of number of words, extension, (co)occurrence frequency, institutional language sharedness, and fixedness and variation degrees (269). Gries (2008) uses the term 'phraseologism' to refer to 'the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance' (6). Besides, he further identifies and exemplifies different kinds of phraseologism on the basis of six criteria, namely nature of the elements, number of elements, frequency of occurrence, distance of the elements, flexibility of the elements and semantic unity (4-5). Hunston (2011) defines phraseology as 'a very general term used to describe the tendency of words, and groups of words, to occur more frequently in some environments than in others'.

Regarding the specialized language domains context, a great deal of research has been conducted in collocations, multi-word units, multi-word expressions and multi-word patterns but few studies have closely examined the phraseology as a theoretical construct. At this point, Pavel (1994) states four selection criteria for LSP phraseologisms which include relevance, degree of specialization, degree of frozenness and usefulness for the target public. Cabré, Lorente and Estopá (1996) present a classification of specialized phraseological unit as a subcategory of phraseological unit which, in turn, stems from a typology of polilexematic units. Bevilacqua (2001) provides a major input in this field through her study about specialized event phraseological unit (UFE) as governed by four criteria: the UFE includes at least a simple or syntagmatic terminological unit; the UFE includes at least one eventive element from which the UFE is organised semantically; the UFE has a certain degree of fixedness that depends on the semantic relationship between the elements; the UFE should have a relevant degree of frequency in the texts (114). On the other hand, in the terminological context, one of the most influential proposals is Meyer and Mackintosh's (1996) terminological phraseme. In this line, Granger and Paquot (2008) are also engaged in phrasemes typology and offer a corpus-based proposal focused on three major categories: referential phrasemes, textual phrasemes and communicative phrasemes (42).

[5] This study provides a further distinction between polilexematic units and specialized phraseological units within the context of the communicative theory of Terminology/ 'Teoría Comunicativa de la Terminología' (TCT) (Cabré, 1998).

The object of our study is the phraseology of a restricted sublanguage in Maritime English radio communications specialized domain known as *Standard Marine Communication Phrases* (*SMCP* onwards) and based on the SMCP-EXT corpus. It attempts to address the issue of *SMCP* computerisation in order to develop a pedagogical training application tool in the field of the Restricted Domain Question Answering (RDQA) system within a corpus-based discursive-pragmatic approach where phraseology is involved in the context of language technologies[6]. To the best of our knowledge, no previous study has accomplished this task in our setting and from this approach.

In our view, SMCP poses special phraseological characteristics that may lead us to consider each phrase as a fixed meaningful block unit and as a matter of phraseological concern. Every phrase can comply with the phraseological criteria, namely extension, fixedness, frequency of (co)occurrence and semantic unity. They are also restricted in commutative properties. On the other hand, they become fixed expressions because there was a need of a community of speakers for a restricted, short and simple way to communicate so that each phrase fulfils a specific role in communication. Interestingly, the pragmatics of their speech acts reveals it requires the use of procedural patterns for the SMCP delivery and this is not required in the realm of speech act pragmatics of unrestricted and specialized natural languages. For these reasons, the phraseological theoretical framework that best fits our SMCP construct is Granger and Paquot's adaptation of Burger's classification (1998) where communicative phrasemes are involved as it is illustrated in figure 1. In this study we support the view that SMCP may well be enlarging the list of the communicative phrasemes.



Referential function	**Textual function**	**Communicative function**
Referential phrasemes	Textual phrasemes	Communicative phrasemes
(Lexical) collocations	Complex prepositions	Speech act formulae
Idioms	Complex conjunctions	Attitudinal formulae
Irreversible bi- and		(including attitudinal sentence stems)

Figure 1. The phraseological spectrum (Granger & Paquot, 2008: 42)

[6] In this context, as Corpas (2013) mentions, either 'computational phraseology' or 'processing of multi-word expressions' are applicable terms though the latter is preferably used in the field of NLP (342).

2. RESEARCH CONTEXT, AIMS AND METHODOLOGY

Within the corpus linguistics and computational linguistics field, this research is based on the controlled language of seafaring for ship-to-ship, ship-to-land-based station or ship-to-aircraft VHF communications known as the *Standard Marine Communication Phrases* (2002) in its external communications section. SMCP is a substantial content of the maritime English curriculum and plays a decisive role for junior deck, engines and marine radioelectronics officers and ratings during their academic and professional training. The International Maritime Organization promotes its mandatory use in training schemes for both officers and ratings and a training material consistent with the peculiar pedagogy that a restricted sublanguage of a specialized domain involves is of utmost importance.

In this line, some research work has been focused on the communicative approach for the development of learning strategies and tools suitable to SMCP communicative restrictive nature (Losey, 1995; Trenkner, 1997; Losey, 2000). At present there are some publications, online resources and computer-based tools oriented to the learning of SMCP terms and phrase structures (Carrasco, 2011; John & Gregorič, 2014[7]), its testing (TOMEC or Test of Maritime English Competencies)[8] or training and self-assessment such as *SMCP Training Tools* (MarineSoft, 2013)[9]. Besides, a recent mainstream of Maritime English projects have led to enhanced learning and/or testing material evolved within a vocation-specific context where SMCP also plays a significant role though not exclusively –MarEng (2006-2010), TUMSAT (2007), MarTEL (2008-2012), CAPTAINS (2010-2012), INTERMAR (2012-2014), SeaTALK (2013-2015)[10]-. However, in our view, there is still a lack of resource tools that get closer to an integral gradual SMCP learning by inserting the phrases in contexts so that they may enhance the learner's task in knowledge integration by the association of questions-to-answers in a proper semantic scenario of virtual, blended or face-to-face courses.

The main aim of this empirical corpus-based study is to provide the overall description of the discursive-pragmatic basis and principles underlying the computerisation of SMCP for a pedagogic resource application, *SMCP-QA*, that may help students' acquisition of a non generative language, restricted in syntax patterns, terms, and grammar. This programme is intended to interact with the user and shall be grounded on the RDQA system (Restricted Domain Question Answering) whose potential contributions in this work will be explored likewise. *SMCP* is an essential part of the academic curricular programme of Maritime English courses for navigation, marine engineering and marine radioelectronics students in Maritime colleges and universities.

The methodology at this exploratory stage of the study is premised on the SMCP analysis to explore all the possible phrase-to-phrase matching that will result into matching rules to associate questions to answers. They will be generated according to the message pattern and its communicative function in combination with the requirements of the Restricted Domain Question Answering (RDQA). In order to prepare the SMCP associative Q-A data a taxonomy of the SMCP communicative functions shall be provided so that it will help to determine the answer candidates expected types (AM) and their

[7] Available at http://www.smcpexamples.com/

[8] Available at http://www.benntec.de/marinesoft/referenzen/tome-test-of-maritime-english/?lang=en

[9] Available at: https://www.yumpu.com/en/document/view/13739873/specification-smcp-training-tools-marinesoft-entwicklungs

[10] The corresponding websites for these projects are: MarEng http://mareng.utu.fi/; TUMSAT http://www2.kaiyodai.ac.jp/~takagi/mei/english/index.html; MarTEL http://www.martel.pro; CAPTAINS http://www.captains.pro/ ; INTERMAR http://www.intermar.ax/; SeaTALK http://www.seatalk.pro/.

individual association to the QM; at the same time, it will also contribute to define the user's input nature. Besides, as part of the research methodology purposes, an inventory containing the most outstanding lexical collocations has been created on the basis of retrieved data from the SMCP-EXT corpus as they will be stored in the system in order to support the refinement of the query. In this study, each individual phrase shall be named "phrase-message" (PM) as per its active communicative function nature; Question-message (QM) will refer to every phrase in the SMCP that is expected to receive a reply and Answer-message (AM) will denote the corresponding reply as to SMCP rules.

The rest of this study is organised as follows: Section 3 provides a SMCP sublanguage definition, a brief outline of its structural organization and of the discursive-pragmatic features of SMCP that shall be at the forefront on devising the system architecture. Next, section 4 covers a description of the SMCP-EXT corpus and of its processing preparation method altogether. Section 5 describes and reasons the RDQA system potential in our work. Finally, section 6 gathers the main conclusions on the subject and sums up open issues and ongoing research.

This research uses the linguistic search engines Wordsmith tool 6.0 and a POS tagger, CLAWS 4 POS tagger for empirically approach the research questions and at present we are testing some domain specific modeling tools for the formalization and representation of the phrase-matching logical rules.

3. THE STANDARD MARINE COMMUNICATION PHRASES

The Standard Marine Communication Phrases[11] are defined in the present research as a restricted sublanguage in the maritime specialized domain that covers major safety related issues for external verbal communication -ship to ship, ship to shore personnel and/or land-based stations- and for onboard verbal communications which have been designed to improve navigational communication and contribute to the safe operations and management of ships, the navigational safety and the protection of human life.

The concept of sublanguage here used makes reference to a purposely created set of simple fixed grammar pattern phrases based on Maritime English in order to facilitate fluent spoken understanding within the sea navigation and transport specific discourse community.

The whole SMCP compendium comprises external and on-board communication sections:

(a) External communication phrases: This section includes the phrases intended to be used via VHF radiotelephony for ship-to-ship or ship-to-shore and viceversa communications.

(b) On-board communication phrases: They are phrases to be used either intra-ship via the ship internal voice systems (for example, between the bridge and the engine-room or briefings to the passengers) or face-to-face/walkie-talkie communication (for bridge talk, pilot requests, engine-room checking status of equipment, for crew dealings with shore-side personnel, etc.).

[11] The full standard phrases are contained in the IMO (2002) *Standard Marine Communication Phrases*. London: IMO.

In our view, according to its communicative range and aim the external communication PMs can be subdivided into four general basic types:

3.1. SMCP communicative functions

According to its communicative range and goal, the external communication PMs can be subdivided into five basic types:

(1) Distress communications[12]: It contains PMs to be used when being threatened by serious and/or imminent danger and of requiring immediate assistance.

(2) Urgency communications: It contains PMs to be used when dealing with a condition concerning the security of a ship, aircraft or other vehicle, or of some person on board or within sight, but which does not require immediate assistance.

(3) Safety communications: It contains PMs to be used in conditions which necessitate the transmission of a message concerning the safety of navigation or providing important meteorological warnings.

(4) VTS communications: It contains PMs for use by the VTS operators or bridge when the vessel is passing congested areas, such as channels or shipping lanes or just for inbound and outbound vessels in ports or harbours. The VTS phrases are devised to aid the safe movement of sea traffic, especially in congested areas, in port and harbour approaches, channels, and shipping lanes. It includes special communications for pilot exchanges and tug requests.

(5) Broadcast service communications: navigational and weather information mainly.

Each of them will be further explored in order to refine and extract its fundamental communicative functions as they will form the knowledge base module for *SMCP-QA* system.

3.2. Discursive-pragmatic features of SMCP

An in-depth study addressing the design of adequate SMCP language learning resources should be preceded by a previous analysis of the main traits of SMCP discursive-pragmatic functions as the focus is on interactional issues (Schneider & Barron, 2014). Examples of early studies in this vein are Losey (1993), Johnson (1994), Trenkner (1997), Pritchard & Kalogjera (2000), to mention but a few. Weeks et al. (1984), the authors of *Seaspeak,*[13] point out that the standardized language was designed 'to reduce possibilities of confusion, to maximize guessability and listeners expectancies, and to concentrate on effective communication' (v). On the other hand, as stated in the IMO SMCP introduction (2002), in contrast to full natural language, SMCP imposes restrictions concerning term selection, syntax, and further communicative features such as 'avoiding synonyms, avoiding contracted forms, providing fully worded answers to 'yes/no' questions and basic alternative answers to sentence questions, providing *one* phrase for *one* event, and structuring the corresponding phrases after the principle: *identical invariable plus variable'* (3). SMCP shows a high degree of narrowness at different linguistic levels as well as formal restrictions to generate or reformulate new phrases, recurrence of a closed loop of syntactic patterns, recurrence of

[12] SMCP handbook (2002) includes the message procedures (106-11) in compliance with the International Telecommunications Union Radio Regulations (2012: 320, 331, 335).

[13] *Seaspeak* (Weeks et al., 1984) is a reference manual containing recommendations for VHF maritime communications concerning VHF procedures, distress, urgency and safety procedures, the exchange procedure, VHF messages –a set of twenty standard short phrases for making and maintaining contact and for conversation controls-, and message markers.

specialized terms, not commutability of phrases, preference for simple phrases and restricted syntax pattern, and avoidance of subordinate sentences (an exception is the use of infinitive of purpose in some cases).

Due to size constraints, this paper does not present the detailed SMCP discursive pragmatic features analysis that has been developed in full. Thus, what follows is a brief summary of the main discursive-pragmatic features: Use of the international maritime alphabet, avoidance of polite formulae, coded call sign, coded caller identifier and addressee, restricted modality, turn-taking procedure or switch-over rules, time periods, question-to-answer correspondence in real time.

4. THE SMCP-EXT CORPUS

The SMC-EXT corpus consists of 14,730 tokens and 874 types were retrieved being the token/type ratio 6.89. It is a written monolingual tagged corpus composed by 1,403 phrase messages out of an overall amount of 1,822 phrases. There is a difference of 419 phrases. The reason is that our study considers that a phrase-message (PM) may consist of more than one utterance in the same SMCP delivery whereas the analyzer counts every individual sentence since the WS analysis was conducted using the 'stop at sentence break' set up. PM can be split into 1 or more phrases as it is illustrated in the samples provided in table 1.

Ref.	SMCP Phrase messages	SMCP individual phrases
A1/ 6.2.2.1.7	I have located you on my radar screen.	I have located you on my radar screen.
A1/ 6.2.1.3.1 2	Small fishing boats in sea area around Grand Sol. Navigate with caution.	Small fishing boats in sea around Grand Sol.
		Navigate with caution.
A1/ 6.2.1.7.2 .1	Tropical storm warning at 1200 hours UTC. Hurricane *Sarah* with central pressure of 980 millibars located in position 35º20'05"N065º15'18"W. Present movement NNW at 30 knots. Winds of 20 knots within radius of 30 nautical miles of centre. Seas over 8 metres. Further information on VHF channel 12 at 1200 hours UTC.	Tropical storm warning at 1200 hours UTC.
		Hurricane *Sarah* with central pressure of 980 millibars located in position 35º20'05"N065º15'18"W.
		Present movement NNW at 30 knots.
		Winds of 20 knots within radius of 30 nautical miles of centre.
		Seas over 8 metres.
		Further information on VHF channel 12 at 1200 hours UTC.

Table 1. SMCP Phrase messages (PMs) vs. SMCP individual phrases

This analysis was performed with Wordsmith Tool 6.0 (Scott, 2012) and the tagger used was a POS tagger or grammatical tagger called CLAWS4 (UCREL) -output obtained in the C7 tagset- as it is shown in figure 2 below.

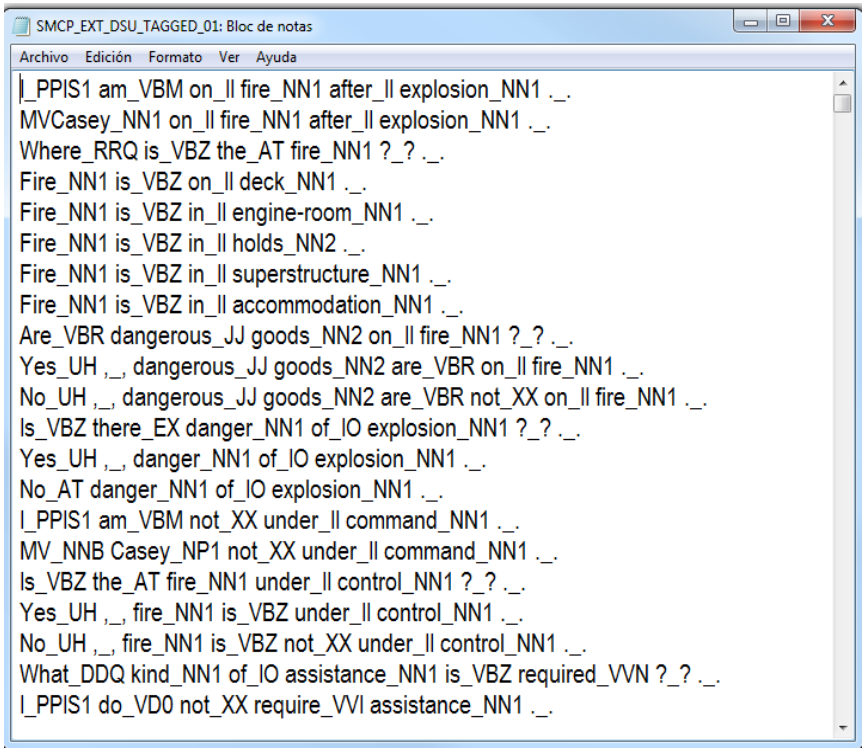


Figure 2. A sample of SMCP-EXT tagged by CLAWS4

This corpus-based study also gives account of the lexical collocational patterns because these data which have been collected in an inventory shall be transferred as input data for the general knowledge base of *SMCP-QA* tool. To this end, an inventory of the SMCP collocations was developed and it was observed that the most frequent collocational pattern was V+NP. This is the case for the search for *require** concordance which resulted in a total amount of 106 items.

The compilation process was easy at hand because we compiled ready-made phrases from the *SMCP IMO* handbook but a number of decisions were necessarily taken during the pre-processing stage concerning its layout and the way in which the phrases' variables were to be included in the corpus under homogeneous criteria. The preparation has been done manually and the corpus includes all the possible phrases to be generated with free information. The full set of criteria applied is displayed in figure 3.

5. RESTRICTED DOMAIN QUESTION ANSWERING SYSTEM

As reported by Skalban et al. (2012), QG has frequently been employed in educational contexts as its applications include the automatic creation of multiple-choice questions tests (MCQ), vocabulary exercises and solutions which promote reading comprehension (1152). Mollá and Vicedo (2007) define QA as 'the task whereby an automated machine (such as a computer) answers arbitrary questions formulated in natural language' (41). The extension of this system to restricted domains is known as RDQA and its task is to find the text that contains the answer to the question and extract it. On the other hand, the goal of a question answering system is to retrieve answers to questions rather than full documents or best-matching passages. Given that one of the main reasons for the development of *SMCP-*

QA tool was to provide for the association of QM to AM, and that the data into play is a set of phrases and not full documents, RDQA seems to suit our needs.

RDQA system consists of three phases or modules for the modeling of the domain information, in this case of the SMCP.

(a) Question classification: This phase classifies the SMCP-QA's user question, derives expected answer types and extract keywords which shall be used as domain knowledge selection for PM's association. As Mollá & Vicedo state, RDQA also requires 'the definition of an internal representational model that allows the integration or combination of the different information sources available for the domain' (Mollá & Vicedo, 2007: 54).

(b) Information retrieval: By this means it is possible to select and retrieve applicable candidate answers.

(c) Answer extraction: This module is in charge of finding and extracting the expected answer type (EAT) by using natural language processing tools (such as POS tagger, syntactical parser, entity annotator, semantic role parser, etc.) to analyse this set of passages.

The user writes a question by means of the user query interface. This query is used to extract all the possible answers for the input question.

6. CONCLUSIONS

This corpus-based phraseology empirical study on the Standard Marine Communication Phrases sublanguage for maritime speech communications has been based on a discursive-pragmatic approach. The results obtained from a corpus of structured documents have contributed to shape a knowledge database in terms of communicative functions of language which shall be applied in a RDQA (Restricted Domain Question Answering) system for the implementation of the *SMCP-QA* resource. It has been proved that the RDQA system may be suited to our research aims as it provides concise answers from a collection of documents (SMCP-EXT PMs) to questions (QM) stated by the user. The answers are expected to be provided to the own questions of the system through the expected answer type (EAT) so that correct feedback and evaluation provided emerge from the adequate matching of the question phrase message (Q-PM) and the reply phrase message (R-PM) that shall be executed by setting logical rules. For the SMCP-QAM system, the question classification operates according to a set of categories based on an analysis of the prominent SMCP communicative functions. At present we are in the implementation phase of the computer programme and working on the logical rules form. Further issues of our concern are (a) the design of the user input taxonomy. User input is defined as all cases of user inputs to the QA system, such as all questions, dialogue sentences, keywords for information search; (b) the design of an advanced metamodel for extra data and (c) extend our research to the SMCP on-board communication phrases.

References

BEVILACQUA, C.R., 2001. Unidades fraseológicas especializadas: elementos para su identificación y descripción. In: M.T. Cabré, J. Feliu, ed. 2001. *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica.* Barcelona: Institut Universitari de Lingüística Aplicada. pp. 113-141.

CABRÉ, M.T., LORENTE, M., ESTOPÀ, R., 1996. Terminología y fraseología. In: RiTERM, ed. 1996. *Actas del V Simposio de Terminología Iberoamericana.* Ciudad de México: Colegio de México. pp. 67-81.

CAPTAINS, 2012. *The Captains English Learning Tool. Standalone Course.* [online] Available at: http://www.captains.pro [Accessed 10 December 2012].

CARRASCO CABRERA, M.J., 2011. *Inglés Técnico Marítimo para titulaciones náuticas profesionales y capitán de yate.* La Coruña: Netbiblio.

COLE, C., PRITCHARD, B., TRENKNER, P., 2007. Maritime English Instruction- ensuring instructors' competence. *Ibérica*, 14, p. 123-148.

COLSON, J.-P., 2008. Cross-linguistic phraseological studies: An overview. In: S. Granger, F. Meunier, ed. 2008. *Phraseology: An interdisciplinary perspective.* Amsterdam/Philadephia: John Benjamins Publishing Company. pp. 191-206.

CORPAS PASTOR, G., 1996. *Manual de Fraseología Española.* Madrid: Gredos.

CORPAS PASTOR, G., 2013. Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In: I. Olza, E. Manero, eds. 2013. *Fraseopragmática.* Berlín: Frank & Timme GmbH. pp. 335-373.

GRANGER, S., PAQUOT, M., 2008. Disentangling the phraseological web. In: S. Granger, F. Meunier, ed. 2008. *Phraseology: An interdisciplinary perspective.* Amsterdam: John Benjamins Publishing Company. pp.27-49.

GRIES, S.T., 2008. Phraseology and linguistic theory: A brief survey. In: S. Granger, F. Meunier, ed. 2008. *Phraseology: An interdisciplinary perspective.* Amsterdam: John Benjamins Publishing Company. pp.3-25.

HUNSTON, S., 2011. *Corpus Approaches to Evaluation. Phraseology and Evaluative Language.* New York/ London: Routledge.

I.M.O., 2002. *Standard Marine Communication Phrases.* London: I.M.O.

I.M.O., 1985. *Standard Marine Navigational Vocabulary.* London: I.M.O.

I.T.U., 2012. *Radio regulations. Articles. Library and Archive Services.* [online] Available at: http://handle.itu.int/11.1004/020.1000/1.41 [Accessed 25 January 2015].

JOHN, P., GREGORIČ, T., 2014. *SMCP examples.* [online] Available at: http://www.smcpexamples.com [Accessed 12 May 2015].

JOHNSON, B., 1994. English in maritime radiotelephony. *World Englishes,* 13(1), p. 83-91.

LIU, M., CALVO, R.A., RUS, V., 2014. Automatic generation and ranking of questions for critical review. *Educational Technology & Society* 17 (2), p. 333-346, [online] Available at: http://www.ifets.info/journals/17_2/27 [Accessed 15 January 2015].

LOSEY LEÓN, M.A., 1993. SEASPEAK: Función estilística del inglés específico para fines marítimos. In: Barrueco, S.; Hernández, E., Sánchez, M.J., Sierra, L. eds. 1993. *Actas de las II Jornadas de Lenguas para Fines Específicos.* Alcalá de Henares: Servicio de Publicaciones. Universidad de Alcalá de Henares. pp. 101-107.

LOSEY LEÓN, M.A., 1995. On the Right Course!: Propuesta de actividades para el laboratorio de idiomas y el ordenador. In: Barrueco, S.; Hernández, E., Sierra, L. eds. 1995. *Lenguas para fines específicos IV. Investigación y enseñanza.* Alcalá de Henares: Servicio de Publicaciones. Universidad de Alcalá de Henares. pp. 289-299.

LOSEY LEÓN, M.A., 2000. Facing new changes in the Maritime English curriculum: Tasks' design towards the acquisition of the Standard Marine Communication Phrases. In: Piniella, F. et al. eds. 2000. *Proceedings of the 2nd. International Congress on Maritime Technological Innovations and Research.* Cádiz: Servicio de publicaciones de la Universidad de Cádiz. pp. 123-129.

MEYER, I., MACKINTOSH, K., 1996. Refining the terminographer's concept-analysis methods: How can phraseology help?. *Terminology*, 3(1), p. 1-26.

MOLLÁ, D., GARDINER, M. 2004. AnswerFinder: Question Answering by Combining Lexical, Syntactic and Semantic Information. *Proceedings of the Australasian Language Technology Workshop.* pp. 9-16.

MOLLÁ, D., VICEDO, J.L., 2007. Question answering in restricted domains: An overview. *Computational Linguistics,* 33(1), p. 41-61.

SCHNEIDER, K.P., BARRON, A., 2014. *Pragmatics of Discourse.* Berlin, New York: Walter de Gruyter.

SCHMIDLIN, R., 2007. Phraseological expressions in German standard varieties. In: H. Burger, D. Dobrovol'ski, P. Kühn, N. R. Norrick ed. 2007. *Phraseologie. Phraseology.* Berlin, New York: Walter de Gruyter. pp. 551-562.

SCOTT, M., 2012. *Wordsmith Tools version 6*. Liverpool: Lexical Analysis Software.

PAVEL, S., 1994. *Guide to Phraseology Research in Languages for Special Purposes.* Montreal: Terminology and Documentation Directorate. Translation Bureau.

PRITCHARD, B., KALOGJERA, D., 2000. On some features of conversation in maritime VHF communication. In: M. Coulthard, J. Cotterill, F. Rock, ed. 2000. *Dialogue Analysis VII: Working with Dialogue.* Tübingen: Max Niemeyer Verlag. pp. 185-196.

RINALDI, F., DOWDALL, J., HESS, M., MOLLÁ, D., SCHWITTER, R., 2002. Towards Answer Extraction: An Application to Technical Domains. *Proceedings of the 15th. European Conference on Artificial Intelligence,* p. 1-5.

SKALBAN, Y., HA, L.A., SPECIA, L., MITKOV, R., 2012. Automatic question generation in multimedia-based learning. *Proceedings of COLING 2012: Posters*, p. 1151-1160, [online] Available at: http://aclweb.org/anthology/C12-2112 [Accessed 19 December 2014].

TRENKNER, P., 1997. The IMO Standard Marine Communication Phrases. *Proceedings of the 9th. Workshop on Maritime English (WOME 9)*. Sweden: World Maritime University, p. 1-9.

WEEKS, F., GLOVER, A., STREVENS, P., JOHNSON, E., 1984. *Seaspeak Reference Manual. Essential English for International Maritime Use*. Oxford: Pergamon Press.

ANNOTATION OF MULTIWORD EXPRESSIONS IN FRENCH

Agnès Tutin
LIDILEM
Université Grenoble Alpes
agnes.tutin@u-grenoble3.fr

Emmanuelle Esperança-Rodier
LIG
Université Grenoble Alpes
emmanuelle.esperanca-rodier@imag.fr

Manolo Iborra
Université Grenoble Alpes
iborra.manolo@gmail.com

Justine Reverdy
Université Grenoble Alpes
Justine Reverdy
@e.u-grenoble3.fr>

Abstract

This paper presents an experiment of annotation of MWEs in French. The corpus used is made of several genres (news, novel, scientific report, film subtitles) and includes a rich annotation scheme including several kinds of MWEs from collocations to routines and full phrasemes. The annotation is performed semi-automatically with finite-state transducers. The inter-annotator agreement score shows that the annotation is quite consistent but the difficulty of the task relies heavily on the textual genre: literary texts are harder to annotate than scientific reports. Besides, two types of categories are difficult to differentiate, collocations and full phrasemes.

1. INTRODUCTION

This paper presents an experiment of multiword expression annotation on the French part of a French-English bilingual corpus. Our aim is to achieve three goals: a) building a corpus-based and robust typology of MWEs; b) providing a basis for linguistic studies on MWEs, especially in relation to diverse textual genres; c) building a corpus of evaluation for Machine Translation (MT) tasks, and especially statistical machine translation (SMT) tasks (e.g. Potet *et al.* 2012).

Every scholar working on MWEs knows that defining clearly different types of MWEs is a complex task. But we think that confronting concrete examples will help to refine typologies of MWEs, and enable to better understand how they work.

This will also help to explore the most frequent MWEs, especially according to the specific genres, in order to answer questions such as the following ones:

- Are collocations really more frequent in general expressions than in idiomatic expressions?
- Are true idiomatic expressions, such as *to break the ice*, more frequent in spoken genres?
- Regarding syntax now, we would like to observe in more detail syntactic properties of MWEs. Are real MWEs highly variable, as suggested by Moon, or not?

Considering now practical goals, we know that there are few annotated corpora with MWEs, especially for French. There are two small corpora with nouns and MWE adverbs (Laporte *et al.* 2008a; Laporte *et al.* 2008b), but these corpora do not include any typology of expressions. The French Treebank (Abeillé *et al.* 2003) includes several kinds of MWEs including verbs, but only on contiguous MWEs such as *faire part* but no discontinuous expressions, e.g. *prendre ce problème en compte*. In English too, there are not many resources. One of the most interesting ones is undoubtedly Schneider *et al.* (2014) social web corpus with MWE annotations, which distinguishes between strong and weak MWEs, but does not include any fine-grained typology.

We obviously need reference corpora for evaluation purposes in MT and especially statistical MT applications, especially to evaluate which MWEs are more difficult to translate. Our hypothesis (which has been partially confirmed) is that contiguous expressions are easier to translate.

But the annotation of MWEs is not a straightforward task, mainly due to three types of problems:

1. Deciding whether an expression is a MWE expression or not. This is easy for compounds such as *as long as* but complex for collocations and routines.

2. Delimiting the expression is also complex: for example, do determiners belong or not to the expression? In our annotation scheme, we decided to include fixed determiners, e.g. *la* in *il fait la fête* (lit. 'he makes the party') but not in examples such as e*lle donne un cours* ('she gives a class') where we could have any determiner.

3. Deciding the type of the expression is also complex and needs clear-cut criteria.

We will first present the corpora and the annotation scheme, and also the semi-automatic annotation process. We will then present the quantitative and qualitative results of the evaluation of the annotation, including inter-annotator agreement.

2. CORPORA AND ANNOTATION SCHEME

Our project is to build a corpus of about 45,000 tokens, freely available, of the French part of a bilingual corpus. This paper only focuses on the annotation of the French part. The next step will be the bilingual alignment.

Our corpus includes several genres of texts, namely:

- a scientific writing (*BAF Citi 1* from Baf corpus, about 14,500 tokens),
- news (journalese, from several news corpora of the WMT evaluation campaign (from 2006 to 2010), almost 12,300 tokens),
- subtitles of a French film, *Amélie Poulain* (9,900 tokens), and
- an extract of Zola's *Thérèse Raquin* (7,260 tokens).

Several MWEs are included and a semi-automatic annotation using Nooj system has been carried out.

2.1. Typology of MWEs

In our annotation scheme, we include a typology of MWEs inspired by Granger & Paquot (2008), Ulrich Heid (2008), Mel'čuk (2013) and a previous work on this topic (Tutin 2010).

Following Mel'čuk, we distinguish « Full phrasemes » which are non-compositional expressions of nouns, adjectives, verbs and adverbs (including compounds) e.g. *dead end* or *to take into account* from collocations, which are compositional but are difficult to predict e.g. a *heavy smoker* in English while we have a *« big » smoker*, *gros fumeur* in French (e.g. Tutin & Grossmann, 2002).

Then, we have a specific category for functional words such as prepositions, conjunctions, determiners, pronouns and discourse and negation adverbs, e.g. *on the one hand, in front of, insofar as, a large number of*.

We also use the category of pragmatemes, which is very frequent in dialogues, for expression with a specific pragmatic function e.g. *You're welcome, See you later*.

Furthermore, we include complex terms, specific to a field, named entities e.g. *Université Stendhal, Laboratoire d'Informatique de Grenoble*, and routine formulae, e.g. *As previously said, force est de constater …* which are prefabricated verbal expressions frequent in a specific genre.

2.2. Annotation scheme

Our annotation scheme is a very simple surface annotation. We are aware that a stand-off annotation would be more suited for our purpose, but such an annotation scheme requires complex annotation tools and the annotation is not very convenient for linguists or people who are not familiar with NLP. Nevertheless, a stand-off annotation could be used at the end of the process.

The annotation scheme includes several features which are illustrated fig. 1 with the help of an example:

- An **identifier**, e.g. id="23" on each element of the MWE.
- The **type of MWE** (e.g. full phraseme, collocation, functional word…). In the example, *pris en compte* is a full phraseme.
- **Syntactic categories of the expression** (here the MWE is a verb) and of **the parts of the expression** (Verb + Prep + Noun).

```
Nous avons <epl id="23" type="fphraseme" catepl="verb" catw="verb" lemma="prendre_en_compte">pris</epl> ce problème <epl id="23" catw="prep">en</epl> <epl id="23" catw="noun">compte</epl>
```

Figure 1. Annotation of *Nous avons pris ce problème en compte*

Besides, overlapping expressions can be annotated. As we can see in the examples below, partial overlapping such as *pay close attention (pay attention + close attention)* as well as inclusions such as *au minimum* which is included in the collocation *réduire au minimum*, can be annotated with several identifiers, types and syntactic categories.

1) Unlike many theorists, he **[paid$_1$] [close$_2$] [attention$_{1+2}$]** to a broad range of experimental evidence…

2) Afin de **[réduire$_1$] [au$_{1+2}$] [minimum$_{1+2}$]** cet effort …

3. ANNOTATION PROCESS

The annotation process is a semi-automatic process. It uses a finite-state tool called Nooj, developed by M. Silberztein (2013), known to be very suited for the treatment of MWEs. Nooj uses a core dictionary of MWEs extracted from several resources:

- The most frequent MWEs from the French Treebank (Abeillé *et al.,* 2003), where MWEs are decomposed;
- Other dictionaries of MWEs such as the *Dictionnaire Electronique des Mots* (Dubois & Dubois Charlier, 2010), Wiktionary or the DELAC (Courtois *et al.* 1997).

Here is an example on Fig. 2 of the annotation process with the finite-state tool.

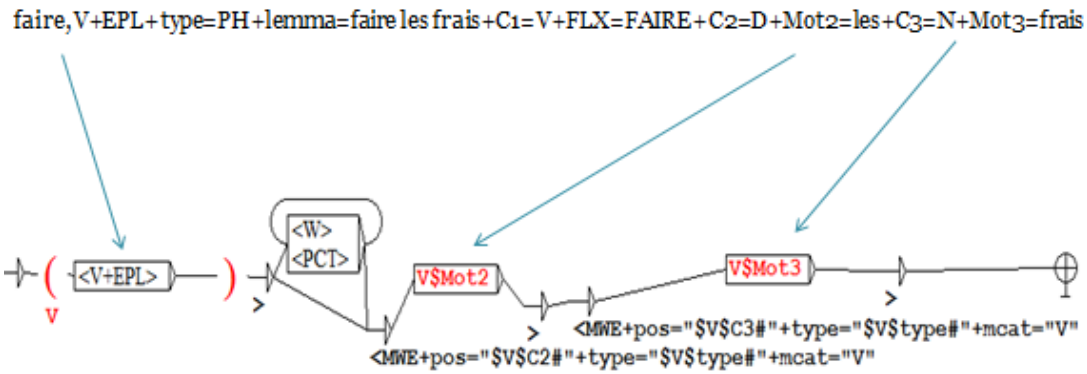


Figure 2. Annotation of MWEs with the NooJ processing system

In the dictionary, the MWE *faire les frais* is decomposed as follow:

- The elements of the expressions: for example, verb, determiner and noun for the expressions *faire les frais*.
- The type of MWE: here, it is a « full phraseme » since the meaning is not compositional.

Then the parts of the expression are associated to elements of a finite state graph, in order to enable the annotation in texts.

In order to achieve the annotation process, a core lexicon of about 5,000 MWEs is used. The semi-automatic annotation is thus performed and checked on concordances. It is then completed with manual annotation by two linguists with an XML editor (Oxygen). About 35% to 50% of MWEs, depending on the kind of text, are semi-automatically annotated.

The automatic annotation is quite good with functional words and frequent phrasemes. However, it has a weak coverage of collocations, pragmatemes and routines. In order to get better results, we could use a complex term extractor for technical terms and a better named entity recognition system.

In Fig. 3; we can see an example of the annotated expressions with a style sheet of subtitles of the film *Amélie Poulain*.

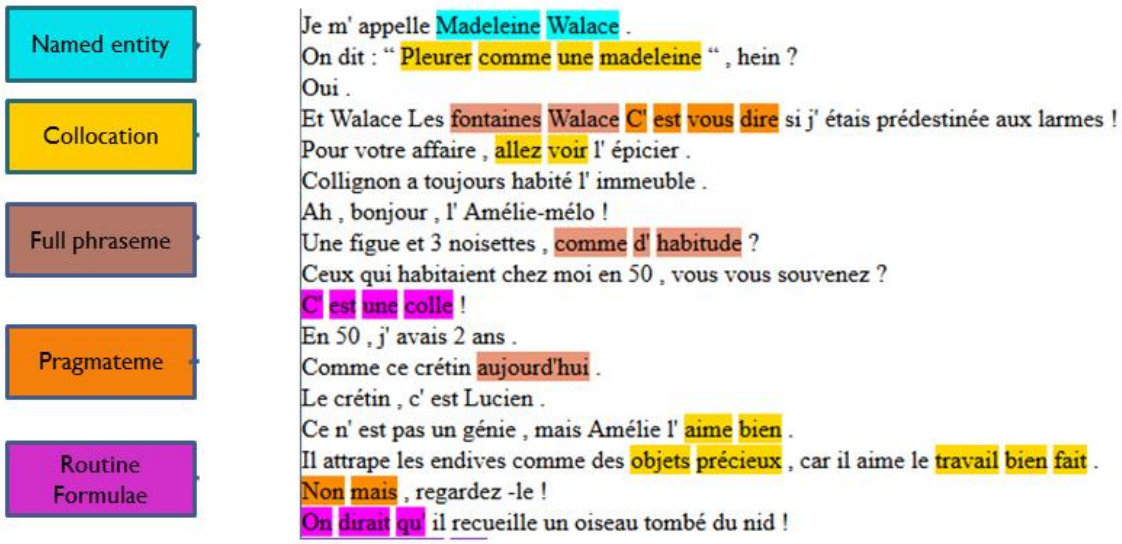


Figure 3. An example of annotation of MWEs: *Amélie Poulain*'s subtitles

Named entities with proper nouns are in light blue e.g. *Madeleine Walace*, full phrasemes e.g. *fontaines Walace* or *comme d'habitude* are in brown. In orange, we can find pragmatemes such as *non mais* or *C'est vous dire* which are typical of dialogues, in yellow, collocations e.g. *pleurer comme une Madeleine* or *aime bien* and finally in purple routines e.g. *C'est une colle !*.

Once the manual annotation performed by the two linguists according to the typology already described, we have evaluated the annotated expressions in texts.

4. QUANTITATIVE AND QUALITATIVE RESULTS

We now come to the evaluation of the annotation process. The objectives of this evaluation are to improve the typology by splitting or merging complex categories, or by adding criteria for the annotation process.

For that, an experiment was carried out on two different extracts of our corpus:

1) the literary text (*Thérèse Raquin*, a Zola's novel) around 2,000 words

2) the scientific report (CITI 1) around 2,000 words

Two annotators, two linguists familiar with the issue of MWEs, were involved in that task. We mainly focused on the type of MWE and not on grammatical issues (e.g. the grammatical category of the MWE).

4.1. Quantitative Results

The first parameter to examine is the proportion of MWEs in texts. On Fig. 4, we can see the percentage of words which belong to a MWE according to each annotator and common to both annotators.

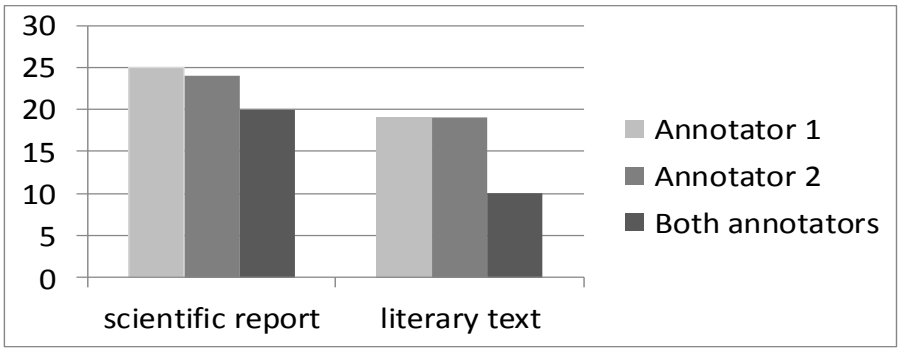


Figure 4. % of words belonging to a MWE

For the scientific report, we have more than 20% of the words from the text which are considered as belonging to a MWE and the agreement (in %) is good for this text. But for the literary text, we have less than 20% of the words and it falls to 10% for both annotators. Therefore, we observe more MWEs in the scientific text, probably due to the formulaic style of this genre, than in the literary extract, which was expected. We also notice that the agreement (in %) on what is a MWE is less good for the extract of the novel.

We observe the same tendencies on the agreement concerning the type of MWE. We computed the Fleiss Kappa score to compare the annotation of the type of MWE (only for MWEs selected by both annotators) (see. Fig 5).

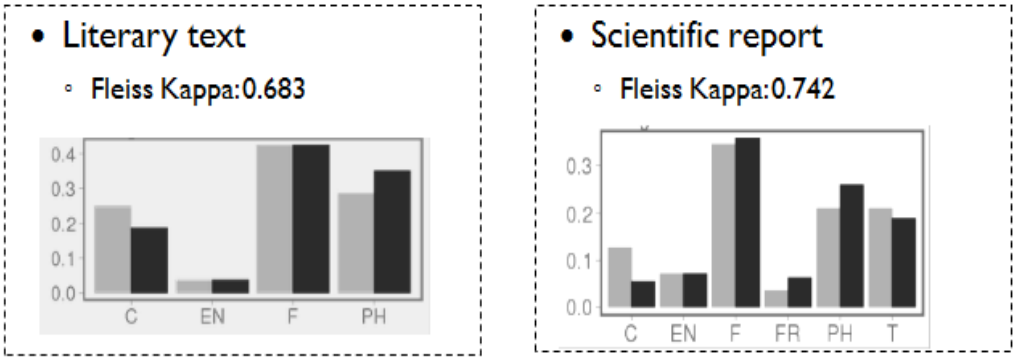


Figure 5. Inter-annotator agreement for types of MWEs

We can observe that the results are not bad: we have a kappa of 0.683 for the literary text and a kappa of 0.741 for the scientific report. There is a good agreement for named entities (EN) and functional words (F) but this is more complicated for collocations (F) and phrasemes (PH), which is not very surprising.

4.1. Quantitative Results

If we now have a look at the qualitative results, we can observe that, as expected, the different kinds of MWEs depend on the genres. This is quite obvious and expected. For example, we do not have any terms (T) or routine formulae (FR) in the literary text.

The disagreements between the two annotators often concern collocations or full phrasemes for several reasons.

Some collocations can have some specific syntactic properties, for example a lack of determiner, which is close from full phrasemes; e.g. in *tirer profit* ('take advantage') even though they are compositional.

A second problem occurs with « complex terms » which are very close to full phrasemes: the decision is determined by the genre of the text: is *emplois salariés* a term or a phraseme? Complex terms generally seem more compositional than phrasemes.

Binomials such as *jour et nuit* ('night and day'), *pur et simple* ('pure and simple') also raise specific problems: should they be considered as a type of collocation or a type of phraseme? This is not a very frequent type of MWEs and creating a specific category would seem artificial.

Finally, nominal hyperonymic collocations (non predicative collocations) such as *cuillère à soupe* ('tablespoon'), *boîte à outils* ('toolbox') also raise interesting problems. They are partially compositional but refer to a single referent and have an intermediate status.

These problems should be examined thoroughly in order to refine, with the help of clear-cut examples, the annotation guidelines.

5. CONCLUSION AND PERSPECTIVES

Annotating MWEs is a stimulating task (it is a deep reflection about MWEs) but this is also a feasible task, even though this is a complex one. The complexity of the task depends on the genre of the text and on the types of MWEs. This is especially difficult for some categories for which we need to develop more detailed criteria. For collocations and phrasemes, the issue of compositionality should be considered more in detail with analyze of complex examples described in the guidelines.

We also saw that this kind of annotation is complex and it seems necessary to carry out a double annotation, and with more annotators in case of disagreement. Therefore the annotation task requires the use of dedicated tools that allow collaborative annotation in order to maintain a good inter-annotator agreement. To our knowledge, such tools are not available yet, and one of our future works will be to characterize specifications for implementing an annotation tool. In addition, the automatic annotation needs to be developed, and this can be done incrementally with annotated corpora.

Observation on MWEs in real texts, especially on challenging cases such as routines or atypical collocations, is necessary to make progress in the field of phraseology and to propose well-suited encoding schemes.

References

ABEILLÉ, A., CLÉMENT, L. AND L. TOUSSENEL, 2003. Building a treebank for French. In: Treebanks. Springer Netherlands. pp. 165-187.

COURTOIS, B., GARRIGUES, M., GROSS, G., GROSS, M., JUNG, M., MATHIEU-COLAS, M., MONCEAUX, A., PONCET-MONTANGE, SILBERZTEIN, M. AND VIVÈS, R., 1997. Dictionnaire électronique DELAC : les mots composés binaires. Technical Report 56, University Paris 7, LADL.

DUBOIS, J. AND DUBOIS-CHARLIER, F. 2010. La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. *Langages*, 179(3, pp 31-56.

GRANGER, S. AND PAQUOT, M. 2008. Disentangling the phraseological web. In Granger, S. & Meunier, F. *Phraseology: An interdisciplinary perspective.* Amsterdam/Philadelphia: John Benjamins.

HEID, U. 2008. Computational phraseology. An overview. In: S. Granger and F. Meunier, *Phraseology. An interdisciplinary perspective*. Amsterdam: Benjamins. pp 337-360.

LAPORTE, E., NAKAMURA, T., AND VOYATZI, S. 2008a. A French corpus annotated for multiword nouns. *In Language Resources and Evaluation Conference*. Workshop Towards a Shared Task on Multiword Expressions. pp. 27-30.

LAPORTE, E., NAKAMURA, T., AND VOYATZI,. 2008b. A French corpus annotated for multiword expressions with adverbial function. In *Language Resources and Evaluation Conference (LREC)*. Linguistic Annotation Workshop. pp. 48-51.

MEL'ČUK, I. 2013. Tout ce que nous voulions savoir sur les phrasèmes, mais... *Cahiers de lexicologie. Revue internationale de lexicologie et de lexicographie*, 102, pp. 129-149.

MOON, R. 1998. *Fixed expressions and idioms in English: A corpus-based approach.* Oxford:Oxford University Press.

POTET, M., ESPERANÇA-RODIER, E., BESACIER, L., BLANCHON, H. 2012. Collection of a Large Database of French-English SMT Output Corrections, *(LREC 2012)*, Istambul, 2012, 21-27 mai.

SCHNEIDER, N., ONUFFER, S., KAZOUR, N., DANCHIK, E., MORDOWANEC, M. T., CONRAD, H., AND SMITH, N. A. 2014. Comprehensive annotation of multiword expressions in a social web corpus. *Proc. of LREC*. Reykjavík, Iceland.

SILBERZTEIN, M., KHURSHUDIAN, V., AND DONABÉDIAN, A., 2013. *Formalizing Natural Languages with Noo.* Cambridge: Cambridge Scholar Press.

TUTIN, A. 2010. *Sens et combinatoire lexicale: de la langue au discours*. Unpublished Dossier en vue de l'habilitation à diriger des recherches). Grenoble: Université Stendhal.

TUTIN, A., AND GROSSMANN, F. 2002. Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée*, *7*(1), pp. 7-25.

RESTRICTED COLLOCABILITY AND ITS USE IN ARABIC CORPUS LINGUISTICS

Petr Zemánek
Institute of Comparative Linguistics
Charles University in Prague
petr.zemanek@ff.cuni.cz

Jiří Milička
Institute of Comparative Linguistics
Charles University in Prague
milicka@centrum.cz

Abstract

Restricted collocability has received some attention, but not as a formalized method. We suggest that it should be used as a metrics for collocations, as well as for other types of usage, both in linguistics and even outside it, as it has great potentials for a plethora of applications. On the examples from a diachronic corpus of Arabic, we show the possibilities of its employment in studying prepositional valency and lexical profiling.

1. INTRODUCTION

It is generally accepted that in every language, some words have restricted collocability and some of those in the extreme. These restrictions are manifested in the ratios of the occurrences of individual items with the frequency of the whole string.

Such restrictions can often be observed at terms. E.g., for English, the British National Corpus (BNC) reports for such collocations as "pulmonary embolism" the following results: 15 occurrences for the whole expression, whereas 21 for "embolism", i.e. 71% of the instances of the word "embolism" occur in the above mentioned phrase. Such examples, on one hand, confirm the position of many who put such phenomenon on the periphery of the language system.

On the other hand, the occurrence of such collocations can be considerably frequent, such as "look forward to", where the intersection of the whole and "look forward" represents almost 92% (BNC search: 1095/1005), thus exhibiting a very strong relation between the two parts. This would, however, contradict the thesis of a peripheral phenomenon; such relation can be observed at many frequently occurring words, as shows our phrase, and it is rather easy to find similar examples.

The usage of the phenomenon in current linguistic research seems to be somewhat limited. Although the description of these multi-word units can be found in lexicographical textbooks (e.g., Granger and Meunier 2008), the usage so far reported in literature is rather restricted to purposes of studying lexicalization (e.g., Barkema 1996 in connection with

terminology), they are part of studies on language teaching (e.g. Howarth 1998) or are used in enriching existing databases with phrases (for WordNet, Bentivogli and Pianta 2003). They are, among other fields, an issue in translatology (e.g., Hansen et al. 2001). All of the above perceive such collocations as semantic units (semantic cohesion criterion). A pioneering contribution that deviates from this mainstream is Čermák 2006, where the phenomenon is used for mapping prepositional valency, or Bentivogli and Pianta 2004 who suggest to enrich the lexical databases with syntactic information based on restricted collocations.

2. THE METHOD

Dozens of collocability metrics have been proposed (Oakes 1998, pp 158–192). This is rather natural, since there is no general agreement what collocability is, and it is usually defined by the formula of the metrics that is being employed, which may result in somewhat circular definition. It should be, however, mentioned that some metrics gained more respect than others.

One of the most popular metrics is the Pointwise Mutual Information (MI-score) (Oakes 1998, p. 89). Its main advantage is the clear interpretation within the Shannon's theory of communication, which is, however, also its main weakness: not every linguist is familiar with the notions of information, entropy, etc., and the results are sometimes not fully understood. It could also be noted that collocability can also be of importance outside purely linguistic usages, e.g. for historical research and digital humanities, and the familiarity with the Shannon's theory in such fields is usually lower than with linguists. Then, the MI-Score results represent just numbers that can be compared, but have no obvious interpretable meaning. Let us imagine typical situation of a researcher who wants to compare two Arabic phrases: the first one is a typical Quranic epiteton of the devil, *aš-šayta:n ar-raji:m* (the-devil the-stoned), the second one, *nazaɣa:t aš-šayta:n* ("insinuations of a devil").

Phrase	Gloss	Count
ar-raji:m	the-stoned	2500
aš-šayta:n	the-devil	52747
aš-šayta:n ar-raji:m	the stoned the devil	2337

Table 1. Absolute frequency of the phrase *aš-šayta:n ar-raji:m* in the corpus (more information on the corpus in the next section).

Phrase	Gloss	Count
nazaɣa:t	insinuations	164
aš-šayta:n	the-devil	52747
nazaɣa:t aš-šayta:n	insinuations of the devil	89

Table 2. Absolute frequency of the phrase *nazaɣa:t aš-šayta:n* in the corpus.

A general formula for acquiring the Mutual Information of the phrase in question is given here:

$$MI(insinuations,\ devil) = \log_2 \frac{\frac{count(insinuations, devil)}{size\ of\ the\ corpus}}{\frac{count(insinuations)count(devil)}{size\ of\ the\ corpus^2}}$$

$$MI(stoned,\ devil) = \log_2 \frac{\frac{count(stoned,\ devil)}{size\ of\ the\ corpus}}{\frac{count(stoned)count(devil)}{size\ of\ the\ corpus^2}}$$

By substitution, we get the following formula:

$$MI(insinuations,\ devil) = \log_2 \frac{\frac{89}{4000000000}}{\frac{164 \cdot 52747}{400000000^2}} = 12\ bits$$

$$MI(stoned, devil) = \log_2 \frac{\frac{2337}{4000000000}}{\frac{2500 \cdot 52747}{400000000^2}} = 12.8\ bits$$

If the comparison is the only interpretation of these two results, then the formula can be far simpler. Let us start with the following inequation:

$$\log_2 \frac{\frac{c\ (insinuations, devil)}{size\ of\ the\ corpus}}{\frac{c(insinuations)c\ (devil)}{size\ of\ the\ corpus^2}} < \log_2 \frac{\frac{c\ (stoned,\ devil)}{size\ of\ the\ corpus}}{\frac{c(stoned)c(devil)}{size\ of\ the\ corpus^2}}$$

Which can be simplified by the following steps:

$$\frac{\frac{c(insinuations, devil)}{size\ of\ the\ corpus}}{\frac{c(insinuations)c(devil)}{size\ of\ the\ corpus^2}} < \frac{\frac{c(stoned,\ devil)}{size\ of\ the\ corpus}}{\frac{c(stoned)c(devil)}{size\ of\ the\ corpus^2}}$$

$$\frac{c(insinuations,\ devil)}{c(insinuations)c(devil)} < \frac{c(stoned,\ devil)}{c(stoned)c(devil)}$$

$$\frac{count(insinuations,\ devil)}{count(insinuations)} < \frac{count(stoned,\ devil)}{count(stoned)}$$

Then, the resulting formula is far simpler and well understandable even for those not comfortable with mathematics. We suggest that the new simplified score is called the RC score (i.e., Restricted Collocability score):

$$RC\ score = \frac{count(insinuations,\ devil)}{count(insinuations)}$$

The ranking of the results would still be the same as the ranking of the MI-Score, but the formula consists only of a single proportion. What does the proportion mean? Let us illustrate it by the diagram in figure 1.

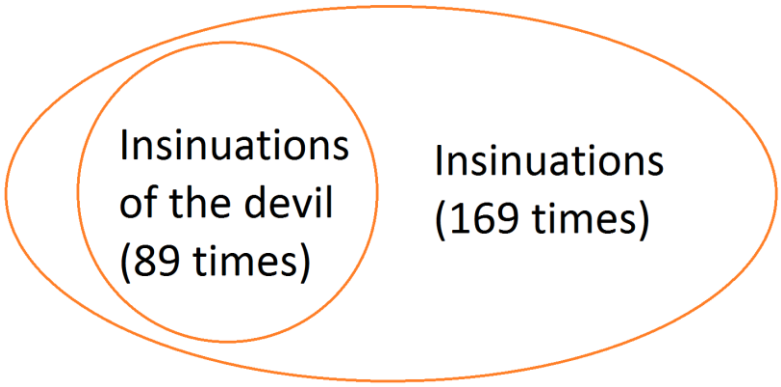


Figure 1. Representation of the sets that take part in RC score.

Basically the cardinality of the set is divided by the cardinality of its subset. More than every second occurrence of the word "insinuations" is a part of the "insinuations of the devil" phrase:

$$RC\ score(insinuations, devil) = \frac{89}{164} = 0.54$$

And nearly every occurrence of the word "the stoned" is collocated with "the devil":

$$RC\ score(devil, stoned) = \frac{2337}{2500} = 0.93$$

We think that the best thing about this metric is that it represents an effect size and that the confidence intervals (Wallis 2013) can be easily calculated so that the randomness can be taken into account. The result can be depicted as follows:

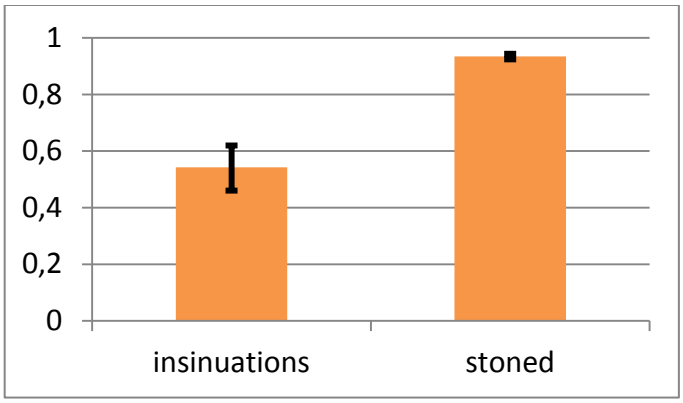


Figure 2. Graphical representation of the two proportions.
The 95 % exact binomial confidence intervals
are represented by the black bars.

Let us get back to the second phrase (the devil the stoned) for a while. The value of the metric is close to 1 which means that any other occurrence of the word "the stoned" except in this collocation is very rare (in our case, only 7% of "the stoned" occur without "the devil"). Such words and the phrases they form can be interesting not only from the theoretical point of view but also for practical reasons. That is why they attracted lexicographers and some dedicated dictionaries can be found (Čermák 2014).

3. RESTRICTED COLLOCABILITY: CASE STUDIES

It should be noted that the metric is not universal and has some prerequisites. One of them is the size of the corpus – the corpus needs to be large enough to reflect such relations satisfactorily. Also, not all words exhibit such behaviour – there are many words whose collocations are numerous, in other words, whose collocability is free. This is, however, natural as the method is intended for studying strong collocability.

The usage in lexicographic applications oriented at searching for collocations and idioms seems rather trivial to us – the method can be used in measuring strong collocational relations basically in the same way as is the case with other metrics, only the results seem to us much more intuitive and thus more friendly to some kind of (correct) hierarchization. That is why we will not try to show such examples.

That is why we want to concentrate on other cases where the method can be used for not just purely lexicographic or dictionary making purposes. There are many of them, e.g., it can be used in searching for personal names consisting of several words (such as the collocability of a name and surname or the collocability of some signaling label such as *ʾabu:* "father" in Arabic names like *ʾabu: bakr*).

In the following, we will concentrate on two possible usages, the study of prepositional valency (still a linguistic task) and a "lexical profiling", aimed at properties of a certain concept – the second application has many usages outside the field of linguistics.

3.1. The Data

Our dataset is based on a historical corpus of Arabic reflecting the use of the language in medieval times. The corpus, called CLAUDia (Corpus Linguae Arabicae Universalis Diachronicum), contains cca 440 million words and covers all the main phases of the Arabic literature, beginning with the text of the Quran (7th century C.E.) to texts dated to

mid 20th century C.E. The corpus is not based on samples, and as such, contains full texts of individual titles, based on edited manuscripts. Altogether, there are texts of more than two thousand books. All the main registers (genres) that played role in the history of the Arabic literature, are represented in the corpus. The corpus is not annotated with linguistic information. This is important for both the algorithm development and postprocessing of the retrieved data (ordering flective forms according to their lemmas).

Generally, it should be noted that here we are interested in cases where the restriction is substantial. Thus, the limit where our data is cut is rather strict, usually below RC score = 0.5.

Data is taken from 2- and 3-gram sets, depending on the quality of rendering of individual cases.

3.2. Prepositional Valency

This topic is not new with the method in question. There have been attempts to study the phenomenon before (e.g., Čermák 2006). The basic idea is that beside content words that are usually in the centre of lexicographers' attention, there are also grammatical words that can be studied by this method.

Such type of focus can help even in purely lexicographic tasks such as finding idioms or sayings: the doublet *ʕala: ɣa:ribi-ka* "on your withers" (the second person singular is far prevalent) with RC score = 1 is part of several sayings (*rama: rusna-ka ʕala: ɣa:ribi-ka* "he threw the halter on your withers", *ħablu-ka ʕala: ɣa:ribi-ka* "your rope is on your withers") used in the sense "to leave thy way free, open", "you are free to do whatever (or to go wherever) you will" (dictionaries note only the second phrase).

Focusing on the grammatical side, which has been already proposed by Čermák, shows that the prepositional valency can be studied in this way, too. In Arabic lexicography, this topic is rather understudied (for the best analysis so far, cf. El-Ayoubi et al. 2010; within the framework of Functional Generative Description, it has also been studied by Bielický 2015).

It should be noted that our data, based on bigrams, reflect only immediate valency, i.e. cases where the string containing the content word is immediately followed by the preposition. This is not completely in accord with the typical word order rules in Classical Arabic, where the VSO order is prevalent, and thus the preposition should, in case the subject is explicit, follow on the third position after the content word.

The following table presents a sample out of 32 occurrences of the root *ħrṣ* ("to desire, covet") in combination with the preposition *ʕala:* "on":

Phrase	Gloss	RC	Phrase count	Count 1	Character
ħaraṣa ʕala:	he desired for	0.57	1100	1945	verbal
yaħriṣu ʕala:	he desires for	0.53	454	851	verbal
ħari:ṣan ʕala:	eager for	0.71	1099	1541	nominal
ħaraṣan ʕala:	in the desire for	0.64	223	348	nominal
ħaraṣu-hu ʕala:	his desire for	0.48	451	931	nominal
ħaraṣu-hum ʕala:	their desire for	0.71	223	316	nominal

Table 3. Sample of the valency forms of the root *ḥrṣ* ("to desire, covet")

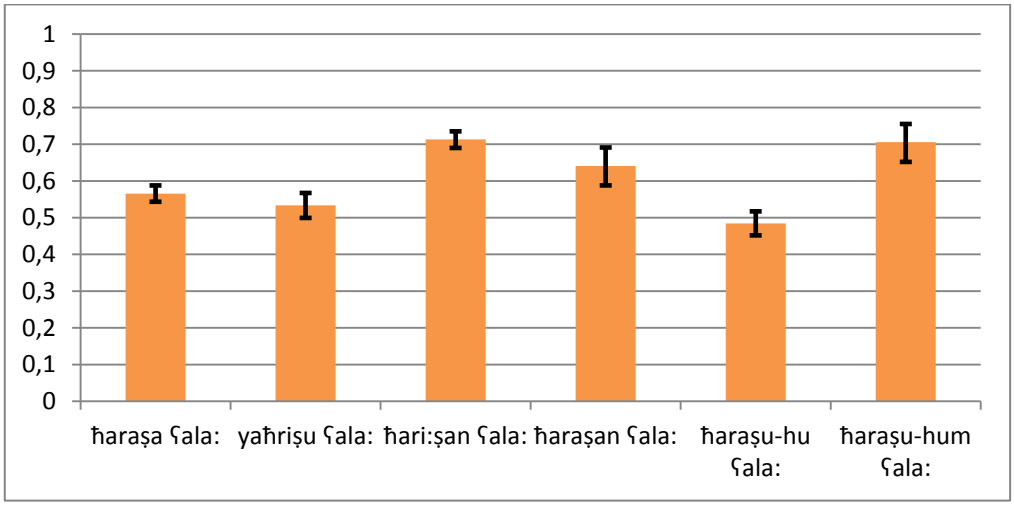


Figure 3. Graphical representation of the proportions given in Table 3. The 95 % exact binomial confidence intervals are represented by the black bars.

The valency of this root is strong in both verbal and nominal spheres. Having in mind the VSO order, one can expect higher RC score with nominal phrases, which can be observed in the table. However, the ratio at verbal forms is still very high.

Another analysis concerns the root *ḥml* ("to bear, carry"). This is, according to dictionary information, connected with the prepositions *ʔila:* ("towards"), *ʕan* ("on") and *min* ("from"), as the last one comes *ʕala:* ("on").

In the immediate context (bigrams), we find 27 high RC score collocations with *ʕala:* (in various morphological forms, both verbal and nominal, RC score ranging from 0.5 to 0.9),while there are only 2 instances of connection with *ʔila:* (RC score 0.54 and 0.69) and no instances of *ʕan* and *min*.

It is clear that in immediate position, the connection with *ʕala:* is prominent, while other connections are negligible. This does not exclude other valency frames, one has to be aware of the fact that some frames appear on more distant positions and one also has to admit that there can be constructions such as (carry (on st.)(for sb)), for example. On the other hand, it shows that there is difference in the positioning of various prepositions. It is

obvious that in this case, *ʕala:* is strongly bound to the immediately following position, which is information useful e.g. in building the valency frames. Such relation should certainly be reflected both in dictionaries, but in other disciplines like language teaching, etc.

3.3. Lexical Profiling

In computational linguistics, this term is sometimes used to describe the concept of keyness used in corpus linguistics. In our case, however, the idea is somewhat different. We consider specific concepts and check whether their collocates exhibit strong relation to such concepts. In other words, if a verb or adjective occurs frequently in collocation with a concept, it can be considered its property or description. Should such a word occur almost exclusively with some concept, in such an extent that can be measured by our metrics, it can be considered a unique, prototypical property of that concept, as it is strongly bound to it. Then, interpretation of such properties of a concept then can go beyond the field of linguistics.

In order to illustrate the principle, we have chosen the concepts of demons and angels. In the following, we will show with which properties these concepts are related in the language use in Arabic as a reflection of human and/or cultural imagination.

3.3.1. The collocations "devil"

The phrases featuring "devil" and exhibiting high RC score are rather numerous. They can be found for devil in both singular and plural, and there is a significant overlap of the two sets. For simplicity, we have chosen only the singular (and one dual) occurrences, summed up in the following table:

Phrase		RC	Count	Source
aš-šayṭa:n al-raji:m	the stoned devil	0.93	2337	Quranic
yataxabbaṭu-ni: aš-šayṭa:n	the devil strikes me	0.91	684	Quranic
istazalla-hum aš-šayṭa:n fa-ʔazalla-hu aš-šayṭa:n	the devil makes him to commit an error	0.76	338	Quranic
šaya: ṭa:ni yatakaðða:ni	two devils accuse of lying	0.88	43	general
yaxtalisu-hu aš-šayṭa:n	the devil misappropriates him	0.83	229	general
rakaḍa:t aš-šayṭa:n	devil's impulses	0.91	86	general

Table 4. Phrases with high RC score featuring "devil"

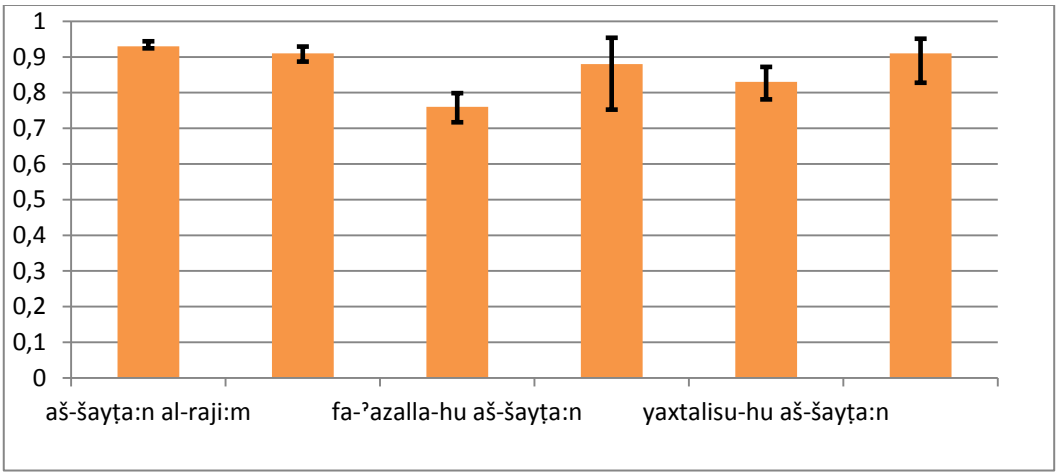


Figure 4. Graphical representation of the proportions given in Table 4. The 95 % exact binomial confidence intervals are represented by the black bars.

The table enables us to draw several types of results. First, the analysis of lexemes exhibiting strong connection with the concept shows that they are negative. Then, the strongest relation tells us that the devil should be stoned, we also learn that the devil does harm or leads humans to errors, etc. In other words, we learn what we should do about the devil and in which situations he can mislead us. Another observation is more culturally based – the (very rough) analysis of the sources where the connections appear shows us that Quran is not the only source of such information, although the strongest connections come from this source.

3.3.2. The collocations of "angels"

Contrary to what we find at "devil", phrases featuring "angels" and exhibiting high RC score are not as numerous. The singular is almost not represented with high RC score, and it is mostly the plural that is usable for further analysis.

Phrase		RC	Count	Source
ħaffat-hum al-mala:ʔika	the angels received them kindly	0.91	107	general
la-rafaʕat-ka al-mala:ʔika	the angels raised you (to heaven)	0.95	36	general
tatanazzalu ʕalay-him al-mala:ʔika	the angels delivered to you (the godly message)	0.91	578	general
tawaffa:-hum al-mala:ʔika	the angels took them to heaven	0.71	369	general
al-mala:ʔika qad tasawwamat	the angels have marked (them) with a good sign	0.94	33	general

Table 5. Phrases with high RC score featuring "angels"

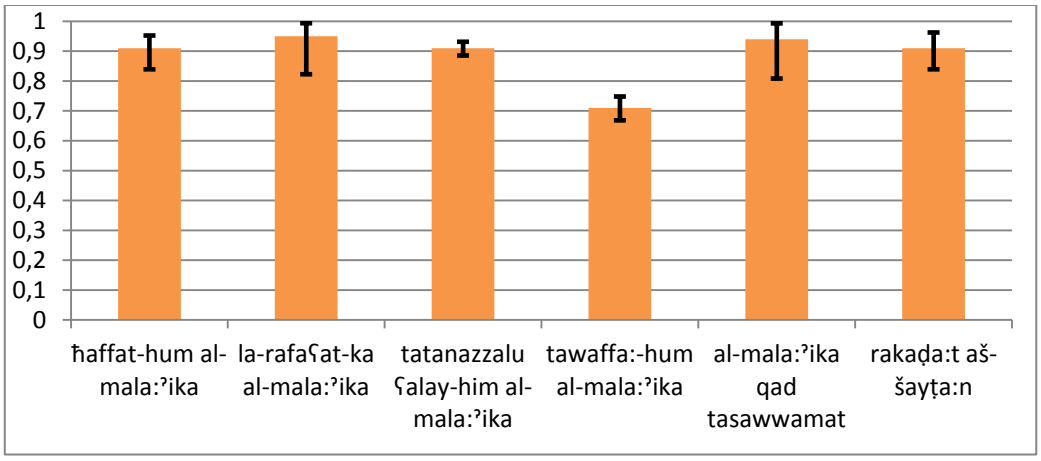


Figure 5. Graphical representation of the proportions given in Table 5.
The 95 % exact binomial confidence intervals are represented by the black bars.

One of the observations has been mentioned above – "angels" are specific rather in plural, thus, they may be seen as acting specifically only in a group, not as individuals. Second, their acts are positive and connected with godly acts (delivering the godly message) or take humans to heaven (interestingly enough, the connection of "devil" and "hell" is not exclusive). A brief look at the sources of such phrases shows us that "angels" as a group are not lexically represented as a specifically acting concept in the Quran (they exhibit free collocability).

3.3.3. Angels or demons?

The above given analyses show several interesting results. As expected, the angels exhibit positive values and acts, the devils are the opposite. However, angels act specifically only in plural, and one devil seems enough to do his specific tasks. This may be considered Quranic, as there is no occurrence of single angel in it. Another interesting difference is in the source of their acts – while in case of a devil, the Quranic influence is clear and some of the standard sayings about the devil stem from this source, it is not the case of the angels. Quranic contexts of angels do not give us these specific collocations.

As pointed out earlier, such uses of the method can go out of the field of linguistics, and such results can be further interpreted with the tools of other scholarly disciplines, such as religion studies in our case, but possibly other methods (historical, literary, etc.) in other cases.

4. CONCLUSION

The method presented in this contribution is not completely novel, but it has great potentials of use in lexicography, other fields of linguistics as well as non-linguistic scholarly disciplines. For most researchers, its interpretation is intuitive and easily graspable, which should prevent misinterpretations. As such, it can contribute to linguistic research, especially the lexicographical one, by widening the palette of metrics used for measuring collocability.

Using the method for extra-linguistic purposes has been demonstrated by the comparison of the unique properties of the two concepts, "angels" and "devil". The

analysis shows interesting results that can be further interpreted by non-linguistic tools and methods.

It should be, however, noted that it is usable only with large corpora (hundreds of millions of words), as the properties of lexemes may not appear correctly with smaller corpus sizes.

Acknowledgements

The research reflected in this article has been supported by the Grant Agency of the Czech Republic, project no. 13-28220S. We are also grateful to the anonymous reviewers for their valuable comments.

References

BARKEMA, H., 1996. 'Idiomaticity and Terminology: A Multi-Dimensional Model.' *Studia Linguistica* 50(2), 125-160.

BENTIVOGLI, L. AND PIANTA, E., 2003. 'Beyond lexical units: Enriching wordnets with phrasets.' *Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL03)*, Vol. 2, 67-70.

BENTIVOGLI, L. AND PIANTA, E., 2004. 'Extending wordnet with syntagmatic information.' In *Proceedings of Second Global WordNet Conference*, 47-53. Brno.

BIELICKÝ, V., 2015. *Valenční slovník arabských sloves*. PhD thesis, Charles University, Prague.

ČERMÁK, F., 2006. 'Collocations, Collocability and Dictionary.' In: In E. Corino, C. Marello, and L. Onsti (eds.), *Proceedings of XII EURALEX International Congress,* Torino: Edizioni dell' Orso, Vol. II, 929–937.

ČERMÁK, F., 2014 *Periferie jazyka – Slovník monokolokabilních slov*. Praha: Nakladatelství Lidové noviny.

GRANGER, S. AND MEUNIER, F. eds., 2008. *Phraseology: An interdisciplinary perspective.* Amsterdam: John Benjamins.

HANSEN, G., MALMKJÆR, K. AND GILE, D. eds. 2004. *Claims, Changes and Challenges in Translation Studies. Selected Contributions from the EST Congress.* Copenhagen 2001. Amsterdam/Philadelphia: John Benjamins.

HOWARTH, P., 1998. 'Phraseology and second language proficiency.' *Applied Linguistics*, 19, 24-44.

EL-AYOUBI, H., FISCHER, W. AND LANGER, M. eds., 2010. *Syntax der Arabischen Schifsprache der Gegenwart: Die Verbalgruppe. Teil II.* Wiesbaden : Reichert Verlag.

OAKES, M., 1998. *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.

THE BRITISH NATIONAL CORPUS, VERSION 3 (BNC XML EDITION). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. [online] Available at: http://www.natcorp.ox.ac.uk/ [Accessed 10 March 2015].

WALLIS, S., 2013. 'Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods.' *Journal of Quantitative Linguistics* 20(3), p 178.

GERMAN-INTO-BASQUE/SPANISH TRANSLATION ANALYSIS OF BINOMIALS IN A PARALLEL AND MULTILINGUAL CORPUS

Zuriñe Sanz Villar
UPV/EHU
zurine.sanz@ehu.eus

Abstract

As Piirainen put it (2012, p. 43); binomials are "sequences of two or more constituents that belong to the same grammatical category, have some semantic relationship and are joined by a conjunction like *and* or *or*". This type of phraseological unit has been analyzed from several perspectives and across different languages (Malkiel, 1959; Gustaffson, 1984; Müller, 2009; Čermák, 2010; Toury, 2012; Pontrandolfo, 2011), but this is the first attempt to carry out a corpus-based translation analysis of binomials in the German-Basque/Spanish language combination.

1. INTRODUCTION

The main goal of the present[14] paper is to analyze the translation of binomials in the German-Spanish-Basque language combination using the AleuskaPhraseo corpus, which — it is important to emphasise — is multilingual, as opposed to bilingual. Since a considerable number of German-into-Basque translations have been labeled as indirect — that is, as translations that have not been made directly from the German source text — in the Aleuska catalogue — which consists of all literary works that have been translated from German into Basque over history —, a number of intermediary texts were added to the corpus in the case of indirect translations. In doing so, it becomes possible to compare both direct and indirect translations.

In the present paper, relevant research lines in multilingual phraseological research will be summarized; then, the creation process of the AleuskaPhraseo corpus, a "small-scale topic-specific PTC [Parallel Translational Corpus]" (Ji, 2010, p. 6) consisting of around 3.5 million words from where the binomials under analysis have been extracted, will be described. Subsequently, the actual process of binomials extraction using AntConc will be outlined. Finally, drawing on the theoretical and methodological framework proposed by Toury (2012), the results obtained from the translation analysis will be reported.

[14] This paper has been done within the framework of the Tralima-Itzulik (GIC12/197, UPV/EHU) research group.

2. THE "TRANSLATIONAL TURN" IN MULTILINGUAL PHRASEOLOGICAL STUDIES

When comparing phraseological units (PU) from two (or more) languages, two different approaches can be adopted: the contrastive and the translational approach. From the point of view of contrastive phraseology, PUs of different languages are studied at the *langue* or system level, which means that PUs are analyzed in isolation, without reference to the surrounding context. The main goal of the contrastive approach is to establish similarities and differences between two (or more) phraseological systems.

Within translation research, on the other hand, the comparison is made at the *parole* level. The goal here is not to compare the phraseological universes in terms of similarity and difference, but rather to state how source text PUs have been translated into the target text, taking into consideration the context in which the corresponding PUs are embedded. Nevertheless, at the Europhras conference held in Maribor in 2012, Korhonen[15] pointed out that actually, in the early stages of translational studies on phraseology, scholars used to use terms and methodologies germane to the ones used in contrastive linguistics. In that sense, the goal of most of the early work on the translation of PUs was to define types of equivalences and to classify phraseological material according to equivalence relations. Context was rarely taken into account when doing that kind of research.

However, in the same speech, Korhonen also pointed out that research studies examining the context in which PUs are embedded had increased from the 1970s onwards. Based on the bibliographical material on phraseological and paremiological research that Wolfgang Mieder[16] published in two volumes in 2009, it can similarly be said that in general the number of studies focusing on translation within phraseology increased significantly from the 1980s on. In order to evaluate these claims I performed a search in Mieder's bibliographical database for publications with the keyword "translation" (702) and organized the results chronologically, which can be seen in the following figure (figure 1).

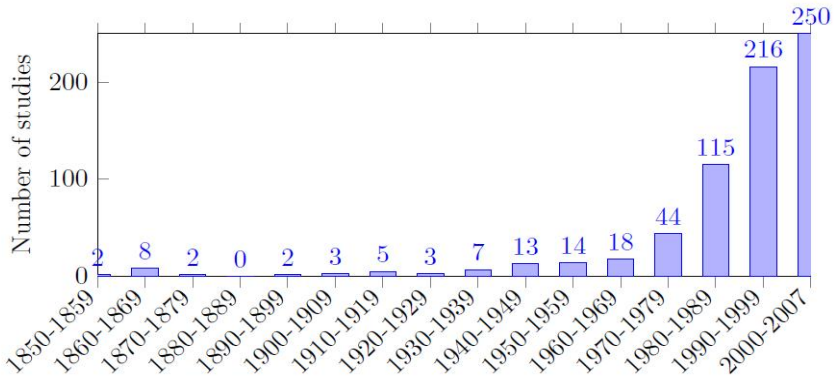


Figure 1. Number of studies featuring the keyword "translation"

Figure 1 provides us with a clear picture of how many studies were carried out in the past which merged both phraseology and translation, but does not shed light on how the

[15]Korhonen's speech can be foundonline at: http://www.youtube.com/watch?v=q9eDgxIMbm0

[16]Mieder (2009) did a tremendous job collecting the bibliography on phraseology and paremiology between 1852 and 2007. The result is the two volumes, mentioned above, which include 10.000 entries. The publications are organized both alphabetically and also labeled according to their topic, which greatly facilitates a search for related publications.

studies were carried out. With the aim of exploring this further, I will now outline two pertinent theoretical and methodological approaches; the prescriptive/descriptive approach and the corpus-based approach.

Within the prescriptive approach, the starting point is usually an ideal PU equivalence between two (or more) languages, and, when in translation such equivalence is not found, there develops a translation problem where the translation is not regarded as the ideal equivalent. As a result, these kinds of studies tend to have a prescriptive character as reflected in Hallsteinsdottir's (2011, p. 286) quote: "The perception of an obligatory equivalence within the linguistic category of phraseology focuses solely on linguistic items in the source text as objects of the language system and it disregards other aspects of translation, both seen as a process and as a text. As a consequence, it generates and maintains an illusion of a mandatory systemic equivalence of phrasemes in translation". Similarly, Farø (2006) describes an unwritten and implicit law, according to which a PU in the source language should be translated through the use of an equivalent PU in the target language, or, at least, the phraseological image should be maintained in the corresponding target language.

In Corpas 2000, the author describes the procedures and techniques that have been used when translating Tom Sharpe's work, Wilt, into Spanish. In her article, she mentions the tension between prescriptive and descriptive studies, especially in cases where the prescriptive approach maintains that the best way of translating a PU is using another PU. However, in her analysis, Corpas concludes that this is in fact not the procedure that is always employed. Furthermore, she makes an interesting point in relation to the use of corpora in translation studies, when analyzing the translation of PUs: "Las conclusiones derivadas de este estudio no son representativas ni tienen validez absoluta, pero sí permiten vislumbrar ciertas tendencias en la traducción de la fraseología, que, curiosamente, ponen en tela de juicio algunas de las opiniones vertidas tradicionalmente por los teóricos de la traducción sobre el tema. En todo caso, será necesario realizar investigaciones a gran escala que confirmen o desestimen, por idiosincrásicos e idiolectales, los resultados de nuestro análisis. Ello sólo será posible cuando se disponga de córpora paralelos de textos traducidos en soporte informático y un *software* adecuado de alineación automática. Hasta entonces, habrá que seguir analizando manualmente binomios textuales concretos que nos permitan vislumbrar cuáles son las estrategias y procedimientos preferidos por los traductores, qué riesgos y pérdidas semánticas entrañan, y de qué manera se pueden aplicar dichos resultados a una incipiente pedagogía de la traducción" (Corpas, 2000, p. 518).[17]

Thanks to recent technical developments, the wish expressed by Corpas in the year 2000 has been realized, allowing us now to talk about a "translational turn" in multilingual phraseological studies. In fact, there are an increasing number of studies within Translation Studies that have used corpora to analyze, among many other linguistic phenomena, PUs.

To that regard, the COVALT research group from the Jaume I University must be mentioned. As can be read in the introduction of the last volume they published (Bracho, 2013), since 2001 the goal of the research group has been to compile and analyze the

[17]"The conclusions drawn from this analysis are neither representative nor do they have absolute validity, but they do show certain tendencies regarding the translation of phraseology, and interestingly enough, they cast doubts on some of the opinions expressed on this matter by translation scholars. In any case, it will be necessary to carry out large-scale studies that will support or reject, given the idiosyncratic and idiolectal character of our study, the results of our analysis. This will be possible when digitized and parallel corpora, and adequate software for the automatic alignment of texts, are available. Until then, we will need to keep manually analyzing specific textual binomials, which will enable us to define the strategies and procedures mostly employed by translators, the risks and semantic losses they present, and how the results can be useful within an emerging translation pedagogy" (my own translation).

COVALT (Corpus Valencià de Literatura Traduïda) corpus, which comprises around four million words and is made up of literary texts published by Valencian publishing houses between 1990 and 2000 (Guzmán and Serrano, 2006, p. 1744).

For conducting corpus queries, a management program called AlfraCOVALT is used, which allows for queries to be made according to authors, translators or languages (Guzmán and Serrano, 2006, p. 170). As for analysis, to my knowledge, three of the members — Josep Marco, Heike van Lawick and Ulrike Oster — have been working within Toury's descriptive approach (Bracho, 2013, p. 2), and their analysis has focused mainly on the translation of somatic PUs from English and German into Catalan.[18]

The research study I have carried out over the past years within the framework of my PhD thesis (Sanz, 2015) shares many similarities with the studies from the COVALT research group, as will be shown in the following sections, and, in that sense, is intended to humbly contribute to the aforementioned "translational turn".

3. THE ALEUSKAPHRASEO CORPUS: A DIY CORPUS

Since there was no parallel corpus containing German-into-Basque translations at the time I began writing the PhD thesis, I created my own corpus using TraceAligner[19], a program developed within the TRALIMA-ITZULIK research group (UPV/EHU) that serves to build up and exploit parallel and multilingual corpora.

The word that explains best the process of creating the corpus is, undoubtedly, "time-consuming". The texts needed to be selected, digitized, cleaned, tagged and aligned, but everything was worth it, since the end product, the AleuskaPhraseo corpus, as well as the tool, the TraceAligner program, are being used and will be used beyond the scope of the aforementioned PhD thesis.

The texts that comprise the corpus were selected according to four main qualitative criteria. Firstly, only texts translated from 1980 on were included in the corpus. This is because a standardized form of the Basque language has only existed since 1969, and after its creation, the need to create texts in Basque, translations included, became urgent. In the case of Basque translations of German narrative texts, for example, the number of translations increased considerably from 1980 onwards. Secondly, texts labeled in the Aleuska catalogue as either direct or indirect translations were included in the corpus. Because Basque is a minority language that coexists with another dominant language, Spanish, and indirect translations have been quite common throughout the history of the Basque translation practice. Since one of the aims of this study was to compare direct and indirect translations, Spanish translations were included, in the case that their Basque translations had originally been made through an intermediary text. The third criterion is related to the genre of texts and ensured that both children's literature (CL) and adult literature (AL) texts were represented; finally, every effort was made to ensure the presence of a diversity of both source and target authors.

As mentioned before, the compilation process of the AleuskaPhraseo corpus consisted of four steps: digitizing, cleaning, tagging and aligning the texts. Obtaining all the texts in

[18]Apart from the English-Catalan and German-Catalan subcorpora, there is also an additional French-Catalan subcorpus.

[19]Although the version I used for the creation and exploitation of the corpus at the time I was elaborating my PhD thesis was the 2.0 version of the program, the one I presented here is the improved 3.0 version.

electronic format was not an easy task, since many of the texts had to be scanned. For the purpose of correcting the formatting errors, an application integrated within the TraceAligner program (the option "Limpiar" in Figure 2) was used, but other errors, such as character-recognition errors, had to be corrected manually. Afterwards, the texts were aligned with the same program (with the options "Etiquetar" and "Alinear"). This figure shows an example of what an alignment of three texts looks like.

Figure 2. Interface of TraceAligner 3.0 showing the alignment of three texts simultaneously

Once the texts were aligned, they were saved into a searchable database. What can be seen in Figure 3 is the search engine's interface along with the result of an example query. This tool provides the researcher not just with the sentence that contains the searched for word or words, but also the preceding and subsequent sentence, as can be observed in the Figure 4.

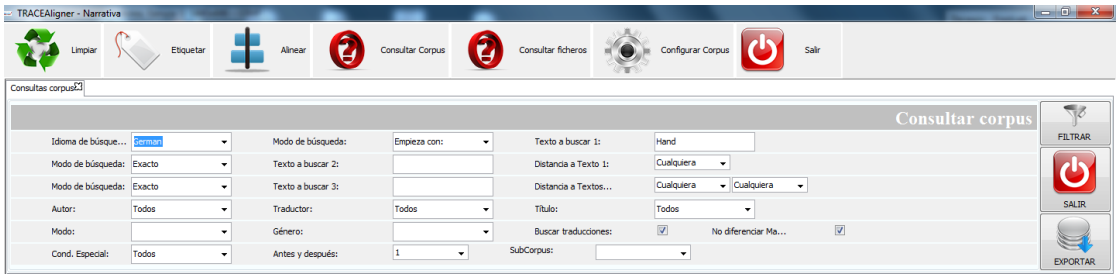


Figure 3. Search engine's interface in TraceAligner 3.0

: (191)	Lilli führt ihn zum Kühlschrank.	: (191)	Kikak hozkailuraino eramaten du,	: (191)	Kika lo conduce hasta la nevera,
: (192)	Sie öffnet die Kühlschranktür, gießt ihm einen winzigen Schluck Mineralwasser in ein Glas und gibt es ihm in die Hand.	: (192)	ur mineral tanta batzuk botatzen ditu baso batean eta Daniren eskuan jartzen du ontzia.	: (192)	echa unas pocas gotas de agua mineral en u mano de Dani.
: (193)	Diesmal ist Leon misstrauisch.	(193)	Oraingoan, mutila ez da horrenbeste fidatzen;	(193)	Esta vez, él ya no se fía tanto,
: (325)	Ob der Bursche sich genau wie sie versteckt? Ist ja nicht weiter schwierig in dieser wild wuchernden Umgebung.	: (325)	baina hainbeste landareren artean ezkutatzea erraza da.	: (325)	aunque con tanta vegetación es fácil esconc
: (326)	Der Stimme nach muss es sich um einen Mann handeln.	: (326)	Gizon baten ahotsa zirudien... .	: (326)	Aquella voz parecia la de un hombre... .
: (327)	Lilli wartet ab.	: (327)	Kika erne dago,	: (327)	Kika espera,

Figure 4. The result of an example query

At present, the corpus contains 110 texts, from which 34 are assumed to be direct (DI) translations and 14 indirect (INDI).[20] Diversity in terms of authors is ensured — with 30

[20] The corresponding source texts — one source text in the case of the direct translations and two (the German and the Spanish intermediary text) in the case of the indirect translations — needed to be added in order to obtain the final total of 110 texts.

different source authors (SA) and 28 different target authors (TA) or translators —, and in total, the corpus consists of around 3.5 million words, as Figure 5 shows.

	Translation mode		Authors		Number of words		
	DI	INDI	SA	TA	DE	EU	ES
AL	19	5	17	15	1,120,534	935,530	198,274
CL	15	9	13	16	593,871	512,204	166,860
Total	34	14	30	28	1,714,405	1,447,734	365,134

Figure 5. Main features of the AleuskaPhraseo corpus

4. EXTRACTING BINOMIALS WITH ANTCONC

Binomials are very productive in different languages, and the translation of binomials has been studied by scholars, such as Toury (2012, p. 131-141), who obtained interesting results on the use of binomials (or *conjoint phrases of near-synonyms*[21]) in Hebrew translations. Aside from their productivity and the interest within the research community, another reason to have selected binomials as object of study is practical: with some corpus-analysis tools — AntConc[22], for instance — the extraction of this type of PU can now be easily achieved.

In this study, the "Clusters/N-Grams" option of the AntConc program was employed to extract the German and Basque binomials from the corpus. All in all, 203 types of German binomials were extracted, and this figure corresponds to 1,456 tokens. In Basque, the number of extracted binomials in terms of types (243) and tokens (2,563) was larger. It is important to note that thus far in the process, the extracted binomials are still only potential binomials, and that contextual analysis is needed in order for a decision to be made regarding whether a word combination is in fact a binomial or not (Müller, 2009, p. 35).

Considering the large number of tokens found for each of the languages, a representative sample needed to be selected. For that purpose, the following statistical formula from Izquierdo (2012) was used. In the formula N represents the number of tokens, E the acceptable margin of error (0.05 %), and n the representative sample to be obtained.

$$n = \frac{N}{(N-1)E^2 + 1}$$

Figure 6. The statistical formula

[21]"Conjoint phrases of synonyms consist in two (occasionally more than two) synonymous or near-synonymous items of the same part of speech, combined to form a single functional unit" (Toury, 2012, p. 132).

[22]AntConc is a multiplatform freeware, which, among other things, enables the user to automatically scan "the entire corpus for 'N' (e.g. 1 word, 2 words, ...) length clusters" (Anthony, 2014), in order to find common expressions in the uploaded corpus.

According to the formula, 314 German and 346 Basque binomials should be extracted and analyzed in order to work with a representative sample. Prior to the selection, the binomials were arranged according to their formal structure, and then, a selection was made for both languages, based on specific criteria (such as formal structure, meaning or frequency). Both lists can be consulted in the Figures 7 and 8.

Binomials	Freq.	Structure
mehr oder weniger	18	X und/oder... Y
Art und Weise	16	X und/oder... Y
einzig und allein	10	X und/oder... Y
an Ort und Stelle	8	X und/oder... Y
nie und nimmer	8	X und/oder... Y
früher oder später	7	X und/oder... Y
fix und fertig	5	X und/oder... Y
warum und wieso	5	X und/oder... Y
(über) kurz oder lang	4	X und/oder... Y
hin oder her	4	X und/oder... Y
tief und fest	4	X und/oder... Y
allem und jedem	3	X und/oder... Y
dieser oder jener	3	X und/oder... Y
für immer und ewig	3	X und/oder... Y
Grund und Boden	3	X und/oder... Y
gut und gern	3	X und/oder... Y
laut und deutlich	3	X und/oder... Y
ruhig und gelassen	3	X und/oder... Y
Ecken und Enden	2	X und/oder... Y
Ehre und Ruhm	2	X und/oder... Y
Frieden und Ruh	2	X und/oder... Y
heil und gesund	2	X und/oder... Y
Herr und Meister	2	X und/oder... Y
klar und deutlich	2	X und/oder... Y
klar und verständlich	2	X und/oder... Y
Rauch und Qualm	2	X und/oder... Y
Ruhe und Ordnung	2	X und/oder... Y
schlicht und einfach	2	X und/oder... Y
Schutt und Asche	2	X und/oder... Y
Sitten und Gebräuche	2	X und/oder... Y
steif und fest	2	X und/oder... Y
Ziel und Ende	2	X und/oder... Y
durch und durch	16	X und/oder X
mehr und mehr	15	X und/oder X
so oder so	7	X und/oder X
Tag für Tag	14	X um/für X
ein für allemal	11	X Präp Y
schwarz auf weiß	4	X Präp Y
Arm in Arm	26	X Präp X
von Kopf bis Fuß	9	von X bis/auf/zu/nach Y
von Zeit zu Zeit	49	von X zu X
nach wie vor	25	X wie/als Y

Figure 7. List of extracted and analyzed German binomials

Binomials	Freq.	Structure
argi eta garbi	89	X eta/edo/ala Y
zur eta lur	59	X eta/edo/ala Y
nola edo hala	56	X eta/edo/ala Y
estu eta larri	14	X eta/edo/ala Y
jira eta bira	12	X eta/edo/ala Y
jaun eta jabe	9	X eta/edo/ala Y
korrika eta presaka	8	X eta/edo/ala Y
su eta gar	9	X eta/edo/ala Y
jira eta buelta	4	X eta/edo/ala Y
egiaz eta benetan	2	X eta/edo/ala Y
guztiz eta erabat	2	X eta/edo/ala Y
jauzika eta itzulipurdika	2	X eta/edo/ala Y
mesedez eta faborez	2	X eta/edo/ala Y
oihuka eta garrasika	2	X eta/edo/ala Y
pozik eta alai	2	X eta/edo/ala Y
saltoka eta brinkoka	2	X eta/edo/ala Y
zalaparta eta iskanbila	2	X eta/edo/ala Y
zuzen eta bidezkoa	2	X eta/edo/ala Y
kosta ahala kosta	26	X eta/edo/ahala/ala X
dir-dir	36	X (-) X

Figure 8. List of extracted and analyzed Basque binomials

5. TRANSLATION ANALYSIS: MAIN RESULTS

In order to analyse the resulting binomial translations, my starting point was with the list of translation techniques as described by Marco (2008, 2009, 2013) in his studies on the English-into-Catalan translation of PUs, which I then adapted according to my own needs. A major difference between Marco's research study and my own is that here, searches were conducted in both directions; that is, not only starting from the German source texts (which is the usual way of designing translation analyses), but also from the Basque target texts. In what follows, firstly the results of the "source text-target text" analysis will be outlined, followed by the results obtained from the "target text-source text" analysis. In both cases, differences and similarities between CL and AL texts will be mentioned, as well as between direct and indirect translations.

5.1. "Source text-target text" analysis

Figure 9 shows that a considerable number of German binomials have been translated with another Basque binomial (50.64%), and this may indicate that there is an attempt on the translator's part to use PUs in the target texts that function as the source-text PUs' counterparts. As far as CL and AL texts are concerned, the translation option "PU-similar PU" has been used more often in adult-literature translations, which may show that the tendency to adhere to the source text is greater in AL translation.[23] This tendency is reinforced by another one: be it similar or different, the inclination to use binomials that are less common in "everyday" Basque is greater in AL texts. The translation of the German binomial *Arm in Arm* (*arm in arm* in English) and *von Zeit zu Zeit* (*from time to time*) is

[23] I am aware of the fact that the margin is quite narrow (18.22 to 13.43), but the same tendency has been observed when analyzing the translation of somatic PUs. In that case, the margin was much wider (32.51% in AL to 12.12% in CL) (Sanz, 2015).

prime evidence of just this. As for the former binomial, the Basque PU *besotik helduta* is more commonly used (22 occurrences) in the translations. Only three occurrences of the less usual *besoz beso* have been found in the translations, and all of them in AL texts. In the case of *von Zeit zu Zeit*, the options in Basque are more diverse — i.e., eight different binomial types were found in the target texts —, and the least common ones, according to the Basque reference corpus ETC — *alditik-aldira*, *noizetik noizera* and *txitean-pitean* —, were only used in AL texts. These results may also indicate that in AL the effort to produce more sophisticated literary texts is greater.

Another difference between both text types is that more direct copies (when the Basque binomials are, to a greater or lesser extent, word-for-word copies of the source-text binomials) have been observed in CL translations (16.42% to 9.31%). These raw figures may indicate that interference is much more accepted in CL translations than in AL. One more aspect that stands out in the data is the relatively high number of omissions, mainly in AL translations.

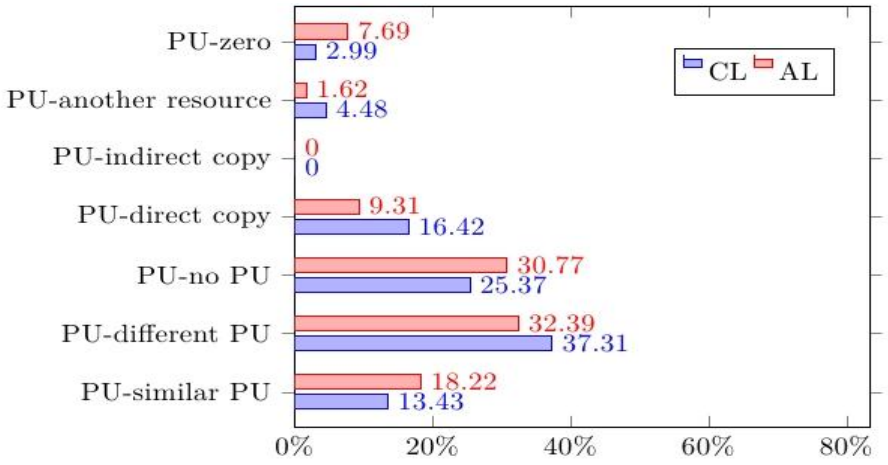


Figure 9. The translation options in CL and AL texts

With regards to direct and indirect translations (see Figure 10), PU-PU translation percentages are very similar in both cases (50.98% to 49.13% respectively). Since indirect translations are conditioned by another source text — the intermediary text — it seems logical that the number of similar PUs is higher when translating directly. Similarly, the fact that less phraseological solutions have been found in the Basque target texts (when there is a PU in the source text) in indirect translations could be explained in the same way. As for the direct copies, I would argue that the much lower percentages of direct copies in indirect translations may function as an indicator of more conventional and standardised language use when translating indirectly. Furthermore, during the translation analysis I have observed that Basque translators, when translating indirectly, sometimes tend to deviate from the Spanish intermediary texts. For instance, in the four cases in which the German binomial *von Kopf bis Fuss* (meaning "from head to foot") is used in the source texts and the Spanish similar PU *de pies a cabeza* in the intermediary versions, the Basque translators have opted to use the Basque binomial *goitik behera* (literally, "from top to bottom"). In doing so, I argue that the Basque binomial that more closely resembles the Spanish (and German) PU — *burutik oinetaraino* — may have been intentionally avoided by the translator. As opposed to, when translating directly from the German source texts, the latter Basque

binomial — *burutik oinetaraino* — is the one most often employed (three out of five times) in the Basque target texts.

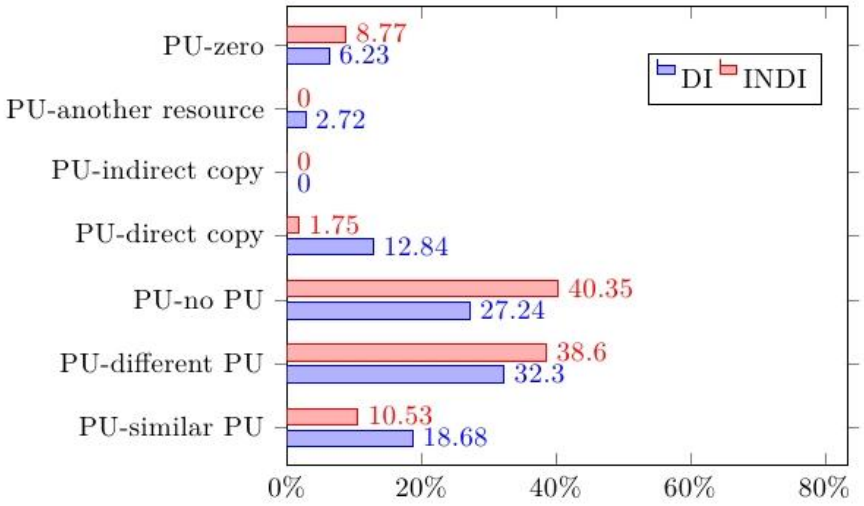


Figure 10. The translation options in texts translated directly and indirectly

5.2. "Target text-source text" analysis

As Figure 11 shows, in the majority of the cases, the counterpart of the extracted Basque binomial was a free word combination in German. At the same time, the cases in which the Basque binomials represent outright additions (because there is no counterpart in the source text) are not to be neglected. In this regard, Toury's remark on the translation of conjoint phrases of near-synonyms seems relevant. According to him, the large number of these word combinations in translations — especially when the counterpart in the source texts are single lexical items or when they represent outright additions — may be regarded as an attempt "to develop indigenous linguistic capabilities" (Toury, 2012, p. 140) in a weak system such as Basque. Although more data should be evaluated, it may, tentatively, be concluded that there is a tendency in Basque to use binomials in translations when there is no phraseological stimulus in the source text, or even when there is nothing at all in the source text. Let us have a look, for instance, at this example:

Ich war sprachlos (Homo Faber, 1957)	Hitzik gabe ni, **zur eta lur**. (Homo Faber, 2001)

Table 1. Example of the translation option "Ø-PU"

The Basque binomial *zur eta lur* (meaning "astounded") functions in this example as an outright addition, since "Ich was sprachlos" [I was speechless] corresponds to "Hitzik gabe ni" in Basque.

In the context of CL and AL text translation, as can be seen in Figure 11, the tendency to use binomials in the translation when there is no PU in the German text is greater in AL texts, whereas the "Ø-PU" translation option is more common in CL translation. As for the direct and indirect translations, on the other hand, the percentages do not differ much this time, as can be seen in Figure 12. However, when doing a more detailed analysis of the

translations, the tendency to deviate from the Spanish intermediary text in indirect translations seems to be confirmed, as explained in Sanz (2015).

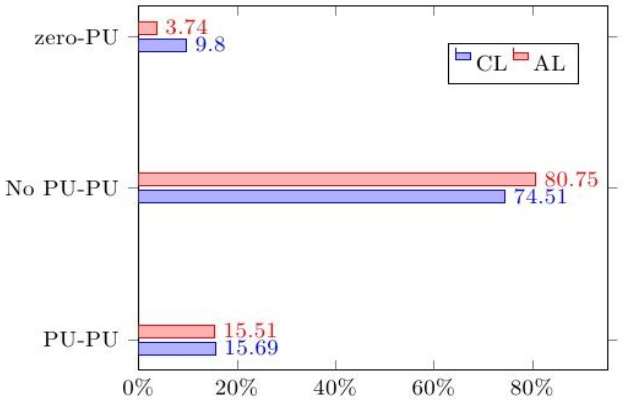


Figure 11. The translation options in CL and AL texts

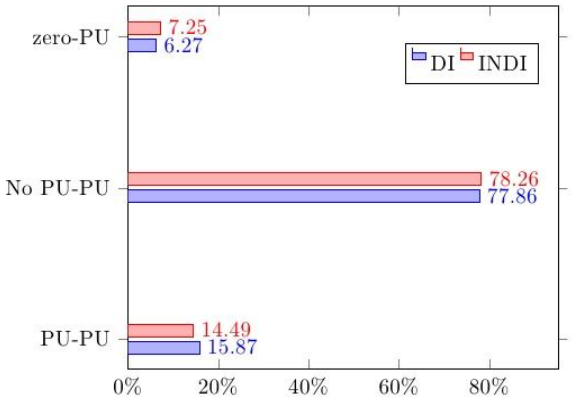


Figure 12. The translation options in texts translated directly and indirectly

6. CONCLUSIONS

Corpora play a very important role in several fields, such as Translation Studies and Phraseology. In my case, the TraceAligner tool and the AleuskaPhraseo corpus have both become an essential part of the German-into-Basque translation analysis of binomials that I have carried out over the past four years and within the framework of my PhD thesis. Thanks to these tools, I was able to create my own corpus, to systematically extract data from a 3.5-million-words database and to reach my research goals. In future research activities, I am sure they will keep playing an essential role. With regards to those future research lines; it will be very interesting to keep improving both TraceAligner and the AleuskaPhraseo corpus, and to explore other experimental methods to analyse the translation process that will serve to better understand the results obtained from corpus analyses.

References

ANTHONY, L., 2014. *AntConc (Version 3.4.3)*. [Computer Software]. Tokyo: Waseda University. Available at: http://www.laurenceanthony.net

BRACHO, L. ed., 2013. *El corpus COVALT: un observatori de fraseologia traduïda*. Aachen: Shaker.

CORPAS, G., 2000. Acerca de la (in)traducibilidad de la fraseología. In: G. Corpas, ed. 2000. *Las Lenguas de Europa: estudio de fraseología, fraseografía y traducción*. Granada: Comares. pp.483-522.

ČERMÁK, F., 2010. Binomials: Their nature in Czech and in general. In: J. Korhonen, W. Mieder, E. Piirainen, and R. Piñel, eds. 2010. *Phraseologie global-areal-regional.* Tübingen: Narr. p.309-315.

FARØ, K., 2006. Dogmatismus, Skeptizismus, Nihilismus und Pragmatismus bei der Idiomübersetzung: Grundfragen zu einer idiomtranslatorischen Theorie. In A. Häcki and H. Burger, ed. 2006. *Phraseology in Motion I. Methoden und Kritik*. Baltmannsweiler: Schneider. pp.189–202.

FRISCH, M., 1957. *Homo Faber. Ein Bericht*. Frankfurt am Main: Suhrkamp.

FRISCH, M., 2001. *Homo Faber*. Translated from German by J.A. Arrieta. Donostia-San Sebastián: Elkarlanean.

GUSTAFFSON, M., 1984. The syntactic features of binomial expressions in legal English. *Text*, 4, p.123–141.

GUZMÁN, J. R. AND SERRANO, À., 2006. Alineamiento de frases y traducción: Alfracovalt y el procesamiento de corpus. *Sendebar*, 17, p.169–186.

HALLSTEINSDÓTTIR, E., 2011. Phraseological competence and the translation of phrasemes. In A. Pamies, L. Luque and J.M. Pazos. eds., 2011. *Multi-Lingual Phraseography: Second Language Learning and Translation Applications*. Baltmannsweiler: Schneider. pp.279-288.

IZQUIERDO, M., 2012. Corpus-based functionality and translatability. English–Spanish progressive constructions in contrast and translation. *Languages in contrast*, 12(2), p.187–210.

JI, M., 2010. *Phraseology in Corpus-Based Translation Studies*. Frankfurt am Main: Peter Lang.

MALKIEL, Y., 1959. Studies in irreversible binomials. *Lingua*, 8, p.113–160.

MARCO, J., 2008. "In my mind's eye": análisis traductológico de algunos fraseologismos prototípicos en el contexto del corpus COVALT (corpus valencià de literatura traduïda). In L. Pegenaute, J.A. DeCesaris, M. Tricás and E. Bernal, eds. *Actas del III Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación.* Barcelona: PPU. pp. 251-262.

MARCO, J., 2009. Normalisation and the translation of phraseology in the COVALT corpus. *Meta: journal des traducteurs / Meta: Translators' Journal,* 54(4), pp.842-856.

MARCO, J., 2013. La traducció de les unitats fraseològiques de base somàtica en el subcorpus anglès-català. In L. Bracho, ed. *El corpus COVALT: un observatori de fraseologia traduïda.* Aachen: Shaker. pp.163-215.

MIEDER, W., 2009. *International bibliography of paremiology and phraseology*. Berlin; New York: Walter de Gruyter.

MÜLLER, H. G., 2009. *Adleraug und Luchsenohr: deutsche Zwillingsformeln und ihr Gebrauch*. Frankfurt am Main: Peter Lang.

PIIRAINEN, E., 2012. *Widespread idioms in Europe and beyond. Towards a Lexicon of Common Figurative Units*. New York: Peter Lang.

PONTRANDOLFO, G., 2011. *La fraseología en las sentencias penales: un estudio contrastivo español, italiano, inglés basado en corpus*. PhD thesis, Università degli studi di Trieste.

SANZ, Z., 2015. *Unitate fraseologikoen itzulpena: alemana euskara. Literatur testuen corpusean oinarritutako analisia*. Servicio Editorial de la UPV/EHU. [online] Available at: https://addi.ehu.es/handle/10810/15128

TOURY, G., 2012. *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

NLP AND/OR CORPUS-BASED IDENTIFICATION AND CLASSIFICATION OF PHRASEOLOGICAL UNITS

IDENTIFICACIÓN Y CLASIFICACIÓN DE UNIDADES FRASEOLÓGICAS BASADA EN CORPUS O MEDIANTE TÉCNICAS DE PLN

TRADUIRE DES IDIOMES FRANÇAIS EN LANGUE ÉTRANGÈRE (ALLEMAND, ESPAGNOL) : TRAITEMENT COGNITIF ET STRATÉGIES D'INTERPRÉTATION

Mariangela Albano
Université de Bourgogne et Université Sorbonne Nouvelle – Paris 3
albanomariangela@gmail.com

Résumé

Cette étude analyse d'un point de vue de linguistique acquisitionnelle les éléments qui participent à la traduction et à l'interprétation des idiomes français par des apprenants germanophones et hispanophones adultes de Français Langue Étrangère. Nous nous proposons de souligner l'importance d'une telle approche pour l'étude de l'apprentissage du FLE en tenant compte des enquêtes menées par la linguistique cognitive. Cette recherche a pour but d'apporter une contribution à l'analyse du traitement sémantique des idiomes.

1. INTRODUCTION

Cette recherche porte sur le traitement sémantique de 20 idiomes français par des apprenants germanophones et hispanophones adultes de Français L2. Cette étude est le résultat d'une première enquête pilote qui naît au sein d'une thèse doctorale en FLE à l'Université Sorbonne Nouvelle, Paris 3. L'enquête pilote résulte de l'observation des interviews menées à des étudiants italiens de FLE et d'Allemand Langue Étrangère. Ces interviews, utilisées comme modèle, représentent le résultat d'un travail abouti avec ma collègue Madame Rosa Leandra Badalamenti pour la rédaction de deux articles à paraître en 2015.

Pour cette étude nous avons choisi de limiter les interviews à 20 apprenants germanophones et hispanophones avancés de Français L2. Premièrement nous avons effectué un choix sur la base de l'instruction et de la formation des étudiants : apprentissage du Français L2 pendant 8/11 ans ; étudiants en Licence 3 ou Master 1 ; étudiants Erasmus en France pendant 3/6 mois ; ayant un niveau européen du FLE correspondant à B2/C1. Afin de répondre à ces paradigmes, nous avons choisi d'interviewer 10 étudiants allemands Erasmus à l'Université de Bourgogne et 10 étudiants espagnols en FLE à l'Universidad de Las Palmas de Gran Canaria.

Deuxièmement, la démarche adoptée pour tester notre hypothèse de travail a été d'enregistrer les interviews (durée 40-60 minutes). Les interviews contiennent des instructions de base qui amènent les étudiants à interpréter, à traduire les textes qui sont brefs et de nature différente (articles de presse, récits, blogs, forums, romans) et à motiver leur interprétation. Du point de vue du corpus, nous avons adopté les études phraséologiques et nous avons concentré notre étude sur 20 idiomes présentant une certaine opacité sémantique, non traduisibles littéralement en langue maternelle (allemande et espagnole) et donnés en contexte.

Pour analyser la traduction des idiomes, nous avons abordé une perspective à la fois acquisitionnelle et cognitive. En particulier, l'approche acquisitionnelle nous amène à comprendre les processus d'identification et d'isolement et l'approche cognitive nous conduit à analyser le traitement sémantique des expressions figées françaises.

2. FIGEMENT ET APPRENTISSAGE : UNE APPROCHE ACQUISITIONNELLE ET COGNITIVE DES IDIOMES

Avant d'analyser le traitement des idiomes que nous avons pris en considération dans cette étude, nous allons faire un examen synthétique de ce que représente actuellement le figement dans le domaine des études sur la phraséologie. Quand bien même il y ait eu, auparavant, certaines tentatives pour le définir, le figement a été considéré comme une tendance mécanique qui règle une bonne quantité du système linguistique (De Saussure, 1967 [1916]). Il est évident, donc, que la lexicalisation soit inévitable dans chaque communauté linguistique (Gross, 1996) et qu'elle constitue un procès dynamique qui s'insinue dans la langue par l'usage (Greciano, 1982 et 1991; Mejri, 1997). Dans ce sens, il est utile de noter que Wray (2000: 465) définit une « formulaic sequence » comme « une séquence, continue et discontinue, de mots ou d'autres éléments signifiants qui est, ou semble être préfabriquée : qui est stockée et récupérée entièrement par la mémoire dans le moment de l'usage, plutôt qu'être sujette à la production ou à l'analyse par la grammaire de la langue »[24]. Toutefois, selon Mejri (1998; 1999; 2000), le figement représente un phénomène impliquant toutes les dimensions du système linguistique du moment que chaque expression figée est une séquence de constituants en origine libre qui s'est fossilisée à travers des transformations morphologiques, phonétiques et orthographiques et a donné lieu à un signifié global. En fait, la littérature scientifique sur le figement nous aide à comprendre que l'identification d'une expression lexicalisée dépend essentiellement de critères suivants : le bloc grammatical ou de ses restrictions syntactiques (Hudson, 1998: 9) ; le blocage syntaxique ou sémantique (Gross, 1996 : 154)[25] ; sa conventionalité (Nunberg et al., 1994 : 493)[26]. Au de là des critères, la littérature phraséologique (Hudson, 1998; Moon, 1998 : 19-25 ; Norrick, 1985 : 72 ; Brinton et Traugott, 2005) s'est interrogée sur les

[24] Orig. angl.: «a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar» (Wray, 2000 : 465).

[25] Gross (1996 : 154) a affirmé qu'une expression figée du point de vue syntaxique ne présente pas d'alternative combinatoire ou transformationnelle ; au contraire, une expression figée du point de vue sémantique fait émerger une énonciation opaque ou non compositionnelle.

[26] Nurberg et al. (1994 : 493) témoignent de cette définition lorsqu'ils affirment que « apart from the property of conventionality, none of these properties applies obligatorily to all idioms ».

typologies des expressions figées comme l'idiome, la collocation, la locution, le proverbe, le phrasème et le gallicisme. Toutefois dans cette étude nous focaliserons notre attention sur l'idiome. Normalement l'idiome est considéré comme une expression relativement figée « whose meaning does not reflect the meanings of its component parts » (Benson, 1985 : 66). Selon Nunberg et al. (1994 : 492) les idiomes « are conventionalized; their meaning or use can't be predicted, or at least not entirely predicted, on the basis of a knowledge of the independent conventions that determine the use of their constituents when they appear in isolation from one another », fr. « sont conventionnels ; leur signifié ou usage ne peut pas être prédit, ou au moins pas entièrement prédit, sur la base d'une connaissance de conventions indépendantes qui déterminent l'usage de leurs constituants quand ils apparaissent isolés l'un de l'autre ».

Le statut des idiomes considérés comme équivalant à des mots uniques, c'est-à-dire « à des suites inanalysables en leurs éléments » (Gaatone, 1993 : 39) pose des problèmes pour l'appropriation d'une langue seconde ou étrangère (González Rey, 2002 : 14 ; Mel'čuk, 1993 ; Nesselhauf, 2005 ; Colson, 2000). En fait, les apprenants du FLE n'ont pas accès direct à la dimension figurée des idiomes (Gibbs, 1986) et ils doivent s'efforcer de comprendre l'expression figée à travers plusieurs efforts qui font appel à des stratégies d'interprétation et traduction. Généralement, les apprenants, percevant une certaine opacité sémantique, essayent d'isoler l'idiome du contexte dans lequel il est placé. Après les processus d'identification et d'isolement, ils abordent des stratégies d'interprétation qui se fondent sur la référence au contexte, sur les analogies avec leur langue maternelle ou d'autres langues étrangères et sur la décodification de l'unité lexicale. Les résultats de telle démarche donnent lieu à des emprunts, à des traductions littérales, à des paraphrases et, parfois, certains étudiants utilisent un idiome dans leur langue maternelle visant à rehausser l'expressivité de la traduction. Pendant le processus de traduction, les apprenants réactivent la sémantique des idiomes. En fait, la compréhension des unités figées est généralement produite grâce à des opérations cognitives qui peuvent montrer une nature analogique, métaphorique ou métonymique. Dans ce contexte, il semble essentiel d'analyser les résultats à travers la sémantique cognitive menée par Lakoff et Johnson (1980 et 1999) et par Fauconnier et Turner (2002). En particulier, nous pouvons affirmer, sur la base de la première théorie (Lakoff et Johnson, 1980 et 1999), que la métaphore représente un transfert de projections, structurées selon des correspondances ontologiques et épistémiques, d'un domaine conceptuel d'origine à un domaine d'arrivée. Cette considération conduit à réfléchir sur le fait que l'interprétation des idiomes laisse émerger des métaphores conceptuelles qui sont des groupes d'informations, acquis soit à travers l'expérience physico-perceptive directe soit à travers la culture, qui produit une quantité innombrable d'expressions linguistiques métaphoriques. Similairement, la théorie de l'intégration conceptuelle, ou du soi-disant « blending » (Fauconnier et Turner, 2002), tente de fournir une explication de ce qui se produit parmi des structures saillantes dans le déroulement de la construction du signifié. Nous pouvons observer des opérations de « blending » pendant l'interprétation d'un idiome parce que les apprenants essayent d'identifier des structures relationnelles similaires dans des contextes différents et de transférer l'information d'un contexte à un autre.

3. LA TRADUCTION DES IDIOMES FRANÇAIS : ANALYSE DU TRAITEMENT COGNITIF EN ALLEMAND ET EN ESPAGNOL

L'analyse proposée dans cette étude vise à décrire l'anatomie du traitement d'un idiome à l'aide d'une approche de linguistique cognitive. Dans les paragraphes précédents nous avons brièvement montré le cadre méthodologique et théorique ; différemment ce

paragraphe vise à donner des exemples du traitement cognitif mené par les apprenants allemands et espagnols. Le nombre des exemples ne sera pas exhaustif à représenter l'ensemble des résultats ; toutefois nous essayons ici de proposer une étude préalable des hypothèses ayant cours dans le domaine de l'acquisition langagière des idiomes.

Les résultats de notre étude ont montré que les apprenants utilisent trois stratégies d'interprétation : référence au contexte, analogies avec leur langue maternelle ou d'autres langues étrangères et décodification de l'unité lexicale à travers des processus métaphorique, métonymique ou analogique. En allemand, par exemple, une étudiante a traduit l'idiome fr. « Mon Ipad a cassé sa pipe », litt. « Mon Ipad est mort » avec l'expression figée allemande « Mein Ipad hat den Löffel abgegeben », fr. « Mon Ipad a remis la cuillère », litt. « Mon Ipad est mort ». Cette traduction montre une double stratégie interprétative : d'abord l'étudiante a fait référence au contexte traduisant le passage donné du français à l'allemand ; après avoir identifié l'idiome comme une expression opaque, elle cherche de la traduire en utilisant une expression figée similaire allemande. Une traduction similaire a été abordée par une autre étudiante, laquelle traduit « Mein Ipad hat den Geist aufgegeben », fr. « Mon Ipad a remis l'âme », litt. « Mon Ipad est mort » à l'aide du contexte.

En espagnol, l'apprenant traduit la phrase « Le Renégat numérote ses abattis » en utilisant la référence au contexte et une analogie avec sa langue maternelle. La traduction espagnole « El Renegado enumera sus victimas », fr. « Le Renégat numérote ses victimes », litt. « Le Renégat se prépare à une lutte » (Rey et Chantreau, 2003 [1989] : 1-2) montre que l'apprenant utilise le mot « victimes » au lieu de le traduire avec un synonyme de « abattis ». La justification de ce choix réside, en premier lieu, dans le contexte et, en deuxième lieu, dans la similarité entre le mot « abattis » et le verbe espagnol « abatir », fr. « abattre ». Selon l'apprenant, le mot « abattis » lui évoque les sujets qui subissent l'action du verbe « abattre » et, par conséquent, des victimes.

Un autre cas montrant similarité avec la langue maternelle est représenté par l'idiome « il doit être piqué par une aiguille phono », litt. « il doit être très bavard » (Ibid. : 10). L'étudiant traduit l'expression en espagnol « el debe ser picado de una inyección de una aguja de tocadiscos », fr. « il doit être piqué par une aiguille de tourne-disque » en utilisant le plus récent « tourne-disque » au lieu du « phonographe ».

Il est également intéressant dans cette étude de souligner les exemples qui concernent la compréhension de l'unité idiomatique à travers des mécanismes métaphorique, métonymique ou analogique. Nous avons observé que les trois mécanismes se mélangent pendant le processus d'interprétation d'un idiome. En allemand, par exemple, un étudiant a traduit la phrase « ici les mecs n'ont pas la grosse tête », litt. « ici les mecs ne sont pas prétentieux » en abordant l'expression allemande « hier haben die Jungs nicht so viel im Kopf », fr. « ici les mecs n'ont pas trop de choses en tête ». L'étudiant a évidemment pris en considération le contexte où les sujets sont des footballers qui s'engagent pour l'équipe mais l'interprétation abordée montre un ancrage à un raisonnement métaphorique et métonymique. En fait, l'étudiant nous a expliqué qu'un premier lieu la tête symbolise le corps tout entier ; en deuxième lieu, les footballers pensent seulement au football et ils n'ont pas d'autres projets dans la tête et, par conséquent, le volume de leur tête est proportionnel à la quantité de pensées. En raison des études de Lakoff et Johnson (1980 et 1999), nous pouvons affirmer que dans ce cas la tête est perçue comme un récipient qui peut contenir des objets concrets représentés par les pensées. Le recours à cette typologie de métaphore conceptuelle est visible dans plusieurs exemples où le corps est le sujet de la phrase idiomatique.

En espagnol, un étudiant a fait recours à des stratégies d'analogie et d'intégration conceptuelle. La phrase « l'homme serait un ours mal léché », fig. « l'homme serait rustre »

est traduite « ser un poco sucio », fr. « être un peu sale ». Ce cas d'interprétation nous montre l'influence d'une idée figée sur l'ours et l'homme : l'ours est considéré sale et, par conséquent, l'homme est sale. La condensation des images concernant la férinité humaine développe dans d'autres cas analysés des interprétations comme, par exemple, « faire l'ours » dans le sens de « être timide » ou « être asocial » jusqu'à prendre en considération l'idée de « être trompeur ».

4. CONCLUSION

Même si les exemples proposés dans le paragraphe précédent ne montrent pas l'ensemble des données, nous pouvons déjà affirmer que, à travers les éléments que nous avons analysés, les idiomes semblent s'installer dans une grammaire de métaphores motivées culturellement. En fait, les apprenants, après un procès d'élimination de l'ambiguïté de l'idiome sur la base du contexte, essayent de comprendre les réseaux sémantiques des unités idiomatiques. L'opacité de l'expression amène les apprenants à formuler des hypothèses à travers la juxtaposition des processus analogiques, métaphoriques ou métonymiques. L'explicitation de la sémantique sous-jacente à l'idiome permet à l'apprenant de construire dialectiquement un parcours gnoséologique qui vise à comprendre la motivation de l'idiome. Cette stratégie sert à renforcer la technique interprétative pendant le travail de traduction et à valoriser le traitement des idiomes pendant l'apprentissage d'une langue-culture.

Le répertoire pris en considération nous a permis aussi d'observer les différences entre les étudiants allemands et espagnols du Français Langue Étrangère. Nous avons remarqué que les apprenants allemands et espagnols partagent la tendance à faire recours au contexte mais pour réfuter ce résultat il faudra soumettre une typologie différente d'interview. Toutefois il semble intéressant que dans la majorité des cas les apprenants allemands focalisent leur attention sur la traduction mot-à-mot de l'expression figée ; par contre les étudiants espagnols essayent de traduire en considérant le répertoire langagière de leur langue maternelle. Il ressort de l'observation de deux cas que les apprenants ont la tendance à construire une analyse qui développe une activation de leurs connaissances linguistiques et culturelles de la langue cible.

Références bibliographiques

BENSON, M., 1985. Collocations and idioms. Dans : R. Ilson, éd., 1985. *Dictionaries, Lexicography and Language Learning* VIII. Oxford : Pergamon. pp. 61-68.

BRINTON, L. J. ET TRAUGOTT E. C., 2005. *Lexicalization and Language Change*. Cambridge : CUP.

COLSON, J.-P., 2000. Les locutions verbales françaises et allemandes dans le discours journalistique : pistes de recherche, fausses pistes, pistes brouillées (I. Pistes pour le français). Dans G. Greciano, éd., 2000. *Micro- et macrolexèmes et leur figement discursif*. Louvain-Paris : Peeters. pp. 173-199.

FAUCONNIER, G. ET TURNER, M., 2002. *The Way We Think*. New York : Basic Books.

GAATONE, D., 1993. Les locutions verbales et les deux passifs du français. *Langages*, 27ème année (109). pp. 37-52.

GIBBS, R. W. JR., 1986. Skating on thin ice : Literal meaning and understanding idioms in conversation. *Discourse Processes*, 7. pp. 17-30.

GONZALEZ REY, I., 2002. *La phraséologie du français*. Toulouse : Presses Universitaires du Mirail.

GONZALEZ REY, I., 2007. *Les expressions figées en didactique des langues étrangères*. Fernelmont : E.M.E..

GRECIANO, G., 1982. *Signification et dénotation en allemand. La sémantique des expressions idiomatiques.* Thèse d'état. Paris-Sorbonne : Klincksieck.

GRECIANO, G., 1991. Valence, version intégrée. *L'information grammaticale* 50. pp. 13-19.

GROSS, G., 1996. *Les expressions figées en français; noms composés et autres locutions*. Paris : Éditions Ophrys.

HUDSON, J., 1998. *Perspectives on fixedness: applied and theoretical.* Lund : Lund University Press.

LAKOFF, G. ET JOHNSON, M., 1980. *Metaphors we live by*. Chicago : The University of Chicago Press.

LAKOFF, G. ET JOHNSON, M., 1999. *Philosophy in the flesh : the embodied mind and its challenge to Western thought*. Chicago-London : University of Chicago Press.

MEJRI, S., 1997. Défigement et jeux de mots. *Études linguistiques*, 3. Tunis. pp. 75-92.

MEJRI, S., 1998. Structuration sémantique et variation des séquences figées. Dans : S. Mejri, M. Gross, A. Clas et T. Baccouche, éd., 1998. *Le figement lexical.* Actes des Premières Rencontres Méditerranéennes, les 17-18 et 19 septembre 1998. Tunis : C.E.R.E.S.

MEJRI, S., 1999. Unité Lexicale et polylexicalité. Dans : G. Petit, éd., Linx, n°40, *Le statut d'unité lexicale*. Université de Paris X-Nanterre.

MEJRI, S., 2000. Figement et dénomination. *Meta, XLV-4*. Montréal : Les Presses de l'Université de Montréal.

MEL'CUK, I., 1993. La phraséologie et son rôle dans l'enseignement/apprentissage d'une langue étrangère. *Étude de Linguistique Appliquée,* 92. pp. 82-113.

MOON, R., 1998. *Fixed expressions and idioms in English*, *a corpus-based approach*. Oxford : Clarendon Press.

NESSELHAUF, N., 2004. How learner corpus analysis can contribute to language teaching : A study of support verb constructions. Dans : G. Aston, S. Bernardini, D. Stewart, éd., 2004. *Corpora and Language Learners*. Philadelphia/Amsterdam : Johns Benjamins/Studies in Corpus Linguistics. pp. 109-124.

NORRICK, N. R., 1985. *How proverbs mean: semantic studies in English proverbs*. Berlin : Mouton.

NUNBERG, G., SAG, I. A. ET WASOW, T., 1994. Idioms. *Language,* 70, 3. Washington DC : Linguistic Society of America. pp. 491-538.

SAUSSURE, F. DE. 1967 [1916]. *Cours de linguistique générale, édition critique préparée par Tullio de Mauro*. Paris : Éditions Payot & Rivages.

WRAY, A. 2000. Formulaic Sequences in Second Language Teaching : principle and Practice. *Applied Linguistics*, 21 (4). Oxford : Oxford University Press. pp. 463-489.

WRAY, A. 2002. *Formulaic language and the lexicon*. Cambridge, UK : Cambridge University Press.

Dictionnaires

IMBS, P., 1974. *Trésor de la langue française. Dictionnaire de la langue du XIX et du XX siècle (1789-1960). Tome troisième*. Paris : Éditions du Centre National de la Recherche Scientifique.

IMBS, P., 1979. *Trésor de la langue française. Dictionnaire de la langue du XIX et du XX siècle (1789-1960). Tome septième*. Paris : Éditions du Centre National de la Recherche Scientifique.

GORCY, G., 1994. *Trésor de la langue française. Dictionnaire de la langue du XIX et du XX siècle (1789-1960). Tome seizième*. Paris : Gallimard.

POTTIER, B., 1983. *Trésor de la langue française. Dictionnaire de la langue du XIX et du XX siècle (1789-1960). Tome dixième*. Paris : Éditions du Centre National de la Recherche Scientifique.

POTTIER, B., 1985. *Trésor de la langue française. Dictionnaire de la langue du XIX et du XX siècle (1789-1960). Tome onzième*. Paris : Gallimard.

RAT, M., 1999 [1957]. *Dictionnaire des expressions et locutions traditionnelles*. Paris : Larousse.

POTTIER, B., 2007 [1957].Dictionnaire des expressions et locutions traditionnelles. Paris : Larousse.

REY, A. ET CHANTREAU, S., 2003 [1989]. *Dictionnaire d'expressions et locutions*. Paris : Les Usuels du Robert.

ROBERT, P., 2007. *Le nouveau petit Robert de la langue française 2007 : dictionnaire alphabétique et analogique de la langue française*. Paris : Le Robert.

Pages web

BIBLIOBABIL, 2011. *Mon Ipad a cassé sa pipe*. En ligne http://bibliobabil.com/2011/05/16/mon-ipad-a-casse-sa-pipe/ [Consulté le 15 juillet 2015].

COMMUNE DE BELLONNE, 2012. *L'ancêtre du G.P.S. à Bellonne*. En ligne http://www.bellonne.fr/pageLibre000103d8.html [Consulté le 15 juillet 2015].

FUTURA FORUM, 2004. *Que se passe-t-il quand on tombe dans les pommes ?* En ligne http://forums.futura-sciences.com/sante-medecine-generale/156930-se-passe-t-on-tombe-pommes.html [Consulté le 15 juillet 2015].

L'ILE DE CRETE, 2006. *Voyage en images. Forum des voyageurs. Re : Athena Palace*. En ligne http://www.ile-de-crete.com/forums/read.php?1,11681,12095 [Consulté le 15 juillet 2015].

LE FIGARO.FR., 2011. *Abidal : «Au Barça, on ne se regarde pas le nombril»*. En ligne http://www.lefigaro.fr/sport/2011/03/07/02001-20110307ARTFIG00665-abidal-au-barca-on-ne-se-regarde-pas-le-nombril.php [Consulté le 15 juillet 2015].

LE FIGARO.FR., 2011. *Les républicains se cherchent un champion contre Obama*. En ligne http://www.lefigaro.fr/international/2011/05/06/01003-20110506ARTFIG00615-les-republicains-se-cherchent-un-champion-contre-obama.php [Consulté le 15 juillet 2015].

LE FIGARO.FR., 2012. *François Hollande en tête-à-tête avec la reine Elizabeth*. En ligne http://www.lefigaro.fr/international/2012/07/10/01003-20120710ARTFIG00591-hollande-a-discute-en-tete-a-tete-avec-la-reine-elizabeth.php?page=&pagination=17 [Consulté le 15 juillet 2015].

PLURIELLES.FR., 2012. *Beauté homme : quels soins basiques pour ses premiers pas ?* En ligne http://www.plurielles.fr/beaute/soins/beaute-homme-quels-soins-basiques-pour-ses-premiers-pas-7219330-402.html [Consulté le 15 juillet 2015].

IMPLEMENTING EUROPEAN PORTUGUESE VERBAL IDIOMS IN A NATURAL LANGUAGE PROCESSING SYSTEM

Jorge Baptista
Univ. Algarve
L2F/INESC-ID
Lisboa
jbaptis@ualg.pt

Graça Fernandes
Rui Talhadas
Univ. Algarve
{gracitafernandes,
rtalhadas}@gmail.com

Francisco Dias
IST/Univ. Lisbon
L2F/INESC-ID
Lisboa
francisco.m.c.dias
@tecnico.ulisboa.pt

Nuno Mamede
IST/ Univ.
Lisbon
L2F/INESC-ID
Lisboa
Nuno.Mamede
@inesc-id.pt

Keywords: European Portuguese, verbal idioms, lexicon-grammar, rule-based parsing, natural language processing

Abstract

This paper is based on an extant *lexicon-grammar* of European Portuguese verbal idioms (*e.g.*, *deitar mãos à obra*, literally, 'to throw hands to the work', 'to start working'.). This a database containing about 2,400 expressions, along with all relevant information on the sentence structure, distributional constraints and transformational properties of these frozen sentences. In this paper, we present a solution to the integration of verbal idioms in a fully-fledged natural language processing system, and a preliminary evaluation using a small, manually annotated corpus.

1. INTRODUCTION

Verbal idioms (e.g. smb. *kill two birds with one stone*) can be defined as frozen sentences where the verb and at least one of its arguments are frozen together, and their overall meaning cannot be derived from the mere composition of the meanings of their individual elements, when used separately (Gross 1982, 1996; Cowie 1998). They constitute a large set of the lexicon and grammar of many languages, in the order of several thousands, though their frequency in texts is often very low. In fact, their occurrence it is highly dependent on text genre/type, being more common in oral than in written texts. The integration of verbal idioms in natural language processing (NLP) systems is relevant for an accurate semantic parsing. However, this integration is a challenge to NLP systems (Sag *et al.* 2002), as these idioms cannot be dealt with like other idioms, as frozen strings of words. In spite of their being semantically non-compositional, they do have syntactic structure, allowing inflection, insertions, several transformations and creative pragmatic reuse.

In this paper, we present a solution to the integration of verbal idioms in a fully-fledged natural language processing system, STRING (Mamede *et al.* 2012)[27]. This work is based on an

[27] string.l2f.inesc-id.pt

extant lexicon-grammar of European Portuguese verbal idioms (Baptista *et al.* 2004, 2005), containing about 2,400 expressions, e.g. *deitar mãos à obra* (lit: throw hand to work) 'start working'. Conceived in a tabular format, and organized in 10 main classes, according to the formal structure of the idioms, these tables present the verb and the frozen arguments of each idiom, along with the encoding of distributional constraints on free arguments and the structural changes (or transformations) the sentence can undergo (passive, pronominalization, etc.). Each idiom is also illustrated by an example.

In order to integrate the lexicon-grammar of verbal idioms in rule-based parsing module of the NLP system the following strategy was adopted: firstly, the general parsing rules are applied, so the frozen sentence is given a structure as any ordinary sentence; then, another set of rules extracts a (semantic) dependency (FIXED), based on the previous parse, and groups together the frozen elements of the idiom, while keeping intact the syntactic structure of the sentence; finally, FIXED dependency is then used to further calculate the sentence's semantics: for example, the semantic roles of the verb's free argument are to be extracted not from the information attached to the simple verb but to those of the verbal idiom; the part-whole semantic relations extraction is blocked; the verb sense statistical disambiguation module is prevented from acting, and so on. A script automatically reads the tabular format and converts it into the syntax of the rule-based parser for the extraction of the FIXED dependency. To assess the conversion process, the set of rules was applied to the examples provided in the lexicon-grammar. A small percentage of errors were detected and some rules were manually adjusted. Most errors, though, were due to incorrect part-of-speech (POS) tagging.

To evaluate the system's new module, sentences including all the key-elements of each idiom were extracted. Then, a team of linguists manually annotated a representative sample, taken from a freely available corpus, in order to build a golden standard. Then the sentences were parsed and results were automatically compared to the golden standard. A detailed error analysis is also briefly presented.

This paper is structured as follows: Section 2 briefly presents the lexicon-grammar of European Portuguese verbal idioms. Section 3 described the construction of an annotated reference corpus of verbal idioms. Section 4 describes the implementation of the verbal idioms' identification module in the STRING system, while Section 5 presents and briefly discusses the evaluation of the module using the annotated corpus. Finally, section 6 concludes the paper and suggests future work.

2. A LEXICON-GRAMMAR OF EUROPEAN PORTUGUESE VERBAL IDIOMS

Frozen sentences or *verbal idioms* can be defined (Gross 1982, 1996) as sentences where the verb and at least one of its argument noun-phrases are frozen together, that is, they are distributionally constraint. In a free sentence, the meaning is determined from the individual meaning of the elements in the construction, but, the meaning of the frozen sentence is non-compositional, *i.e.* it cannot be directly calculated from the meaning that the component elements may present when used separately. For example, in *brincar com o fogo* (lit: "to play with fire"), neither the verb *brincar* 'to play' nor the noun *fogo* 'fire' can be substituted by any other word, unless a change is seen in the sentence's overall meaning: 'to do dangerous, risky things'.

The structure of these type of sentences is similar to that of free sentences, but syntactic and distributional constraints cannot be calculated or predicted from the formal properties that the elements of expression may have when constructed separately. For example:

(1) *O Pedro saiu* (*do armário* + °*da sala* + °*da loja*)

literally: 'Peter left/exited [from] the closet/room/shop'

'to assume one's (homo)sexuality'

(2) *A Maria fechou-se em* (*copas* + **paus* + **espadas*)

literally: 'Mary closed herself in hearts/clubs/spears'

'to be silent, not to disclose information'

In (1), the verb *sair* 'leave/exit' requires a locative-source complement, which the noun *armário* 'closet' can fill in, but in the frozen idiomatic interpretation only this noun can occur, otherwise the sentence's interpretation becomes literal (signaled by '°'). In (2), the noun *copas* 'hearts' cannot be replaced by any other deck of cards, and it does not corresponds to the ordinary distribution of the reflexive use of the verb *fechar* 'close'.

The European Portuguese verbal idioms were classified into 10 formal classes according to their structure and distributional constraints. The theoretical and methodological framework here adopted is the Lexicon-Grammar (M. Gross 1982, 1996), based on the Harrissian transformational operator-grammar (Z. S. Harris 1991). Table 1 shows the structure of each class studied in this work[28].

Class	Structure	Example	Count
C1	N_0 *V* C_1	*O Pedro bateu o pé*	
		to beat the foot 'to refuse to do smthg'	491
CDN	N_0 *V* (*C de N*)$_1$	*O Pedro aprendeu o bê-á-bá da gramática*	
		'to learn the basic concepts of smthg'	45
CAN	N_0 *V* (*C de N*)$_1$ = C_1 *a* N_2	*O Pedro cortou as asas* (*do João* = *ao João*)	
		to cut the wings to/of smbd	
		'to prevent smb from acting freely'	175
CNP2	N_0 *V N1 Prep*$_2$ C_2	*O Pedro não tirava a Ana da cabeça*	
		not to take smbd/smthg from the head	
		'think continuously on smb/smthg'	172
C1PN	N_0 *V* C_1 *Prep*$_2$ N_2	*A Rita cravou as unhas na fortuna do João*	
		to dig the nails into smthg	
		'to acquire/steal smthg'	233
C1P2	N_0 *V* C_1 *Prep* C_2	*O Pedro deu o dito pelo não dito*	
		to give the said for the non-said	
		'to change one's opinion, not to be true to one's word'	288
CPPN	N_0 *V* C_1 *Prep* C_2 *Prep* C_3	*O Pedro deitou fora o bebé com a água do banho*	
		to throw away the bay the bath water	26

[28] The code for each class is purely conventional; *N* and *C* stand for noun phrases; *N* is a free constituent and *C* is frozen noun phrase; N_0 is the subject, N_1 and N_2 the first and second complement; *V* is the verb and *Prep* a preposition. These codes and the defined classes are the same as the ones proposed initially by M. Gross (1982, 1996). The example (in *italics*) is followed by a literal translation and gloss.

		'to lose the important along with the non important'	
CPP	*N_0 V Prep C_1 Prep C_2*	*O Pedro deu com o nariz na porta*	
		to hit with the nose on the door	
		'to go somewhere in vain'	201
CP1	*N_0 V Prep C_1*	*O Pedro passou pelas brasas*	
		go through the embers	
		'take a nap'	703
CPN	*N_0 V Prep (C de N)$_1$*	*O Pedro não chega aos calcanhares do João*	
		not to reach the heels of smb	
		'not to be a match for smb'	95
		Total	2417

Table 1. Classification of Frozen Sentences of the European Portuguese.

The relevant linguistic information has been encoded in binary matrices, where each line corresponds to a lexical entry, and the columns contain either the lexical elements of the expression or the signs '+' and '-' to encode the relevant linguistic properties it presents. Properties include the distributional constraints (human/non-human) of the free argument slots and certain transformations, such as *Passive* or obligatory complement permutation. For this paper, the most relevant properties are:

(a) *intrinsic reflexive constructions* (noted *V se*), where the verb presents a reflexive pronoun that can not be derived by pronominalizing a free noun phrase, *e.g.*

(3) *O Pedro pôs-se em bicos dos pés* (lit: 'Pedro put himself on the tip of the feet')

'to pretend exaggerated self-importance'

cp. **O Pedro pôs o João em bicos dos pés* (lit: 'Pedro put João on the tip of the feet');

(b) *obligatory negation constructions* (noted *NegObl*), where an expression can only be used in the negative:

(4) *O Pedro não dá para as encomendas* (lit: 'Pedro does not give to the requests')

'to be unable to meet the demands'

cp. **O Pedro dá para as encomendas.*

3. BUILDING AN ANNOTATED REFERENCE CORPUS OF VERBAL IDIOMS

A small corpus was prepared to evaluate the performance of the parser. This evaluation corpus was retrieved from the CETEMPúblico corpus (Rocha & Santos 2000), a publicly available corpus of journalistic text (dated from 1991 to 1998). The corpus is made of unrelated text extracts. For this paper only the first fragment of the corpus was used, featuring over 12 million words. For the evaluation corpus, the selection was made retrieving the extracts containing simultaneously the idioms' main verb and their frozen elements (prepositions and frozen NP head nouns) in the manner described below.

The UNITEX 3.0 linguistic development platform (Paumier 2003, 2014)[29] was used to retrieved these expressions from the CETEMPúblico fragment. UNITEX is based on finite-state automata (FSA) technology and it is able to intersect the linguistic data encoded in the lexicon-grammar matrices with finite-state transducers, which are then used to find patterns in texts and modify them, as well as to extract matching textual units (sentences) from larger texts. First, reference graphs are built, one for each class of idioms. Fig. 1 illustrates the reference graph for class **CP1**.

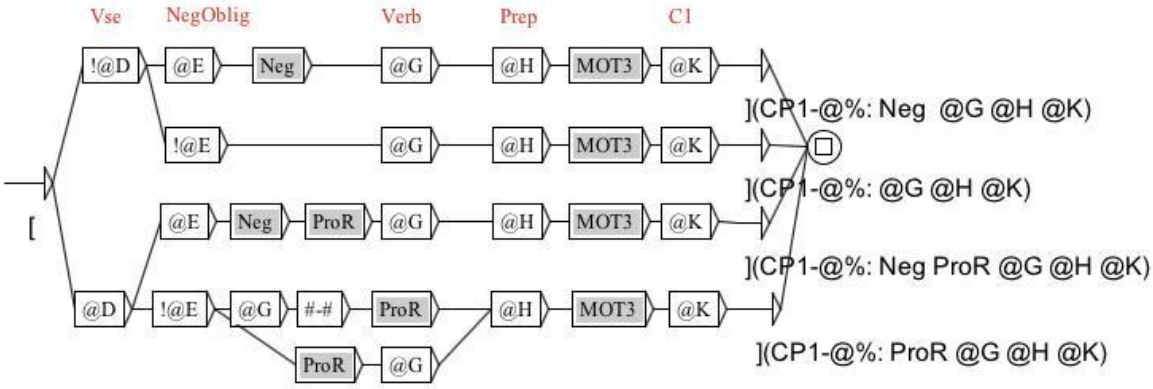


Figure 1. Reference graph for class CP1 (*e.g.* O Pedro passou pelas brasas, lit: Pedro passed through/over the burning coals 'Pedro took a nap')

These graphs refer to the data in the matrices by way of a set of variables @X, where 'X' stands for the corresponding column in the matrix. The variables can also be used as switches, allowing a path to be followed if the given cell contains a plus '+' sign or collapsing the path at that point if the cell contains a minus '-' sign. Switches can also be denied '!@X', which as the opposite effect. The graph reads as follows (symbols in red are just comments, to help reading the graph, and are not taken into account): variables @D and @E refer, respectively, to columns D and E of the lexicon-grammar matrix of idioms' class **CP1**, where the intrinsic reflexive and the obligatory negation properties, respectively, have been encoded. They function as switches. There can be four combinations of these property-value pairs, each with the corresponding path in the graph. Sub-graphs are represented by grey nodes: Neg includes the most common negation adverbs, ProR lists the reflexive pronouns and MOT3 is an insertion window from 0 to 3 words. Variables @G, @H and @K stand for the verb, the preposition and the head noun of the frozen complement. In the output, the matched is delimited by square brackets, while a simple tag is added after it, consisting of the class code, the line number in the matrix (variable @%) and the key elements of the idiom. An equivalent graph is hand-built for each class. The reference graph is then intersected with its respective matrix, reading it line-by-line and building an FST for each idiom (see Fig. 2). A general graph is then produced containing all the FSTs for that class.

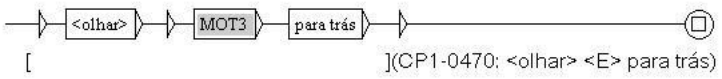


Figure 2. Finite-state transducer automatically built from the CP1 lexicon-grammar matrix for the idiom *olhar para trás* 'to look back (=into the past)' .

[29] www-igm.univ-mlv.fr/~unitex

The resulting transducer can thus be used to retrieve, delimit and tag the matching sequences from the corpus. Table 2 shows the breakdown of the raw 562 matching sequences per class and their distribution. There were 270 different expressions receiving different tags and their frequency ranged from a single instance (or *hapaxes*; 159 unique instances) to 42 (*v.g.* CAN-0045:<*chamar*> <DET> *atenção*); the difference between the number of matches and counts per class is due to the fact that some idioms could correspond to more than one lexical entry:

Class	**C1**	**CDN**	**CAN**	**CNP2**	**C1PN**	**C1P2**	**CPPN**	**CPP**	**CP1**	**CPN**	**Total**			
Matches	170	170	84	24	42	12	0	9	98	20	629			
Bin	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**25**	**42**	**Total**
Count	159	48	28	7	8	4	1	5	2	3	3	1	1	562

Table 2. Breakdown of the matching sequences per class and per frequency bins

The text units (sentences) matched by UNITEX were then manually annotated by 3 linguists, all well experienced in corpus annotation tasks and very familiar with the concepts involved in the study of idioms. A set of guidelines with examples was also provided and the annotation process was carried out independently. Each sentence could be tagged as follows:

tag	**description**
fixed :	the matched string corresponds to the targeted idiom;
fixed-different :	the matched string corresponds to an idiom but not to the targeted idiom (lexical elements involved can be slightly different, and/or the class is different); this happens because some idioms share the same lexical items, and usually the system chooses the longest match;
literal :	the matched string corresponds to the targeted idiom (eventually, it only partially corresponds), but this sequence of of words is being used in a literal, non-idiomatic way (e.g. *O Pedro deu um berro* 'Pedro gave a scream (='yell') vs. *O portátil deu o berro* 'The laptop broke');
false-positive :	the matched string contains elements of an idiom, but in that context it has nothing to do with the target idiom;
PoS-error :	the matched string includes an incorrectly PoS-tagged item;
other :	other problems not mentioned above.

Table 3 shows the distribution of the tags by the 3 annotators. The inter-annotator agreement was measured using ReCal 0.1 Alpha for 3+ Coders[30]. Table 4 shows the percentage of agreement between annotators:

tag	**Annotator 1**	**Annotator 2**	**Annotator 3**
other	0	15	33
fixed	399	373	398
fixed-different	37	41	55
false-positive	103	4	32
literal	63	158	86
POS-error	27	37	25

Table 3. Distribution of the tags by the annotators.

[30] dfreelon.org/recal/recal3.php

Average Pairwise Percent Agreement			
average	**annotators 1 & 3**	**annotators 1 & 2**	**annotators 2 & 3**
0.67	0.70	0.66	0.65
Average Pairwise Cohen's Kappa			
average	**annotators 1 & 3**	**annotators 1 & 2**	**annotators 2 & 3**
0.435	0.471	0.432	0.403

Table 4. Inter-annotator agreement.

The inter-annotator agreement can be considered as only moderate and it is similar pairwise; the Fleiss Kappa is 0.431, for an observed agreement of 0.671 and an expected agreement of 0.432, which is usually interpreted as 'fair'; the average pairwise Cohen Kappa is also deemed as 'fair'. This level of agreement may due to the difficulty of the task, particularly with the distinction between a literal and a figurative use, and the tag 'false-positive', which are very differently distributed among annotators, perhaps in a significant way. There were also an important number of cases where the annotators were unsure of the tag (and selected 'other').

A golden standard was established based on the consensual or most voted tag. In the end, 'PoS-error' and 'other' sentences were removed. The resulting corpus contains 602 sentences, 432 positive instances (noted 'fixed' and 'fixed-different') and 170 negative instances (false-positives and literal). The positive instances contain the target expression delimited and annotated for its class, ID number, key lexical elements (verb and frozen prepositions and head nouns). The negative instances contain just the tag 'non-fixed'.

4. IMPLEMENTING VERBAL IDIOMS IN THE XIP FORMALISM

Dealing with idioms in natural language processing systems is difficult, among other reasons, because their architecture must be conceived in such a way that it should not preclude the processing of both free word combinations and these, more constraint, expressions. On the other hand, many idioms do have syntactic structure, and can undergo several types of formal variation, thus making them hard to identify in a strictly string pattern-matching approach. Furthermore, many of these expressions are ambiguous between a literal (non-idiomatic) and figurative, non-compositional (idiomatic) use, depending of many linguistic and extra-linguistic factors.

In this section, we present the way European Portuguese verbal idioms have been integrated in STRING (Mamede *et al.* 2012), a hybrid, statistical and rule-based, fully fledged natural language processing system. STRING has a modular structure, shown in Fig. 3, that performs all the basic NLP tasks in four main steps: (1) preprocessing (tokenization, sentence splitting and lexical analysis; (2) rule-based and context-depending PoS disambiguation, MWE detection and context-depending contraction splitting; statistical PoS disambiguation; and (4) parsing, using the rule-based XIP parser (Xerox Incremental Parser, Ait-Moktar *et al.* 2002). Additional external modules operate on the output of XIP to perform other NLP-specific tasks, such as anaphora resolution, time normalization or slot filling.

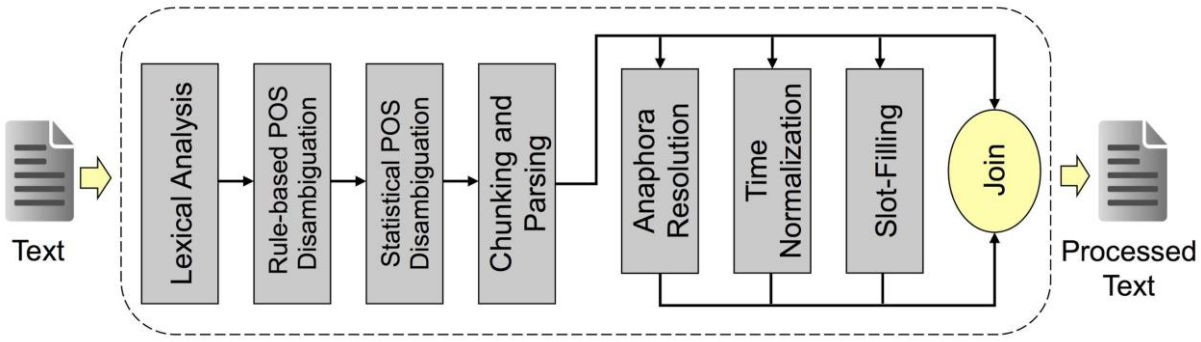


Figure 3. STRING architecture (Mamede *et al.* 2012)

The rule-based Portuguese grammar used by XIP was developed by the L2F (INESC-ID Lisboa) team in collaboration with Xerox and, besides adding to the lexicon some syntactic-semantic information required for the parsing stage, its processing consists of two main steps: (1) *chunking* (or *shallow parsing*), that is, the delimitation of elementary constituents (or chunks), such as NP, PP, etc.; and (2) *deep parsing*, that is, the extraction of syntactic dependencies between the chunks heads e.g. SUBJect, MODifier, CDIR (direct complement), *etc.*

Next, the general strategy for processing verbal idioms in STRING is laid out in order to better understand the automatic conversion process of the idioms' lexicon grammar matrices into the formalism XIP dependency rules.

Verbal idioms are identified in STRING by means of a dependency FIXED linking the key elements of the structure (the main verb, the prepositions and frozen head nouns). Below is an example of a sentence, the corresponding chunking and the (relevant) dependencies the parser extracted, including the FIXED dependency that identifies the idiom *medir as palavras*, literally 'to measure one's words', 'to be prudent when speaking':

```
SUBJ_PRE(mediu,Pedro)
CDIR_POST(mediu,palavras)
FIXED(mediu,palavras)
0>TOP{NP{O Pedro} VF{mediu} NP{as palavras} .}
```

This FIXED dependency is extracted based on the linguistic information already available in the system and the syntactic structure already calculated at the time the idioms identification module is applied. In this example, the verb and the direct object have been correctly PoS-tagged, the corresponding verb and noun phrase chunks properly identified, and a direct complement (CDIR) dependency between the verb and the head of the noun phrase has already been extracted by the time the idioms identification module comes into play. A dependency rule in this module then identifies the idiom by extracting the FIXED dependency:

```
if (  VDOMAIN(#?,#2[lemma:medir]) &
         CDIR[post](#2,#3[surface:palavras]) )
     FIXED(#2,#3)
```

This rule, which has been simplified here for clarity purposes, reads as follows: first a set of conditions (if) is stated, involving several variables; variables are identified by *#n* ; the first condition captures a verbal chain formed by a string of auxiliary verbs and a main verb, whose lemma is *medir* 'measure'; the next condition verifies if there is a CDIR dependency between the main verb and the noun *palavras*; finally, the dependency is extracted between the verb and the adverb.

This rule formalism requires that the information encoded in the lexicon-grammar matrices be converted into the XIP syntax. Writing rules for XIP grammar is a hard and time-consuming task. Filling a table is a much simpler task to a human than writing rules on XIP syntax. A faster approach for rule implementation could be to write an intermediary representation of the patterns. Then, this intermediary representation could be automatically converted into XIP rules. In order to represent these patterns, a table with the expected elements of the syntactic dependency links for each frozen expression was built. In this table each column represents a given element of a chunk of the syntactic dependencies structure. Any number of chunks can be concatenated to the table. An element of a chunk can be represented in such table in five different ways, which can be combined, as presented in Table 5:

Identification of the element	**Example**	**Description**
Surface	*andar*	Accepts only a word with *andar* 'walk' as surface
Lemma	<andar>	Accepts any inflection of lemma *andar* 'walk'
PoS tag	<DET>	Accepts any determinant (DET is the POS tag for determinant)
Dependency relation	MOD:<andar>	Word with lemma *and*ar 'walk' must be the chunk head which is a modifier (identified by the prefix MOD)
Grammatical features	<DET:ms>	Accepts any determinant with traces masculine and singular

Table 5. Intermediate representation of the information present in the lexicon-grammar.

The presence of a given element can be also negated using the identifier <E>. By default all modifiers are linked to the previous NP or PP (as in the case of *de N* 'of N' so-called determinative complements). That, however, is not always true due to the well-known PP-attachment ambiguity problem. To avoid PP-attachment ambiguity, a special flag (AttachV) can be placed to link a modifier directly to the verb of the sentence. Table 6 shows one entry and some of the columns of the lexicon-grammar for class CP1 and the idiom *ir desta para melhor* (lit: to go from this one to a better one, 'to die'):

SubjHum	Verb	Prep1	Det1	Mod1	Prep2	Det2	Mod2	AttachV2
+	<ir>	de	<E>	esta	para	<E>	melhor	+

Table 6. An idiom entry in the lexicon-grammar.

The resulting FIXED dependency is slightly different from the previous example:

```
FIXED_NORMALIZED(morrer,foi,esta,melhor)
```

In this case, the feature _NORMALIZED is added to the dependency, and a conventional classifying word, *morrer* 'to die' is added to its argument set. This information is encoded in the lexicon-grammar for certain types of semantic predicates that are relevant for the extraction of relations having to do with the major events in peoples biography, such birth and death, marriage or divorce, family ties, etc.

Any number of modifiers can be added in any order to the table. This allows any of the 10 classes of Portuguese frozen expression to be represented in the same table, by filling different elements for each class.

The system was first tested on the illustrative examples provided with the lexicon-grammar entries. These are artificial examples, where the idiom is expressed in the barest of forms, usually in the past or present tense, without any modifiers or adjuncts. Next, we describe the results of this preliminary test and the solutions adopted for the problems found. Table 7 presents the number of sentences used for testing and the errors found for each class of fixed expressions.

Class	Entries	Errors	% Error
C1	491	10	0.02
CDN	45	4	0.11
CAN	173	7	0.05
CNP2	172	22	0.13
C1PN	233	17	0.08
C1P2	288	26	0.09
CPPN	26	2	0.12
CPP	201	16	0.08
CP1	703	53	0.08
CPN	95	6	0.07
Total	2,417	163	0.07

Table 7. Error rate in the parsing of the lexicon-grammar's examples.

Overall, the percentage of correct cases is 93%, which constitutes a relatively high accuracy. It should be noted that the idioms identification module of the XIP parser Portuguese grammar is executed at a very late stage of parsing, thus it is hindered by all the errors that have accumulated in the pipeline up until then.

In the following, we investigate the main causes for the cases where the parser was unable to detect and extract the FIXED dependency. A preliminary classification of the error types was used to guide the assessment of this phase. Errors can be divided into 4 different types: (a) incorrect part-of-speech (PoS) tagging; (b) incorrect dependency extracted; (c) incorrect chunking; and (d) lexical gaps. Table 8 presents the breakdown of the types of errors, which are then discussed below.

Type of Error	Count
incorrect POS	62
incorrect dependency	47
incorrect chunking	41
lexical gaps	12
Total	162

Table 8. Distribution of the error types.

(a) incorrect PoS tagging

The STRING system automatically assigns a PoS tag to each token, using both a rule-based and a statistical PoS tagger. In average, the tagger achieves a state-of-the-art precision of 98%.

Naturally, like any other part of a text, frozen sentences can feature PoS-tagging errors. For example, in (1):

(1) *O João meteu baixa* (lit: 'João put a sick-leave') 'João called off sick'

the word *baixa* 'sick-leave' should have been classified as a noun, but it was tagged as an adjective, *baixa* 'short'. After this, the chunking module of the parser builds an AP chunk (adjectival phrase) instead of a NP, thus precluding the direct complement dependency (CDIR) from being extracted. As the frozen sentence dependency rule is built upon the extraction of the CDIR, the system fails to extract the FIXED dependency. This is the most frequently occurring error in the examples here tested. One of the possible solutions for this is to improve the rule-based PoS disambiguation, using contextual rules whenever the word combination is PoS-unambiguous.

(b) incorrect dependency extracted

The parsing performed by the STRING extract syntactic dependencies between the constituents of the sentence, using the syntactic and semantic features that have been added to the word in the lexicons. Since, at this time, word-sense disambiguation is only performed by the system regarding verbs (Travanca 2013, Suíssas 2014), certain ambiguous words can lead to an error when extracting dependencies. Sentence (2) is a good example of this:

(2) *O João tirou partido da notícia* 'João took advantage from that news'

Since the PoS-tagging and the chunking are correct, the relation CDIR should have been extracted between the verb *tirar* 'take' and the noun *partido* 'advantage'. However, *partido* is an ambiguous noun: besides this use as an abstract-uncountable noun, it also has the feature *group-of-things*, for it can be a human collective (a political party, for example). This feature triggers a QUANTD dependency to be extracted instead. This would correspond to an interpretation where *partido* would function as a quantifying determiner, like group of things, *e.g.*, as in the phrase *um partido de pessoas de esquerda* 'a party of left-wing people'. Therefore, *notícia* becomes instead the direct complement of *tirar*, which is incorrect. In order to capture this expression the noun *partido* should have been word-sense disambiguated first, removing its feature *group-of-things*.

(c) incorrect chunking tree

For each sentence being analyzed, STRING presents the resulting chunking tree. Errors may occur in this previous parsing task, which can lead to failure in recognizing the idiom. Consider sentence (3):

(3) *O João lava daí as suas mãos* (lit: João washes from there his hands)

'to wash one's hands (like Pilatus)'

In the lexicon-grammar, the idiom constituents are presented in their canonical order, which would be **O João lava as suas mãos daí*. However, in this idiom, the direct complement and the prepositional phrase are obligatorily reversed. This leads to the following chunking:

```
0>TOP{NP{O João} VF{lava} PP{de aí NP{as suas mãos}} .}
```

where a single PP is chunked incorrectly, integrating the adverb *aí* 'there' and the NP *as suas mãos* 'his hands'. Since the identification of the idiom depends on the previous correct identification of the two constituents, the CDIR {*as suas mãos*} and the MODifier PP {*de aí*}, the system fails to capture it. Chunking rules are a key element in the parsing process and are not to

be changed lightly, so in this cases, a manual rule has to be written, based on the idiom's (incorrect) chunking.

(d) lexical gaps

In spite of its large lexical coverage, STRING lexicons, especially multiword expressions, may still have some *lacunae*. The compound preposition *por debaixo* de in (4) was an example of such missing items:

(4) *O Pedro passou o dinheiro por debaixo do pano*.

(lit.: 'Pedro passed the money under the rug') 'Pedro made bribe'

In this case, it suffices to add the missing item to the lexicon, and the idiom can be properly identified. Other such cases involved compound nouns like *quadratura do círculo* 'circle's quadrature', (*discutir*) *o sexo dos anjos* '(to discuss) the angels' sex'.

During error analysis, some errors in the lexicon-grammar were also spotted (and corrected). For example, the codes for possessive determiners, as in *gastar a minha saliva* (lit: 'spend my salive') 'talking idly', were missing and had to be introduced in the appropriate cells, in order to capture such expressions. In other cases, a more general intervention in the grammar was required. For example, in the idiom *não ligar nenhuma a N* 'not minding nothing-fem.sg. to smthg' the indefinite *nenhuma* 'none' is obligatorily in the feminine-singular form and the verb prepositional (free) complement is introduced by *a* 'to'. However, a partitive quantifying determiner dependency was being extracted between *nenhuma* 'none' and the head noun of the PP. This should only happen in cases like *nenhum dos livros* 'none of the book', where gender agreement is required between the indefinite and the head of the PP and the preposition must always be *de* 'of'. In this case, the rule for extracting partitive quantifying determiner was made more precise.

5. EVALUATION AND DISCUSSION

The sample of 602 sentences that constitute our evaluation corpus were then processed by STRING and compared against the golden standard. Results are presented in Table 9:

Run	TP	FP	TN	FN	Precision	Recall	F-measure
1st	81	8	165	348	0.91	0.19	0.31
2nd	174	44	131	253	0.79	0.41	0.54

Table 9. Evaluation results.

The initial recall (1st run) was very low, though precision was high. We have undertaken the error analysis to understand the reasons for these suboptimal results. The main reason seems to be the fact that rules require a human subject to be explicit in the sentence. However, subject drop is a frequent phenomenon in Portuguese, so this condition should not be part of the rules (at least not at a first step). The same happens with several free complements (both NP and PP). After reformulating the script that generates the rules, when the system ignores the distributional constraints (2nd run) recall, though low improved to the double, while precision dropped 0.12.

Another reason for low results is the variation of the preposition *a/para* 'to', very common in the corpus, but missing altogether in the lexicon-grammar. The later preposition is also more common in Brazilian Portuguese. The annotators noticed a large number of sentences from the Brazilian Portuguese variant in the evaluation corpus, and, though there are significant differences between the idioms of each variant (Baptista 2008), this did not seem to have a significant impact in the results.

6. CONCLUSION AND FUTURE WORK

This paper presented the complex issues involved in the integration of a large-sized lexicon-grammar of European Portuguese verbal idioms into a natural language processing system, STRING, evaluating the resulting identification module of the system's parser on a manually annotated corpus. While precision was high, recall is pretty low: these results were hindered by the too restrictive rules, imposing the verification of distributional constraints on subject and complement, when these elements (especially the subject) can often be omitted in Portuguese. This is the single most important task to be completed in the near future. The evaluation has also shown that in the pipeline structure of the system, as errors tend to accumulate, some idioms are difficult to capture due to errors in previous processing stages, especially in PoS disambiguation and chunking. Some errors were also due to the incomplete word-sense disambiguation.

The paper showed that it is possible to identify idioms in texts, with a rule-based approach, based on lexical information and on the syntactic structure, as this has been previously calculated by the natural language processing chain, using the grammar of the general language. Thus, the system uses (and maintains) the syntactic structure of idioms (only a few cases present divergent syntactic constraints and require manual encoding).

In this way, the linguistic description is also kept apart from the processing issues. On one hand, this will benefit the continuing process of collecting the rich inventory of verbal idioms in the language. On the other hand, it will improve the semantic processing of Portuguese texts, as these meaning units are now better identified, affecting such disparate tasks as word-sense disambiguation, coreference resolution and semantic role labeling.

Acknowledgments

Research for this paper was partially funded by Portuguese national funds through Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2014.

References

AIT-MOKHTAR, S; CHANOD, J.; ROUX, C. 2001. Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering* 8 – 2/3: pp. 121-144.

BAPTISTA, J. 2004. Compositional vs. Frozen Sequences. Laporte, Eric; Ting Au-Chen, (eds). Proceedings of the Lexicon-Grammar Workshop. Beijing 14-18 de Outubro de 2004. *Journal of Applied Linguistics*, Special Issue on Lexicon-Grammar. Papers presented at the Lexicon-Grammar Workshop, pp. 81-93 (Chinese version).

BAPTISTA, Jorge; MAMEDE, Nuno; MARKOV, Ilia. 2014. Integrating verbal idioms into an NLP system. In: Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago Pardo, Maria das Graças Volpe Nunes, (Eds.). *Computational Processing of the Portuguese Language.* 11th International Conference

PROPOR'2014, São Carlos – SP, Brazil, October 8-10, 2014. Proceedings. pp. 251-256. *Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence* 8775: Berlin: Springer.

BAPTISTA, Jorge; CORREIA, Anabela; FERNANDES, Graça, 2004. Frozen Sentences of Portuguese: Formal Descriptions for NLP. *Workshop on Multiword Expressions: Integrating Processing, International Conference of the European Chapter of the Association for Computational Linguistics*, Barcelona (Spain), July 26, 2004. – ACL: Barcelona, pp. 72-79.

BAPTISTA, Jorge; CORREIA, Anabela; FERNANDES, Graça, 2005. Léxico Gramática das Frases Fixas do Portugués Europeo. *Cadernos de Fraseoloxía Galega* 7, Santiago de Compostela, Xunta de Galicia/Centro Ramón Piñero para a Investigación en Humanidades, pp. 41-53.

COWIE A., 1998. *Phraseology. Theory, analysis, and applications*. Oxford: Oxford University Press.

GROSS, M. 1982. Une classification des phrase "figées" du français. *Revue Québécoise de Linguistique* 11-2: pp. 151-185.

GROSS, M. 1996. Lexicon-Grammar. *Concise Encyclopedia of Syntactic Theories*. Cambridge. Pergamon. pp. 244-258.

HARRIS, Zellig S. 1991. *A Theory of Language and Information: A Mathematical Approach*. Oxford: Clarendon Press.

MAMEDE, Nuno; BAPTISTA, Jorge; DINIZ, Cláudio; CABARRÃO, Vera. 2012 STRING – An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. in Caseli, H.; Villavicencio, A.; Teixeira, A.; Perdigão, F. (Eds.) *Computational Processing of the Portuguese Language*, Proceedings of the 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. http://www.propor2012.org/demos.html.

PAUMIER, S. 2003. De la reconnaissance des formes linguistiques à l'analyse syntaxique. PhD thesis, Université de Marne-la-Vallée, 2003.

PAUMIER, S. 2014. *Unitex 3.0 - User's Manual*. Paris: Université Paris-Est Marne-la-Vallée.

ROCHA, P. AND SANTOS, D. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Nunes, M.G. *et al.*, eds., *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, São Paulo: ICMC/USP. pp. 131–140.

SAG, I.A., BALDWIN, T., BOND, F., COPESTAKE, A.; FLICKINGER, D. 2002. Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.) *Proceedings of the Third International Conference, CICLing - Computational Linguistics and Intelligent Text Processing*, Mexico City, Mexico, February, 2002, pp. 1-15.

POS-PATTERNS OR SYNTAX? COMPARING METHODS FOR EXTRACTING WORD COMBINATIONS

Sara Castagnoli
University of Bologna
s.castagnoli@unibo.it

Gianluca E. Lebani
University of Pisa
gianluca.lebani@for.unipi.it

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Francesca Masini
University of Bologna
francesca.masini@unibo.it

Malvina Nissim
University of Groningen
m.nissim@rug.nl

Lucia Passaro
University of Pisa
lucia.passaro@for.unipi.it

Abstract

This paper reports on work carried out in the framework of an ongoing project aimed at building an online, corpus-based lexicographic resource for Italian Word Combinations. Our aim is to compare two of the most commonly used methods for the automatic extraction of word combinations from corpora, with a view to evaluate their performance – and ultimately their efficacy – with respect to the task of acquiring word combinations for inclusion in the lexicographic combinatory resource.

1. WORD COMBINATIONS: LEXICOGRAPHY AND NLP

It is widely acknowledged that lexicographers' introspection alone cannot provide comprehensive information about word meaning and usage, and that investigation of language in use is fundamental for any reliable lexicographic work (Atkins and Rundell 2008). This is even more true for dictionaries that record the combinatorial behaviour of words, where the lexicographic task is to detect the typical combinations a word participates in. In fact, it was much harder to study lexical combinatorics empirically before the advent of large corpora and the definition of statistical techniques for the analysis of word associations (Hanks 2012).

This paper reports on work carried out in the framework of an ongoing project called CombiNet[31] aimed at building an online, corpus-based lexicographic resource for Italian Word Combinations. We use the term **Word Combinations** (WoCs) to encompass both Multiword Expressions (MWEs) – namely WoCs characterised by different degrees of fixedness and idiomaticity that act as a single unit at some level of linguistic analysis, such as idioms, phrasal lexemes, collocations, preferred combinations (Calzolari et al. 2002, Sag et al.

[31] **PRIN Project 2010-2011** *Word Combinations in Italian* (n. 20105B3HE8) funded by the Italian Ministry of Education, University and Research (MIUR). URL: http://combinet.humnet.unipi.it.

2002, Gries 2008) – and the distributional properties of a word at a more abstract level (argument structure, subcategorization frames, selectional preferences), along the lines of Benson et al. (2010).

The specific aim of this paper is to compare two of the most commonly used methods for the automatic extraction of WoCs from corpora (cf. 1.1), with a view to evaluate their performance – and ultimately their efficacy – with respect to the task of acquiring WoCs for inclusion in our lexicographic combinatory resource. More specifically, we calculate the recall of the two methods using as benchmark the list of combinations recorded in *Dizionario Combinatorio Italiano* (DiCI, Lo Cascio 2013), the largest existing Italian combinatory dictionary. In addition, manual inspection of the top candidates in both datasets is used to assess the proportion of *valid* WoCs that are extracted from the corpus but unattested in DiCI.

1.1. Comparing methods for WoC extraction

Currently, apart from purely statistical approaches, the most common methods for the extraction of WoCs involve searching a corpus via sets of patterns and then ranking the extracted candidates according to various association measures (hybrid method) in order to distinguish meaningful combinations from sequences of words that do not form any kind of relevant unit (Villavicencio et al. 2007, Ramisch et al. 2010).

Generally, the search is performed for either shallow morphosyntactic (POS) patterns (**P-based approach**) or syntactic dependency relations (**S-based approach**). In the case of P-based methods, one needs to have a POS-tagged corpus and to draw a list of POS-patterns assumed to be representative of WoCs in a given language (see e.g. (1)). In the case of S-based methods, one needs to have a parsed corpus and to identify syntactic relations that may give rise to meaningful WoCs (see e.g. (2)).

(1) a. NOUN PREP NOUN

punto di vista

'point of view'

b. NOUN ADJ

anno accademico

'academic year'

(2) a. SUBJ – VERB

guerra – scoppiare

'war – burst'

b. VERB – OBJ

perdere – vista

'lose – (one's)sight'

Most studies so far have concentrated on P-based approaches, which yield satisfactory results for relatively fixed, short and adjacent WoCs. More recently it has been suggested that syntactic dependencies might be helpful to also capture discontinuous and syntactically flexible WoCs, because they can extract syntactically related words irrespective of their surface realizations (Seretan 2011).

Clearly, both methods have cons. In the case of P-based methods, POS-patterns need to be specified a priori. Moreover, not every extracted combination is a WoC, even using a detailed list of patterns and even after applying association measures (cf., among others, Nissim et al. 2014). Finally, without considering syntactic information, it is difficult to extract complex and flexible WoCs (especially verbal ones), let alone more schematic combinatory information (e.g. argument structure). As for S-based methods, abstracting away from specific constructs and information (e.g. linear order, morphosyntactic features, etc.) may result in little information about how exactly words combine. Moreover, it is hard to distinguish frequent, regular combinations from highly fixed, idiomatic ones with the very same syntactic structure.

Overall, the two methods seem to be highly complementary rather than competing with one another. In fact, various attempts are currently being proposed to put them together (cf. the SYMPATHy method discussed in Lenci et al. 2014, 2015; cf. also Heid 2015 and Squillante 2015), and the results of our experiment also point in this direction.

2. THE EXPERIMENT

In order to test and compare the performance of the two above-mentioned methods with respect to the task of extracting WoCs for lexicographic purposes, we selected a sample of 25 Italian target lemmas (TLs) – including 10 nouns, 10 verbs and 5 adjectives (listed in Table 1) – and we extracted P-based and S-based combinatory information from *la Repubblica* corpus[32] (Baroni et al. 2004). TLs were selected by combining frequency information derived from the *la Repubblica* corpus and inclusion in Lo Cascio's (2013) DiCI, which is used for (part of the) evaluation.

Nouns	**Verbs**	**Adjectives**
anno 'year'	*parlare* 'talk / speak'	*economico* 'economic'
governo 'government'	*prendere* 'take'	*giovane* 'young'
casa 'house'	*tenere* 'keep / hold'	*basso* 'low / short'
fine 'end / goal'	*vivere* 'live'	*facile* 'easy'
guerra 'war'	*perdere* 'lose/miss'	*rosso* 'red'
famiglia 'family'	*uscire* 'go out'	
mano 'hand'	*lavorare* 'work'	
situazione 'situation'	*costruire* 'build'	
morte 'death'	*pagare* 'pay'	
stagione 'season'	*leggere* 'read'	

Table 1. Target lemmas for the experiment

[32] The *la Repubblica* corpus (approx. 380M tokens) contains texts from the homonymous Italian daily newspaper. The version of the corpus we used was POS-tagged with the tool described in Dell'Orletta (2009) and dependency-parsed with DeSR (Attardi and Dell'Orletta, 2009).

As regards the P-based method, we extracted all the occurrences of each TL in a set of 122 pre-defined POS-patterns deemed representative of Italian WoCs. The set includes:

- POS sequences mentioned in existing combinatory dictionaries (previously collected in Piunno et al. 2013) and relevant theoretical literature (e.g. Voghera 2004; Masini 2012);

- "new" patterns identified through corpus-based, statistical experiments (Nissim et al. 2014);

- more patterns added manually by elaborating on the previous lists.

For the actual extraction, we used the EXTra tool (Passaro & Lenci 2015). EXTra retrieves all occurrences of the specified patterns, only as linear and contiguous sequences (no optional slots can be included), and ranks them according to a variety of association measures, among which we chose Log Likelihood (LL). The search considers lemmas, not wordforms. Finally, only sequences with frequency >5 have been considered. See Table 2 for an example of data extracted with EXTra.

LL	FREQ	W1	POS	W2	POS	W3	POS
5176.86	702	medico	s	**di**	e	famiglia	s
5176.86	100	medico	s	**in**	e	famiglia	s
5176.86	9	medico	s	**di**	ea	famiglia	s
3205.18	6	amico	s	e	cc	famiglia	s
3205.18	82	amico	s	di	ea	famiglia	s
3205.18	545	amico	s	di	e	famiglia	s
2983.87	403	famiglia	s	cristiano	a		
2615.41	80	cassaforte	s	di	ea	famiglia	s
2615.41	152	cassaforte	s	di	e	famiglia	s
2537.23	600	intero	a	famiglia	s		
2315.64	1154	grande	a	famiglia	s		
2114.56	234	gioiello	s	di	e	famiglia	s

Table 2. Examples of candidates extracted by EXTra for the TL *famiglia* 'family'

Note that the same combination of lemmas can be listed more than once in the results: take for instance the lemma sequence *medico+di+famiglia* ('doctor'+'of'+'family'), which appears in row 1 and row 3 in Table 2. The two hits represent two separate candidates because of the different morphosyntactic configurations of the combination, i.e. because of the different preposition intervening between 'doctor' and 'family': a simple preposition in the first case (cf. *medico di famiglia* 'general practitioner, GP', which is indeed a MWE), an articulated preposition in the second case (cf. *medico della famiglia* 'doctor of the family', which is a normal phrase). Although the two candidates have the same LL value (5176,86), because the preposition is ignored when computing the association strength, their respective frequency (702 vs. 9) appears indicative. Moreover, since data extraction is based on shallow sequences, word order is strictly preserved in the output: hence, the combination *intero+famiglia* ('entire'+'family') represents only the occurrences of the two lemmas in this order (*intera famiglia* 'entire family', A+N), despite the reversed one (*famiglia intera*, N+A) would also be possible.

As regards extraction based on syntactic dependencies (S-based), we extracted the distributional profile of each TL using the LexIt tool (Lenci et al. 2012), which works with Italian nouns, verbs and adjectives. The LexIt distributional profiles contain the syntactic slots (subject, complements, modifiers, etc.) and the combinations of slots (frames) with which words co-occur, abstracted away from their surface morphosyntactic patterns and actual word order. For instance, *Gianni ha dato volentieri un libro a Maria* 'John has willingly given a book to Mary' and *Gianni ha dato a Maria un libro* 'John has given Mary a book' are both mapped onto the syntactic frame "subj#obj#comp_a", despite the different order of their slots and the presence of adverbial modifiers. Moreover, each slot is associated with lexical sets formed by its most prototypical fillers. The statistical salience of each element in the distributional profile is estimated with LL.

For each TL we extracted all its occurrences in different syntactic frames together with the lexical fillers (lemmas) of the relevant syntactic slots, abstracting away from their surface morphosyntactic patterns. As for the P-based settings, only combinations with frequency >5 have been considered.

LL	FREQ	W1 (POS)	SYNT_REL	W2 (POS)
8939.28	1258	famiglia (s)	**modadj-post**	reale (a)
7084.59	1577	famiglia (s)	**modadj-pre**	grande (a)
6364.01	1657	famiglia (s)	modadj-post	italiano (a)
4543.05	719	famiglia (s)	modadj-pre	intero (a)
4271.25	548	famiglia (s)	modadj-post	cristiano (a)
3740.05	514	famiglia (s)	modadj-post	mafioso (a)
3708.22	465	famiglia (s)	**comp_di**	vittima (s)
LL	FREQ	W1 (POS)	SYNT_REL	W2 (POS)
15128.3	1180	perdere (v)	comp_di	vista (s)
15118.06	2615	perdere (v)	**obj**	occasione (s)
12066.27	3539	perdere (v)	**obj**	tempo (s)
11360.72	1831	perdere (v)	obj	terreno (s)
6504.6	1475	perdere (v)	obj	testa (s)

Table 3. Examples of candidates extracted by LexIt for the TLs *famiglia* 'family' and *perdere* 'to lose'

Table 3 shows an example of data extracted with LexIt. Although word order is generally underspecified, in some cases it is indicated in the syntactic relation itself: for instance, the "modadj-post" relation indicates that the first candidate – composed of *famiglia*+*reale* 'family'+'royal' – is *famiglia reale* 'royal family' in the N+A order, whereas "modadj-pre" indicates that the second candidate – composed of *famiglia*+*grande* 'family'+'big' – is *grande famiglia* ('big family') as A+N. Also, note that, in LexIt frames intervening tokens between slots (e.g. determiners, adverbial modifiers, etc.) are not recorded. Hence, the difference between *perdere*+*occasione* 'miss'+'chance' (which normally requires a determiner: *perdere un'occasione* 'miss a chance') and *perdere*+*tempo* 'lose'+'time' (where *tempo* is typically a bare noun: *perdere tempo* 'waste time') is not captured.

3. EVALUATION

The performance of the two extraction methods was assessed by means of a twofold evaluation. First, we calculated precision and recall using as benchmark dataset the list of combinations for our TLs recorded in DiCI. This was expected to shed light on the independent performance of the two methods overall, and with respect to the extraction of different types of WoCs. In addition, human evaluation of the top P-based and S-based candidates was carried out to assess the proportion of valid WoCs that are extracted from the corpus but unattested in a manually compiled resource like DiCI, thus providing information towards improving dictionary coverage.

3.1. Evaluation against DiCi

As DiCI is a traditional paper dictionary, we first built our gold standard benchmark dataset by digitizing the relevant entries and stripping off irrelevant information to obtain bare WoC lists. In order to enable automatic comparison with candidates from the two extraction systems, we then obtained a lemmatized version of benchmark combinations by performing POS and lemma annotation with the same tools used for corpus processing. Then we calculated global recall, overall precision and R-precision, as discussed in the following sections.

3.1.1. Recall

Recall is calculated as the percentage of extracted candidates out of the combinations found in the gold standard. For example, for the TL *rosso* 'red', EXTra extracts 23 of the 32 entries included in DiCi, thus its recall is 71.9% (23/32).

Recall points to an overall complementarity of the two systems, which are biased towards targets with different POS. As shown by the dark grey cells in Table 4, apart from cases in which the two systems have a very close performance (light grey cells), EXTra (P-based) performs better than Lexit (S-based) for nominal and adjectival TLs, whereas LexIt has a higher recall for virtually all verbal TLs.

Lemma		DiCi	EXTra_cand	Over	Rec	LexIt_cand	Over	Rec
rosso	A	32	805	23	0,719	476	22	0,688
situazione	N	149	2518	96	0,644	1343	79	0,53
stagione	N	64	1020	41	0,641	644	32	0,5
governo	N	144	5826	90	0,625	1327	50	0,347
anno	N	113	8762	63	0,558	3223	38	0,336
famiglia	N	130	2340	69	0,531	716	28	0,215
morte	N	83	1403	43	0,518	510	15	0,181
casa	N	356	3734	170	0,478	1092	95	0,267
mano	N	252	2555	117	0,464	934	34	0,135
facile	A	36	876	16	0,444	549	10	0,278
fine	N	71	2801	26	0,366	1017	11	0,155
economico	A	84	2384	62	0,738	981	62	0,738
guerra	N	62	2480	45	0,726	899	44	0,71
perdere	V	145	1557	96	0,662	2437	92	0,634
basso	A	72	668	46	0,639	457	44	0,611
giovane	A	50	1566	20	0,4	926	23	0,46
uscire	V	116	2010	66	0,569	2749	72	0,621
pagare	V	120	1474	61	0,508	1786	76	0,633
prendere	V	237	2831	109	0,46	4813	140	0,591
lavorare	V	98	1553	45	0,459	2218	53	0,541
vivere	V	197	1717	86	0,437	2517	106	0,538
leggere	V	117	1091	50	0,427	1514	68	0,581
costruire	V	90	1095	36	0,4	1249	53	0,589
parlare	V	194	3813	73	0,376	5896	87	0,448
tenere	V	159	1859	58	0,365	3569	97	0,61

Table 4. Comparing Extra and LexIt: Recall

3.1.2. Precision

Overall precision is not very significant as the number of extracted candidates for the two systems varies a lot, and it is generally very high. A better indicator of precision is *R-precision*, a measure borrowed from information retrieval and useful when assessing the quality of ranks. R-precision measures precision at the rank position corresponding to the number of combinations found in the gold standard, in our case DiCI. The rationale behind this is that an optimal system would place in the top *n* hits exactly all *n* entries found in the gold standard. Because our entries are ranked via association measures, and because both systems extract a large number of candidates, R-precision is a useful indicator of how well both methods perform and compare. To give an example, for the TL *pagare* 'to pay' there are 120 WoCs in the benchmark dictionary. In the top 120 candidates for Extra and LexIt we find 22 and 38 of these WoCs, respectively. So R-precision is higher for LexIt. Indeed, R-precision is almost always higher for LexIt (S-based) than for Extra (P-based), irrespective of POS, since Lexit performs better for all verbs and adjectives, as well as for most nouns (see dark grey cells in Table 5).

Lemma		total_DiCi	Extra	R-Prec	%	Over	Prec	Lexit	R-Prec	%	Over	Prec
pagare	V	120	1474	22	0,183	61	**0,041**	1786	38	**0,317**	76	0,034
tenere	V	159	1859	30	0,189	58	**0,031**	3569	41	**0,258**	97	0,016
perdere	V	145	1557	32	0,221	96	**0,062**	2437	36	**0,248**	92	0,039
costruire	V	90	1095	11	0,122	36	**0,033**	1249	21	**0,233**	53	0,029
vivere	V	197	1717	25	0,127	86	**0,05**	2517	43	**0,218**	106	0,034
prendere	V	237	2831	40	0,169	109	**0,039**	4813	50	**0,211**	140	0,023
uscire	V	116	2010	21	0,181	66	**0,033**	2749	22	**0,19**	72	0,024
leggere	V	117	1091	14	0,12	50	**0,046**	1514	21	**0,179**	68	0,033
lavorare	V	98	1553	16	0,163	45	**0,029**	2218	17	**0,173**	53	0,02
parlare	V	194	3813	20	0,103	73	**0,019**	5896	29	**0,149**	87	0,012
economico	A	84	2384	7	0,083	62	0,026	981	28	**0,333**	62	**0,063**
basso	A	72	668	13	0,181	46	0,069	457	18	**0,25**	44	**0,101**
rosso	A	32	805	2	0,062	23	0,029	476	6	**0,188**	22	**0,048**
giovane	A	50	1566	4	0,08	20	0,013	926	6	**0,12**	23	**0,022**
facile	A	36	876	2	0,056	16	0,018	549	4	**0,111**	10	**0,029**
situazione	N	149	2518	21	0,141	96	0,038	1343	40	**0,268**	79	**0,071**
guerra	N	62	2480	1	0,016	45	0,018	899	16	**0,258**	44	**0,05**
stagione	N	64	1020	8	0,125	41	0,04	644	16	**0,25**	32	**0,064**
casa	N	356	3734	43	0,121	170	0,046	1092	73	**0,205**	95	**0,156**
governo	N	144	5826	11	0,076	90	0,015	1327	22	**0,153**	50	**0,068**
famiglia	N	130	2340	17	0,131	69	0,029	716	19	**0,146**	28	**0,096**
anno	N	113	8762	2	0,018	63	0,007	3223	16	**0,142**	38	**0,02**
morte	N	83	1403	13	**0,157**	43	0,031	510	10	0,12	15	**0,084**
mano	N	252	2555	57	**0,226**	117	0,046	934	25	0,099	34	**0,125**
fine	N	71	2801	4	**0,056**	26	0,009	1017	3	0,042	11	**0,026**

Table 5. Comparing Extra and LexIt: R-precision

3.1.3. Thresholds

Obviously, precision and recall vary as we examine more candidates. This sort of information is useful when automatically extracted data then need to be analyzed manually by lexicographers. We therefore calculated both precision and recall at different thresholds (viz. every 250 hits). Figure 1 shows how they vary with increasing batch sizes; figures are averaged across different TLs with the same POS, so that we have one curve for nouns, one for adjectives, and one for verbs.

As expected, recall increases and precision decreases for both EXTra and LexIt . However, some interesting remarks can also be made. For example, apart from a few isolated cases which are represented by only few data points, recall for EXTra (top left) for nominal and verbal TLs seems to plateau after about 2,000 hits: this might suggest that a lexicographer could obtain a good coverage by concentrating on the manual evaluation of about 2,000 candidates per such TLs.

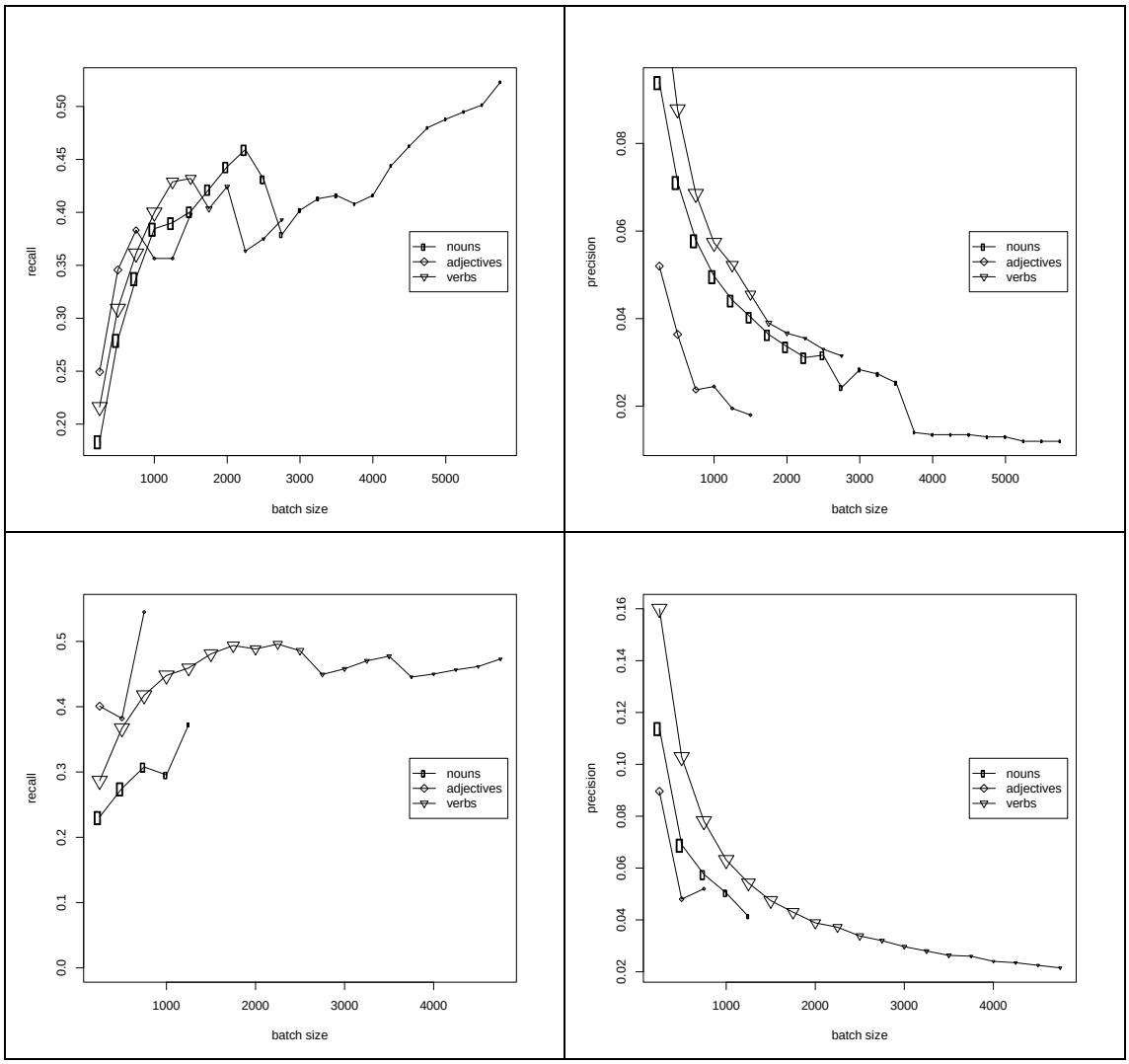


Figure 1. Precision and recall for EXTra (top) and LexIt (bottom) plotted against batch sizes. The size of the data points indicates the number of TLs included in the counts. The maximum size is 10 for nouns and verbs, and 5 for adjectives. The minimum number is 2. Batches with only one TL are not shown.

3.1.4. P-based/S-based overlap

Total overlap is calculated as the percentage of cases in which EXTra and LexIt retrieve/don't retrieve the same gold standard combinations. For instance, the benchmark entry for the adjectival lemma *giovane* 'young' contains 50 combinations: out of these, 20 are retrieved by both EXTra and LexIt, 27 are retrieved by neither of the two systems, and LexIt only extracts 3 further WoCs. This means that the performance of the two systems is identical for 94% of DiCI combinations for the TL. This is the case of the highest overlap between the P-based and the S-based system, which is however quite high (76.05% on average, spanning between 59.07% and 94%, see Table 6). Random manual observations were made to explore possible causes for cases of "negative overlap", that is gold standard combinations that neither of the systems extracts. On the one hand, these appear to include e.g. WoCs with corpus frequency ≤5, as well as proverbs/idioms, thus pointing to a possible impact of corpus type and size. On the other hand, the actual WoC-ness/representativeness of some combinations in the gold standard is somewhat debatable.

When there is no overlap, i.e. when the two systems extract different gold standard combinations, the data (see Table 6) confirm that:

- the independent contribution of EXTra is higher (grey cells) for nouns and most adjectives, and in many such cases LexIt's contribution is minimal (often less than 1/5 of the number of "new" WoC retrieved by EXTra).

- the independent contribution of LexIt is higher for virtually all verbs, and in most such cases the contribution of EXTra is less than half the independent contribution of LexIt.

Lemma		total_DiCI	both	EXTra_only	LexIt_only	none	%overlap
casa	N	356	81	**89**	14	172	71,07
mano	N	252	29	**88**	5	130	63,10
governo	N	144	46	**44**	4	50	66,67
famiglia	N	130	26	**43**	2	59	65,38
anno	N	113	34	**29**	4	46	70,80
morte	N	83	14	**29**	1	39	63,86
situazione	N	149	68	**28**	11	42	73,83
perdere	V	145	72	**24**	20	29	69,66
fine	N	71	11	**15**	0	45	78,87
stagione	N	64	30	**11**	2	21	79,69
facile	A	36	9	**7**	1	19	77,78
basso	A	72	40	**6**	4	22	86,11
guerra	N	62	40	**5**	4	13	85,48
rosso	A	32	20	**3**	2	7	84,38
economico	A	84	57	5	5	17	88,10
prendere	V	237	76	33	**64**	64	59,07
vivere	V	197	70	16	**36**	75	73,60
parlare	V	194	60	13	**27**	94	79,38
leggere	V	117	43	7	**25**	42	72,65
pagare	V	120	56	5	**20**	39	79,17
lavorare	V	98	40	5	**13**	40	81,63
uscire	V	116	61	5	**11**	39	86,21
costruire	V	90	34	2	**19**	35	76,67
tenere	V	159	57	1	**40**	61	74,21
giovane	A	50	20	0	**3**	27	94,00

Table 6. Comparing EXTra and LexIt: overlap and differences in WoC extraction

A quick look at the different combinations extracted by the two systems suggests that the results might be influenced by the specific features and settings of the tools. For instance, verbs ending in *-si* (e.g. *prender-**si** un raffreddore* 'catch a cold') are not captured by EXTra because they are not lemmatized as such in the corpus, while LexIt extracts them thanks to dedicated frames (e.g. subj#si#obj). EXTra also does not capture complex complements as in *prendere con le mani nel sacco* 'catch (s.one) red-handed', as long, complex

patterns like V+PREP+DET+N+PREP+N were not included in the POS pattern set used for candidate extraction, due to their great variability and questionable productivity. Moreover, the possibility to include optional slots in LexIt, contrary to EXTra's fixed POS patterns, might favor the better performance by the former with verbal TLs. The picture is different for nominal/adjectival TLs, where variation and flexibility are less marked than with verbs.

Further investigation is needed to assess the exact impact of these features and settings on the results. Some problems may be solved by varing the extraction parameters, while others directly relate to intrinsic limits to either P-based or S-based approaches.

3.2. Human evaluation

Manual inspection of the top candidates in both datasets was used to assess the proportion of *valid* WoCs that were extracted from the corpus but unattested in DiCI. We obtained human judgments over 2,000 candidates for 10 TLs (1,000 from EXTra and 1,000 from LexIt, taking the top 100 results for each TL from each system).

Nouns	**Verbs**	**Adjectives**
guerra 'war'	*prendere* 'take'	*basso* 'low / short'
famiglia 'family'	*tenere* 'keep / hold'	*rosso* 'red'
mano 'hand'	*uscire* 'go out'	
stagione 'season'	*pagare* 'pay'	

Table 7. Target lemmas used for human evaluation

Annotators were linguists, not necessarily working on WoCs, mainly with a background in translation and/or corpus work. We collected two judgments per candidate. Possible annotations included: **Y** (yes, this is a valid WoC), **N** (no, this is not a valid WoC) and **U** (uncertain, not sure/this may be *part* of a valid WoC). We considered as *valid* candidates only those which received either YY or YU. Table 8 summarizes the results:

	Valid candidates extracted from corpus	**Valid candidates not recorded in DiCI**
EXTra	408 (/1000)	273 (/408)
LexIt	447 (/1000)	261 (/447)
EXTra+LexIt	**855** (/2000)	**534** (/855)

Table 8. Results of human evaluation

Out of 2,000 total candidates, we obtained positive evaluations for 855 combinations (408 from EXTra, 447 from LexIt). Out of these 855 WoCs deemed valid by the annotators, 534 are not recorded in DiCI: 273 from EXTra, 261 from LexIt. If we intersect the two sets, we find that only 80 WoCs are in common, which means we have **454** actual *new* WoCs, which are retrieved thanks to the two corpus-based methodologies. This again confirms their complementary contribution to WoC mining.

4. DISCUSSION AND CONCLUSION

The goal of this paper was to compare two commonly used methods for the automatic extraction of WoCs from corpora – the P-based method and the S-based method – with a view to evaluate their performance and efficacy. To this aim, we set up a twofold evaluation of candidates extracted by two systems – EXTra and LexIt – implementing the two approaches.

As for automatic evaluation (cf. 3.1), recall against DiCi is good for both EXTra (P-based) and LexIt (S-based). In addition, the data suggest a **complementarity** of the two systems, as recall appears to be related to the POS of the TL: EXTra performs better than LexIt for nominal and adjectival TLs, whereas LexIt has a higher recall for virtually all verbal TLs. However, further investigations might be needed to ascertain the extent to which the results are influenced by corpus type, by the specific features and settings of the extraction tools, as well as by the quality of the gold standard.

As for human evaluation (cf. 3.2), our experiment shows that over 40% of WoCs extracted by EXTra and LexIt are deemed valid by human annotators, and that more than half of these valid candidates are not attested in DiCI. This result is even more remarkable if we consider that we only evaluated the top 100 candidates for each TL/system. Automatic extraction of data from corpora therefore proves to be potentially very fruitful for lexicography, since it adds a high number of WoCs that are not recorded in traditional dictionaries, even comprehensive ones such as DiCI. Human evaluation also confirms the complementarity of the two systems, since out of the total number of *valid* WoCs extracted by the two systems and not recorded in DiCI (534), only 80 combinations overlap.

These findings make us all the more convinced of the need for hybrid systems that simultaneously take into account information targeted in P-based and S-based approaches.

References

ATKINS, B.T.S. AND RUNDELL, M., 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

BARONI, M., BERNARDINI, S., COMASTRI, F., PICCIONI, L., VOLPI, A., ASTON, G. AND MAZZOLENI, M., 2004. Introducing the "La Repubblica" Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. *Proceedings of LREC 2004*, pp.1771–1774.

BENSON, M., BENSON, E. AND ILSON, R., 2010. *The BBI Combinatory Dictionary of English*. 3rd revised edition. Amsterdam/Philadelphia: John Benjamins.

CALZOLARI, N., FILLMORE, C.J., GRISHMAN, R., IDE, N., LENCI, A., MACLEOD, C. AND ZAMPOLLI, A., 2002. Towards best practice for multiword expressions in computational lexicons. *Proceedings of LREC 2002*, pp.1934–1940.

GRIES, S. TH., 2008. Phraseology and linguistic theory: a brief survey. In: S. Granger and F. Meunier, eds. *Phraseology: an interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins. pp.3–25.

HANKS, P., 2012. Corpus evidence and Electronic Lexicography. In: S. Granger and M. Paquot eds. *Electronic Lexicography*. Oxford: Oxford University Press. pp.57–82.

HEID, U. 2015. *Extracting linguistic knowledge about collocation from corpora*. Plenary talk delivered at the EUROPHRAS 2015 Conference. Malaga, June 29 – July 1, 2015.

LENCI, A., LAPESA, G. AND BONANSINGA, G., 2012. LexIt: A Computational Resource on Italian Argument Structure. *Proceedings of LREC 2012*, pp.3712–3718.

LENCI, A., LEBANI, G.E., CASTAGNOLI, S., MASINI, F. AND NISSIM, M., 2014. SYMPAThy: Towards a comprehensive approach to the extraction of Italian Word Combinations. *Proceedings of CLiC-it 2014* (Pisa, 9-11 December 2014), pp.234-238.

LENCI, A., LEBANI, G.E., SENALDI, M.S.G., CASTAGNOLI, S., MASINI, F. AND NISSIM, M., 2015. Mapping the Constructicon with SYMPAThy: Italian Word Combinations between fixedness and productivity. *Proceedings of the NetWordS Final Conference* (Pisa, March 30-April 1, 2015), pp.144-149.

LO CASCIO, V., 2013. *Dizionario combinatorio italiano*. Amsterdam/Philadelphia: John Benjamins.

MASINI, F., 2012. *Parole sintagmatiche in italiano*. Roma: Caissa.

NISSIM, M., CASTAGNOLI, S., MASINI, F. 2014. Extracting MWEs from Italian corpora: A case study for refining the POS-pattern methodology. *Proceedings of the 10th Workshop on Multiword Expressions* (MWE2014), pp.57–61.

PASSARO, L. C. AND LENCI, A. 2015. *Extracting Terms with EXTra*. Paper presented at the EUROPHRAS 2015 Conference. Malaga, June 29 – July 1, 2015.

PIUNNO, V., MASINI, F. AND CASTAGNOLI, S., 2013. *Studio comparativo dei dizionari combinatori dell'italiano e di altre lingue europee*. CombiNet Technical Report. Roma Tre University and University of Bologna.

RAMISCH, C. VILLAVICENCIO, A. AND BOITET, C., 2010. mwetoolkit: a framework for multiword expression identification. *Proceedings of LREC 2010*, pp.662–669.

SAG, I.A., BALDWIN, T., BOND, F., COPESTAKE, A. AND FLICKINGER, D., 2002. Multiword expressions: A pain in the neck for NLP. *Proceedings of CICLing 2002*, pp.1–15.

SQUILLANTE, L., 2015. *Polirematiche e collocazioni dell'italiano. Uno studio linguistico e computazionale*. Ph.D. dissertation Università di Roma "La Sapienza".

SERETAN, V., 2011. *Syntax-based collocation extraction*. Dordrecht: Springer.

VILLAVICENCIO, A., KORDONI, V., ZHANG, Y., IDIART, M. AND RAMISCH, C., 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp.1034–1043.

VOGHERA, M., 2004. Polirematiche. *Linguistica Pragensia*, 67(2), pp.100–108.

SEMANTIC STRUCTURING OF VERBAL IDIOMS FROM THE CONCEPTUAL DOMAIN {DEATH} A RUSSIAN-PORTUGUESE-ENGLISH CONTRASTIVE APPROACH

Svitlana Chornobay
Crimean Federal
V.I.Vernadsky University
svetoniy@gmail.com

Jorge Baptista
University of Algarve
L2F/INESC-ID Lisboa
jbaptis@ualg.pt

Keywords: Verbal Idioms, Semantic Structure of Conceptual Domain, Russian, Portuguese, English

Abstract

This paper deals with the comparative study of the semantics of Russian, Portuguese and English verbal idioms, constituting the conceptual domain {death}. Being one of the most basic concepts, spanning various linguistic cultures, the paper investigates its linguistic expression across distantly related languages, presenting the semantic structure of the idioms from this conceptual domain via a network of connotative relations. Cases of idiomatic asymmetry, based on the peculiarities of connotative processes involved in idioms decoding mechanisms, are also presented. The contrastive approach highlights the differences across languages and the semantic structuring can be used to capture better the complex discursive relations idioms hold with their context.

1. INTRODUCTION

The topicality of the research is determined by the growth of interdisciplinary linguistic studies as well as comparative ones. The importance of languages comparison was highlighted by W. von Humboldt (1985), as it gives the opportunity to demonstrate

> "which various ways a human created a language and which part of the world of thoughts he managed to transfer into it, how the individuality of the nation influenced the language and which reverse impact the language had on the people's individuality" (Humboldt, 1985, p. 311).

In this respect, it is of great importance to investigate the information got as a result of human's cognition of the life's essence and its ontologically important aspects, which are conceptualized and categorized in the mind and language according to geographical and cultural-historical peculiarities of the *ethnos*.

The concepts of {life} and {death} are the basic core concepts of the conceptual sphere of any linguistic culture. Being archetypical images, {life} and {death} are ontologically interlaced and can reflect the peculiarities of the national consciousness presented in the language idiomatic stock (Karakevych, 2009, p. 131). The contrastive scrutiny of these concepts in French and Russian has been done (Grabarova, 2005). It is remarkable that idioms verbalizing the concept {life} in Russian linguistic culture have negative connotation. It may be explained by the fact that, being born at a definite time, place, to a definite family, a person has no right to choose: s/he is doomed to fight against his vicious traits and has to put up with such a situation (Grabarova, 2005, p. 276). Linguo-cognitive characteristics of metaphoric representation of the concept {death} in English have been presented by L. Hnapovs'ka (2008). Following these previous works, this paper's research is focused on the contrastive analysis of the concept {death} in Russian, Portuguese and English idiomatics.

The goals of this paper are twofold: on the one hand, (i) we propose a strategy to the semantic structuring of verbal idioms from the conceptual domain associated with {death}, based on the connotative processes involved in idioms decoding mechanisms; on the other hand, (ii) we assess the usefulness of this strategy by comparing data from three distantly related languages, Russian, English and Portuguese.

The remainder of this paper is structured as follows. Section 2 briefly presents the basic definitions used in the paper, the data sources for the description and the method of semantic structuring of the conceptual domain. Section 3 breaks down the idioms of the {death} conceptual domain in its main subsets.

2. BASIC DEFINITIONS AND METHODS

A verbal idiom can be defined (M. Gross, 1996) as a syntactic-semantic lexical-grammatical unit, composed of a verb and at least one main constituent (usually a complement) that is distributionally frozen with the verb; the overall meaning of the idiom cannot be calculated by composition of the meaning of the individual elements of the expressions, as they are interpreted when used separately in other contexts. Hence, in the idiom RU[33]: *отбросить коньки* (*otbrosit' kon'ki*), literally 'to throw off the skates', 'die' (cf. EN: *pop one's clogs, kick the bucket*) the overall meaning is not derived from the meaning of its lexical constituents.

The methods here used are based on componential and contrastive analyses. Componential analysis is the methodology of describing structural organization of the meaning as a set of minimal semantic components, each fulfilling its function and being connected with other definite hierarchic relations (Selivanova, 2006, p. 230). This method gives the opportunity to reveal some culturally important features in a definite domain. Contrastive method is aimed at identifying common and specific features of the analysed languages at all their levels (Selivanova, 2006, p.165). These are valuable approaches to studying non-related languages and understanding specific semantic domains.

Though several collections of idioms are available for Portuguese, including dictionaries, they usually do not distinguish verbal idioms from many other formal types of idiomatic expressions, including compound words. A survey of many of these sources (Baptista et al.

[33] Examples are preceded by the abbreviation of the respective language: EN for English, PT, for Portuguese and RU for Russian. A transliteration and a free translation are provided, and eventually a literal, word-by-word translation is also provided to better illustrate the phenomena being discussed. In this paper, the Passport (2013), ICAO Romanization norm of the Russian script was adopted.

2004, 2005) resulted in the extensive encoding in tabular format (in view of its computational processing) of approximately 2,400 verbal idioms, organized in several formal classes, depending on the structure of the verbal idiom (one or more fixed complements, its relative position, transformations they accept, etc.).

Scrutiny of idioms in Russian and other European languages, namely English, German and French, their structure, semantics and classification problems by the Russian scholars is quite numerous, going back to the first half of the XX century. However, the issues of defining and classifying idioms are still open for discussion. A review of the fundamental theoretical works on *Idiomatics* in native and foreign Linguistics has been done (Chornobay, 2010a, 2010b) along with the survey of different approaches to defining the term *idiom* (Chornobay, 2011a) and the main trends in current phraseological studies (Chornobay, 2012).

The main idea supporting this paper proposes that it is possible to structure a semantic domain, in this case, the idioms associated with the concept of {death} and obtain coherent subsets by relying on the tropological processes involved in the idiomatic interpretation of these expressions. This approach has two advantages over a simple, flat, classification that would only rely on the generic concept, or even one that would distinguish features associated with the main concept, such as {cause}: *send smb. into the bottomless pit*; {agent-volition}: smb. *put the horse/dog/etc. to sleep*; or even aspectual features, {unaccomplished}: (RU) *глядеть в гроб*, (gliadet' v grob) literally, 'look into coffin'. The adoption of this strategy has to do with the dual readings of many idioms: while their meaning is non-compositional, their insertion in discourse still holds a link with the non-idiomatic, literal meaning of the components, and the tropological relations they hold with that literal meaning, must be accounted for in many discursive situations.

The material under analysis includes 76 Russian verbal idioms (Fedorov, 2001), 56 from Portuguese (Baptista *et al.* 2004, 2005), and 107 from English (Kunin, 1984; Sinclair *et al.*, 2001; Ayto, 2010).

3. CONCEPTUAL STRUCTURING OF THE {DEATH} DOMAIN

In the process of structuring the idioms associated with the conceptual domain of {death}, based on the tropological processes involved in the idiomatic interpretation of these expressions, the following subsets were found:

1. {DIE-GO}

This metaphor involves *verbs of movement*, among which one finds the basic equivalent to 'go' (PT: *ir*). An inchoative aspect is often associated to other verbs of this type, such as PT: *partir* (depart) or RU: *отправиться* (otpravit'sia), literally 'set out', *отойти* (otoiti) literally 'go to':

RU: *отправиться на тот свет* (otpravit'sia na tot svet) 'go to the other world'

EN: *go to the other world*

PT: *ir/partir para o outro mundo* 'go to the other world'

This metaphor is related to joining the ancestors:

RU: *отойти к праотцам* (otoiti k praottsam) 'go to the forefathers'

EN: *go to the forefathers; to join the majority*.

A {causative} value may be added to this basic metaphor. In Portuguese, this is carried out by verbs such as PT: *mandar* '*send*', with a human (agentive) or a non-constraint subject, properly a {cause}:

PT: *Isso/O Pedro mandou o João para o outro mundo*

'That/Pedro send João to the other world'

In Russian, the causative construction is expressed by the verb *отправить* (otpravit') 'send':

RU: *Когда через несколько месяцев маленькая царевна умерла, в народе пошли толки, что Борис отправил ее на тот свет* (Kogda cherez neskol'ko mesiatsev malen'kaia tsarevna umerla, v narode poshli tolki, chto Boris otpravil ee na tot svet)

'When some months later a young queen died, rumours spread that Boris had sent her to the other world'

However, some idioms resource to this construction with a reflexive suffix on the verb RU: *отправляться* (otpravliat'sia) *'go myself'*. The subject and the direct object are the same:

RU: *Целые три года оставался он в самом жалком положении: и если бы он не получил от природы железного телосложения, то верно бы отправился на тот свет* (Tselie tri goda ostavalsia on v samom zhalkom polozhenii: i esli bi on ne poluchil ot prirodi zheleznogo teloslozheniia, to verno bi otpravilsia na tot svet)

'For three years he stayed in a pitiful condition: if he had not been of a strong constitution, he would have gone to the other world'

The locative-destination complement can be varied:

PT: *o outro mundo*

RU: *мир иной* (mir inoi), *тот свет* (tot svet) '*the other world*'

EN: *the other world, a better world*

PT: *céu*

RU: *небеса* (nebesa) 'heaven'

EN: *heaven*

PT: *paraíso*

RU: *рай* (rai) 'paradise'

EN: *kingdom come, one's last home*

and, with the same meaning but a negative connotation,

PT: *o* inferno

RU: *преисподняя* (preisponiaia) 'hell'

EN: hell '*go to hell*' (not interj.)

One also finds metonymy-derived expressions of some of the above:

RU: *райские кущи* (raiskie kushchi) 'bushes of the paradise'
RU: *отойти к праотцам* (otoiti k praottsam) '*go to the forefathers*'
('to the forefathers' = 'to the place where the forefathers are'; very rare in PT)
EN: *go to the forefathers; go to meet one's Maker*
PT: *mandar alguém para os anjinhos* 'send smb to the little angels'
(= 'the place where the angels are)

A subtype of locatives involves the concept of *inhumation* with verbs such as PT: *baixar/descer* 'descend' and RU: *сойти* (soiti) 'idem', and a variety of synonyms for '*grave*' (the verb-locative noun combinations are quite idiosyncratic):

PT: *O corpo do Pedro desceu à sepultura*

'The body of Pedro descendend into the tomb/grave'.

RU: *Он раньше времени сошёл в могилу* (On ran'she vremeni soshel v mogilu)

'He went untimely to the grave'

The causative construction is also attested in:

PT: *Isso levou o Pedro à cova* 'That led Pedro to the grave'

Compare in Russian:

RU: *Загнать/вогнать в гроб* (zagnat'/vognat' v grob) 'drive into the coffin'

RU: *свести в могилу кого-л* (svesti v mogilu kogo-l) 'bring smb to his grave';

RU: *вогнать в землю* (vognat' v zemliu) 'bring smb to the ground'.

Another subtype involves several euphemistic or dysphoric designations for 'graveyard':

PT: *a quinta das tabuletas*, literally 'the farm of the signs (=tombstones)',

PT: *a cidade dos pés juntos* 'the city of the feet-together'.

As a locative-transference predicates, synonyms of 'go' are also used with {source} and {destiny} complements:

PT: *O Pedro foi desta para melhor*

lit. 'go from this one (life?) to a better one (life?), 'go to a better place'.

In this latter example, the locatives are represented by a demonstrative and an adjective, the former in the feminine-singular, hence the proposed reconstruction of *vida* (life). With the locative-source complement only, one also finds:

PT: *abandonar este vale de lágrimas* 'to abandon this valley of tears',

an idiom with biblical consonances. Finally, one finds yet:

PT: *sair com os pés para a frente* 'to leave with the feet first',

where a verb, usually selecting a locative-source complement, has an adverbial modifier of manner.

The idioms EN: *cross the Stygian ferry*, *take the ferry* convey the idea of travelling to the other world. Originally, these idioms refer to the Ancient Greek mythology, according to which, Charon ferried the souls of the dead over the Styx in the underworld, where the dead gathered. The idiom *to cross the Styx* also means 'die'. These examples also imply the need of overcoming challenges or difficulties.

2. {DIE-SLEEP}

The metaphor of {death} as {sleep} is also a recurrent trope in several languages. The metaphor implies the process of rest or relaxation. In Russian one finds:

RU: *Почивать в боге/ бозе (*pochivat' v boge/ boze*)* 'rest in god';

*Почить вечным сном (*pochit' vechnym snom*)* 'rest the eternal sleep'

while in Portuguese there is:

PT: *O Pedro dorme o sono eterno* 'Pedro sleeps the eternal sleep'.

In English these correspond to the idiom *to fall asleep*. The idiom *to put something to sleep* euphemistically expresses the idea of painlessly killing (an animal, especially an old, sick, or badly injured one) in order to stop one's suffering. The idiom can also be included into the group {DIE-KILL}, denoting the manner of killing. There is a second figurative meaning of the idiom which has nothing to do with dying, but denotes the action of putting a machine (e.g., a computer) on standby, while it is not being used.

3. {DIE-Stretch Legs}

In this case, Russian has two verbs *протянуть* (protianut') 'stretch-unwind' and *вытянуть* (vitianut') 'stretch-extend', both with the same frozen direct object. However, in Portuguese, only the first connotation is involved, in PT *esticar o pernil*. Besides, in PT, the noun *pernil* literally designates 'ham', so there is a pejorative connotation associated to this word in the idiom, which is absent in the Russian counterparts. Russian idioms also imply the notion of 'getting the horizontal position in order to relax', which is absent in the Portuguese idiom, since it implies 'walking'. Apparently, there are no exact equivalents in English, though idioms with a similar meaning, also related to 'feet' have been detected:

EN: *to pop one's clogs* (where clogs = shoes, taken off the dead)

cf. RU: *коньки отбросить* (kon'ki otbrosit'), literally, 'throw off the skates'.

Another connotative meaning of these idioms is 'to get rid of something unnecessary'. The latter English expression, which was first recorded in 1970 (Ayto, 2010), probably comes from the idea of 'popping' (i.e. *pawning*) a person's clogs after s/he had died (and therefore have no further use for them).

4. {DIE-Surrender/Give-Spirit}

A family of expressions involve the concept of spirit/soul and surrendering it to the deity:

RU: *Отдать дух* (otdat' dukh); *Отдать душу* (otdat' dushu); *Отдавать богу душу* (otdavat' bogu dushu) *'give the spirit/soul to god'*

PT: *O Pedro deu a alma ao criador* 'Pedro surrender the (=his) soul to the maker'

EN: *give up the ghost/ soul, yield up the ghost*.

The Old English meaning of *ghost*, 'the soul or spirit as the source of life', survives only in these idioms.

The metaphor involved in these idioms identifies life with a token (the soul, the spirit), which is (implicitly) borrowed temporarily (by God) and must be given back. Most of these connotative constructs are present in all the languages under scrutiny. The recipient (God, the Creator) can often be omitted but remains implied.

5. {DIE-Breathe Away}

This metaphor takes *breathing* as the manifestation of life so that dying corresponds to stop breathing; the metaphor introduces a terminative aspect (stop living):

PT: *exalar/soltar/dar o último suspiro* 'exhale/release/give the last breath';

RU: *испустить последнее дыхание* (ispustit' poslednee dikhanie);

EN: *breathe one's last*

In PT, the construct associated to 'breath' is represented by the noun *suspiro* (sigh/moan), but the negative connotation is absent from the idiom, being used as a euphemistic periphrasis, while in Russian and English these idioms just imply the relief.

6. {DIE - SACRIFICE}

The Russian idioms:

RU: *отдавать душу* (otdavat' dushu) 'to give the soul',
RU: *отдавать жизнь за кого/что* (otdavat' zhizn' za kogo/chto) 'to give life for smb/ smth',
RU: *положить живот/ жизнь* (polozhit' zhivot'/ zhizn'),
RU: класть душу и живот (klast' dushu i zhivot) 'give (up) one's life/stomach for smth'

express the idea of 'dying for the sake of others', the readiness to sacrifice life for people, native land or ideology. Some Russian idioms highlight the importance and honour of sacrificing life in a battle, defending the Motherland:

RU: *пасть в боге* (past' v boge) 'fall in god',
RU: *пасть смертью храбрых* (past' smert'iu khrabrikh) 'perish the courageous death',
RU: *пасть как герой* (past' kak geroi) 'perish as a hero'.

The use of the verb 'пасть' (past') 'perish' and the noun 'живот (zhivot)' 'stomach' in the meaning of 'life' in these idioms refers to the bookish style. Having a separate verb for denoting heroic death in a battle, probably, shows that defending the Motherland has a slightly different meaning for the Russians than for the Europeans, namely the French (Grabarova, 2005, p.283). No idioms with such meaning have been identified yet, neither in Portuguese nor in English.

7. {DIE – DROWN}

The idea of drowning is expressed euphemistically through the metaphoric idioms both in English and Russian conveying the meaning of 'going down to the bottom of the sea':

EN: *feed the fishes*; *be/ become food for fishes*;
RU: *пойти на корм рыбам* (poiti na korm ribam) 'become food for fishes';
RU: *рыб кормить* (rib kormit') 'feed the fishes'.

The idiom:

EN: *go to Davy Jones's locker*

has a national colouring, referring to the creature of English nautical mythology, Davy Jones, identified as 'the fiend that presides over all the evil spirits of the deep', but the origin of the name is uncertain (Ayto, 2010). *Davy Jone's locker* denotes the bottom of the sea, especially regarded as the grave of those drowned at sea (extension of early 18th-cent. nautical slang *Davy Jones*).

8. {DIE – KILL}

Quite numerous is the group of English idioms denoting the concept of 'killing' or 'execution'. In the idiom EN: *get yours* the deserved punishment is implied. The idiom EN: *cheat the gallows* 'die in a natural way' has a negative connotation, expressing disapproval,

since the convicted had not been punished, and avoided execution. The implicit seme is 'the thirst for justice'.

The idioms EN: *die with your boots on*; *cop a packet*; *go for your tea* convey the meaning of 'being killed while still active, especially in a battle'. The latter idiom arose among members of the IRA in the end of the 20th century.

9. {DIE – COMMIT A SUICIDE}

Only some few idioms of this group have been detected yet, though they are culturally marked:

EN: *do the Dutch* (AmE),

where *Dutch* is short for *the Dutch act*: apparently in the 19th century, when the expression originated, the Dutch had a reputation in America for attempting suicide. English idioms with the component 'Dutch' usually have negative connotation due to the navigation rivalry between the English and the Dutch for centuries (Chornobay, 2011b).

In Russian, one finds, conveying this meaning:

RU: *кончать себя* (konchit' sebia) 'end oneself';
RU: *кончить с собой* (konchit' s soboi) 'end with oneself'

In Portuguese, several idioms can also be found:

PT: *pôr termo a/acabar com a vida* 'put an end to/finish with (one's) life

10. {DIE – LOSE CONNECTION WITH THE SOURCE OF LIFE}

In English, a separate group of idioms, denoting the loss of connection with the object that gives life (*drop off the hooks; pop off the hooks; slip off the hooks*) or the connector breaking with the source of life (*to cut one's cable; slip one's cable*), exists. Still no Russian nor Portuguese idioms have been identified conveying the same metaphors.

11. {ISOLATED METAPHORS}

A certain number of idioms do not form paradigms or only present a short number of variants. These outliers require specific description that, for lack of space, we cannot enter into:

RU: *дать дуба* (dat' duba) 'give the oak';

RU: *сыграть в ящик* (sigrat' v iashchik) 'play the box'

Taking as a basis the synonymic relations of these idioms, showing the specifics of a funeral in a coffin, it is possible to suggest the usage of oak as a material for making coffins is the source of the idioms. The metonymic transference is observed here (object – material of the object produced; Grabarova, 2005, p. 284).

PT: *bater a bota/caçoleta* (lit: 'kick/beat the boot/frizzen'), 'kick the bucket'

PT: (*estar a*) *fazer tijolo* (lit: 'to be making brick')

Unlike the previous idioms, this last example expresses a durative aspect (the state of being dead) and not the process of dying. Aspectual values associated with idioms may also be a valid dimension for structuring conceptual domains.

4. CONCLUSIONS AND FUTURE WORK

Semantic classification of verbal idioms can be approached from different perspectives. The first step can be to try and build general, semantically course classes, such as the one here discussed, {death}. We posit that the metaphoric processes underlying the idiomaticity these expressions convey can be a useful strategy, especially in a contrastive setting, such as the Russian/Portuguese/English comparison here made. Detailed tropological analysis is required; this paper showed it could be done. Extension to other semantic constructs, such as {BIRTH}, {MARRIAGE}, {DIVORCE}, etc. should provide a better perspective of this approach, especially if the contrastive setting is extended to other, non-related languages, such as Ukrainian and Modern Greek, aspects that will be tackled in future work.

Acknowledgments

Research for this paper was partially funded by Portuguese national funds through Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2014.

References

AYTO, John, 2010. *Oxford Dictionary of English Idioms*. Oxford; New York: Oxford University Press.

BAPTISTA, Jorge; CORREIA, Anabela; FERNANDES, Graça, 2004. Frozen Sentences of Portuguese: Formal Descriptions for NLP. *Workshop on Multiword Expressions: Integrating Processing, International Conference of the European Chapter of the Association for Computational Linguistics*, Barcelona (Spain), July 26, 2004. – ACL: Barcelona, pp. 72-79.

BAPTISTA, Jorge; CORREIA, Anabela; FERNANDES, Graça, 2005. Léxico Gramática das Frases Fixas do Portugués Europeo. *Cadernos de Fraseoloxía Galega* 7, Santiago de Compostela, Xunta de Galicia/Centro Ramón Piñero para a Investigación en Humanidades, pp. 41-53.

CHORNOBAY, S.E., 2010a. The Main Tendencies of Phraseological Studies in Native and Foreign Linguistics. (Чернобай, С.Е. Основные тенденции изучения фразеологии в отечественной и зарубежной лингвистике). *The Culture of Black Sea Nations,* (*Культура народов Причерноморья*), 182, pp.131-134.

CHORNOBAY, Svitlana, 2010b. The Main Tendencies of Idioms Research in Foreign Linguistics. *The Culture of Black Sea Nations,* (*Культура народов Причерноморья*), 184, pp. 110-112.

CHORNOBAY, S.E., 2011a. The Definition of Idiom in Modern Phraseology. (Чернобай, С.Е. Определение идиомы в современной фразеологии). *Scientific Notes of Taurida National V.I.Vernadsky University: Philology. Social Communications.* (*Вчені записки Таврійського національного університету ім. В. І. Вернадського. Серія «Філологія. Соціальні комунікації»*). Vol. 24 (63), 3, pp. 317-322.

CHORNOBAY, S.Y., 2011b. The Role of Geographic Factor in the Formation of the British Mentality (on the Material of English Phraseology). (Чернобай, С.Е. Роль географического фактора в формировании британского менталитета (на материале английской фразеологии)). *The Culture of Black Sea Nations,* (*Культура народов Причерноморья*), 199, Vol. 2, pp.161-163.

CHORNOBAY, S.E., 2012. The Study of Phraseology within Linguocultural Paradigm. (Чернобай, С.Е. Изучение фразеологии в свете лингвокультурологической парадигмы). *The Culture of Black Sea Nations,* (*Культура народов Причерноморья*), 224, pp. 181-184.

FEDOROV, A.I., 2001. *Phraseological Dictionary of the Russian Literary Language*. (Федоров, А.И. *Фразеологический словарь русского литературного языка*). Moscow: ACT.

GRABAROVA, E.V., 2005. Concept "savoir vivre" – "ability to live". (Грабарова, Э.В. Концепт "savoir vivre" – "умение жить"). In: Karasik I.V., et al. 2005. *Other Mentality* (Карасик И.В., и др. *IIная ментальность*). Moscow: Gnozis. pp.257-333.

GROSS, M. 1996. Lexicon-Grammar. *Concise Encyclopedia of Syntactic Theories*. Cambridge. Pergamon. pp.244-258.

HNAPOV'KA, L., 2008. Lingvocognitive signs of English-speaking metaphorical representation of the concept "death". (Гнаповська, Л. Лінгвокогнітивні ознаки англомовної метафоричної репрезентації концепту «смерть»). *Scientific Notes of Kirovograd State Pedagogical University: Philological Sciences. (Наукові записки Кіровоградського державного педагогічного університету: Філологічні науки). Vol. 75 (4),* pp. 55-58.

HUMBOLDT, W. Von, 1985. *Language and Philosophy of Culture*. (Гумбольдт, В. фон Язык и философия культуры).Translated from German by I.I.Levina et al. Moscow: Progress.

KARAKEVYCH, R.O., 2009. Conceptual fields "LEBEN"/ "LIFE"in German and Ukrainian language pictures of the world: asymmetry of phraseological culturemes. (Каракевич, Р.О. Концептополя "LEBEN"/ "ЖИТТЯ" в німецькій та українській мовних картинах світу: асиметрія фразеологічних лінгвокультурем). *Materials of X International scientific practical conference "Semantics of Language and Text" 21-23 September 2009, Ivano-Frankivs'k. Vol. 2*. (Матеріали X Міжнародної науково-практичної конференції «Семантика мови і тексту» 21-23 вересня 2009 року, Івано-Франківськ. Част. 2). pp.131-135.

KUNIN, A.V., 1984. *English-Russian Phraseological Dictionary*. (Кунин, А.В., *Англо-русский фразеологический словарь*). Moscow: Russian Language.

KYTAYHORODS'KA, K., 2008. Concepts "life" – "death" in phraseological pictures of the world of the English and Ukrainian languages (Китайгородська, К. Концепти «життя» і «смерть» у фразеологічних картинах світу англійської та української мов). *Scientific Notes of Kirovograd State Pedagogical University: Philological Sciences. (Наукові записки Кіровоградського державного педагогічного університету: Філологічні науки). Vol. 75 (4)*, pp. 205-209.

SELIVANOVA, Olena, 2006. *Modern Linguistics: Terminological Encyclopoedia*. (Селіванова, Олена. *Сучасна лінгвістика: термінологічна енциклопедія*). Poltava: Dovkillya-K.

SINCLAIR, John; FOX, Gwyneth; MOON, Rosamund, *et al.*, 2001. *Collins COBUILD Dictionary of Idioms*. Glasgow: HarperCollins Publishers.

LEXICON-GRAMMAR OF RUSSIAN VERBAL IDIOMS

Tetyana Fukova
University of Algarve
tatyanafukova@gmail.com

Svitlana Chornobay
Crimean Federal
V.I.Vernadsky University
svetoniy@gmail.com

Jorge Baptista
University of Algarve
L2F/INESC-ID Lisboa
jbaptis@ualg.pt

Keywords: Russian, Verbal idioms, phraseology, lexicon-grammar, automatic identification

Abstract

This paper describes an on-going project to build a lexicon-grammar of Russian verbal idioms for natural language processing. The aim is to produce a language resource that can be used to automatically identify these idioms in naturally occurring texts. Such resource can also be useful to several fields of research, such as language acquisition, or language learning and teaching, natural language processing, among others.

1. INTRODUCTION

Though many definitions, and also many conceptual and terminological disputes, can be found in the literature around the term *idiom* (*idioms*, *collocations*, *phrasemes*, etc.), for this paper we adopt the definition of *verbal idioms* given by Baptista (2004), who also called it *frozen sentences*:

> "*Frozen sentences* are elementary sentences where the main verb and at least one of its argument are distributionally constraint, and usually the global meaning of the expression cannot be calculated from the individual meaning of its component elements when they are used independently."

For example, in Russian, an idiom like держать язык за зубами (deržat' jazyk za zubami)[34], literally 'to hold one's tongue behind one's teeth', and corresponding to the English idiom 'to keep one's tongue between one's teeth', has nothing to do with the relative position of the *tongue* and the *teeth*, but rather 'to be prudent when speaking, not saying things that should not be said'. The body-part nouns are frozen with the verb and the locative preposition. However, the nouns establish a part-whole relation with a free human noun (in the form of a possessive), as the translation illustrates, and this have referential value – possessives are obligatorily co-referent with the free human subject of the verb. Therefore, there is some syntactic structure in the idiom, and some lexical and structural variation, which must be adequately captured, and not just try to match the sequence as a

[34] In this paper, the Passport (2013), ICAO romanization norm was adopted.

whole (Baptista *et al.*, 2004; Cowie, 1998). The idiom, however, can be parsed like any ordinary, that is, semantically compositional, sentence. In fact, this expression could be uttered, and literally interpreted, in an appropriate context (for example, by a dentist to his patient). Ambiguous idioms like this also need to be specifically marked, so that the adequate reading be found from context, if possible.

The automatic identification of the meaning units in texts involves the correct delimitation and tagging of idioms in texts. Using available linguistic resources, such as phraseological dictionaries (Molotkov, 1986; Fedosov and Lapisky, 2003) and the linguistic development platform UNITEX (Paumier 2003, 2014)[35], along with the machine readable dictionary distributed with this software, we intend to determine the relevant linguistic information required to process this type of expressions, and to formalize it into a database of idioms. This we will call the Lexicon-Grammar of the Russian idioms, in the sense that it contains not only the lexical entries of the idiomatic frozen sentences, but also the relevant linguistic properties that describe their behaviour in texts.

For the remainder of this paper, we describe preliminary experiments on the application of this linguistic resource to the task of identifying the Russian idioms described in the lexicon-grammar to real texts. The paper is structured as follows: In Section 2, the methods, the process of data collection and the formal classification of the idioms are presented (§2.1). Then, the corpus collection and annotation process is explained (§2.2). Next, the reference graphs building process is presented and exemplified (§2.3). Section 3 presents the evaluation of some preliminary experiments on automatic identification of Russian idioms in the corpus. Finally, in Section 4, the paper concludes with some perspectives for future work.

2. METHODS

2.1. Data collection

To this date, we collected almost 1,000 Russian verbal idioms from phraseological dictionaries and other sources, and classified most of them using the Lexicon-Grammar framework (Gross, 1996). The idioms were formalized into a tabular format, aiming at computational processing and automatic identification in texts. This database is called a *Lexicon-Grammar* and it consist of a fine-grained description of the syntactic structure of those idioms, the lexical content of their frozen elements, the distributional constraints on their free syntactic slots (human/non-human), and the transformational properties of the construction, that is, the alternative, paraphrastically equivalent, forms (or alternations) they can yield (*e.g.* Passive).

For each idiom, a word-by-word English translation and the relevant morphosyntactic (part-of-speech) information is provided, along with a free translation (or gloss) or the English equivalent, when available or known. An illustrative, 'artificial' example, produced with just the essential elements of the idiom is also provided. In these examples, verbs are usually in the past tense; the human noun free slots are fulfilled by a proper noun, etc.

The classification here adopted is inspired in that proposed by M. Gross (1982; see a later synthesis in M. Gross 1996), and already adopted for several other languages or language varieties: French and the four main varieties, from France, Belgium, Switzerland and Québec (Lamiroy, 2010), Greek (Fotopoulou, 1993), Italian (Vietri, 2015), Portuguese, both European and Brazilian (Baptista, 2004, 2014; Vale, 2001).

[35] http://www-igm.univ-mlv.fr/~unitex/

Since Russian nominal morphology includes cases, the classification was adapted so that, instead of (or along with) prepositions, cases are used to mark the syntactic function of the verbs' arguments. Table 1 presents a gist of this classification:

Class	Structure	Example	Count
C1	N_0 *V C-acc*$_1$	*Бить баклуши* (bit' baklushi) N_0 *beat/V spoons/*C_1*-acc* to twiddle one's thumbs, to be idle	311
CP1	N_0 *V (Prep*$_1$*)* C_1	*Влететь в копеечку* (vletet' v kopeechky) N_0 *fly/V in/Prep penny/*C_1*-acc* to cost smb. a pretty penny	220
CAN	N_0 *V (C-acc N-gen)*$_1$ = N_0 *V (C-acc N-dat)*$_1$	*Капать на мозги* (kapat' na mozgi) N_0 *drop/V on/Prep brain/*C_1*-acc smb/N-dat/gen* to assert sth. over and over to s.o	50
CPN	N_0 *V Prep (C-acc N-gen)*$_1$	*Играть на нервах* (igrat' na nervah) N_0 *play/V on/Prep nerves/*C_1*-obliq N-gen* to jangle on someone's ears/nerves	6
C1PN	N_0 *V C-acc*$_1$ *(Prep*$_2$*)* N_2	*Задать пару* (zadat' paru) N_0 *set/V steam/*C_1*-acc smb/*N_2*-dat* to give smb. hell	80
CNP2	N_0 *V N-acc*$_1$ *(Prep*$_2$*)* C_2	*Взять под крыло* (vzyat' pod krilo) N_0 *take/V smb/*N_1*-acc under/prep wing/*C_2*-acc* to take smb. under one's wing	183
C1P2	N_0 *V C-Acc*$_1$ *(Prep*$_2$*)* C_2	*Брать быка за рога* (brat' bika za roga) N_0 *take/V bull/*C_1*-acc of/Prep horns/*C_2*-acc* to take the bull by the horns	99
CPP	N_0 *V w (Prep*$_1$*)* C_1*(Prep*$_2$*)* C_2	*Лезть в душу без мыла* (lezt' v dushu bez mila) N_0 *get/V into/Prep soul/*C_1*-acc without/Prep soap/*C_2*-gen* to try to gain smb.'s favor or trust by cunning	13
CADV	N_0 *V Adv*$_1$ *w*	*Выходить боком* (vihodit' bokom) N_0 *appear/V sideways/Adv* to turn out badly	24
		Total	987

Table 1. Classification of Russian verbal idioms

N0 stands for the subject, *N1* and *N2* for the first and second complement, respectively. C_1 and C_2 indicate the constant (frozen) element of the complement; *Prep* is a preposition and *Adv* and adverb; w represents an unspecified sequence of complements; the cases are shortened: *acc* for accusative, *dat* for dative and *gen* for genitive.

Notice that in class CP1, the first complement is either introduced by a preposition or it does not receive the accusative case (*acc*). The same happens in all the remaining classes. No cases were yet found of frozen subject.

The work of collecting the idioms from dictionaries and other sources is still going on but, based on the length of the dictionaries already consulted, a reasonable estimate would consider around 2 thousand and up to 3 thousand of frequently occurring, frozen verbal idioms to attain a reasonable lexical coverage.

2.2. Corpus collection and annotation

For this paper, we used the Russian National Corpus[36], henceforward *RNC*, which covers primarily the period from the middle of the 18th to the early 21st centuries. According to the corpus' website:

> "[t]his period represents the Russian language of both the past and the present in a wide range of sociolinguistic variants: literary, colloquial, vernacular, in part dialectal. The corpus includes original (non-translated) works of fiction (prose, drama and poetry) of cultural importance, which are interesting from a linguistic point of view. Apart from fiction, the corpus includes a large volume of other sources of written (and, for the later period, spoken) language: memoirs, essays, journalistic works, scientific and popular scientific literature, public speeches, letters, diaries, documents".

The full RNC corpus contains 85,996 documents, 19,362,746 sentences and 229,968,798 words. We chose the 10 most frequent verbs from the Lexicon-Grammar table, excluding verbs that are often support verbs, namely брать (brat) 'to take' (26 entries), давать (davat) 'to give' (19), and делать (delat) 'to do' (12). The most frequent verbs are : держать (derzhat) 'to hold', идти (idti) 'to go', играть (igrat) 'to play', бить (bit) 'to beat', смотреть (smotret) 'to look', класть (klast) 'to put', лезть (lezt) 'to climb', лежать (lezhat) 'to lie', выйти (viiti) 'to go out', жить (zhit) 'to live'.

Using the RNC web interface in its default options, several queries were conducted, extracting the sentences where each one of these verbs (the lemma and the associated inflected forms) occur (we called this **Corpus 1**). Next, we retrieved from the website the top search results (about 2,000 sentences) in a spreadsheet format. Random numbers were accorded to the sentences and, after sorting them, the first 50 were selected. We then manually annotated the idioms found, by delimiting the verb and the frozen elements with underscore, '_'. Table 2 shows the results from this data collection and annotation process.

[36] http://www.ruscorpora.ru/

verb	translit	gloss	RNC	in sample (n=50)	diff. idioms	diff. LG entries	Total LG entries w/ V
держать	derzhat'	to hold	50 643	5	4	4	33
идти	idti	to go	241 225	1	1	1	23
играть	igrat'	to play	66 077	0	0	0	14
бить	bit'	to beat	33 393	6	5	4	12
смотреть	smotret'	to look	157 516	0	0	0	11
класть	klast'	to put	10 458	1	1	1	11
лезть	lezt'	to climb	11 273	3	3	3	9
лежать	lezhat'	to lie	80 235	2	2	2	9
выйти	viiti	to go out	165 311	1	1	1	8
жить	zhit'	to live	187 841	2	2	2	7
		Total	**1 003 972**	**21**	**19**	**18**	**137**

Table 2. Corpus 1 data collection (from Russian National Corpus).

The 500 instances of these 10 verbs retrieved from the corpus constitute just a small sample, which cannot be considered representative of the total of occurrences of those verbs in RNC corpus. Our point was to have a glimpse of how complete the Lexicon-Grammar may already be. Still, 4.2% of the retrieved instances contained idioms, attesting the frequency of the phenomenon. The 21 instances correspond to 19 different idioms, which signals few repetitions. These idioms, in their turn, represent only 18 entries on the Lexicon-Grammar, but only one idiom, бить поклоны (bit' pokloni) 'pay smb. one's respects', had not been previously registered, thus indicating adequate lexical coverage. Still, the 18 entries found are just a fraction (13%) of all the entries involving those same verbs that had already been collected in the Lexicon-Grammar.

In a second moment, we retrieved from the RNC all the sentences containing the verbs of the previous list and the first frozen nominal element of the 137 idioms listed in the Lexicon-Grammar involving those verbs, allowing for a window from 0 up to 3 intervening words. As in the sampling procedure above, we randomly selected the top 50 sentences; if the result yields less than 50 occurrences, all of them were considered. These sentences constitute **Corpus 2**. This second sample also underwent manual classification of the idioms found therein, using the same delimitation procedure, with underscores '_'. Since some expressions can be ambiguous between a literal and a figurative (idiomatic) interpretation, literal expressions were delimited with '#'. Results of this procedure can be found in Table 3.

Verb	Translit	Gloss	diff. idioms	matches	idioms/50
Держать	derzhat'	'hold/keep'	29	1,048	594
Идти	idti	'go'	23	988	490
Играть	igrat'	'play'	13	468	189
Бить	bit'	'beat'	11	501	295
Смотреть	smotret'	'look'	9	296	158
Класть	klast'	'put'	9	229	71
Лезть	lezt'	'climb'	5	225	154
Лежать	lezhat'	'lie'	6	253	152
Выйти	viiti	'go out'	6	229	151
Жить	zhit'	'live'	6	193	80
			117	4,430	2,334

Table 3. *Corpus 2* data collection (from Russian National Corpus) (extract)

In the Appendix, the full list of the 117 idioms is presented, with the breakdown of matches

2.3. Building reference graphs

In order to automatically retrieve from texts the idioms represented in the Lexicon-Grammar in a tabular format, we used UNITEX (Paumier 2003, Paumier 2014), an open-source linguistic development platform. One of its functionalities is to intersect data matrices with finite-state transducers (FST), which can then be used to match and label complex patterns in texts. This is done by first building reference graphs, which are Directed Acyclic Graphs (DAGs) where each variable (noted @X) refer to the corresponding column in the matrix. Figure 1 illustrates the reference graph for class **C1**:

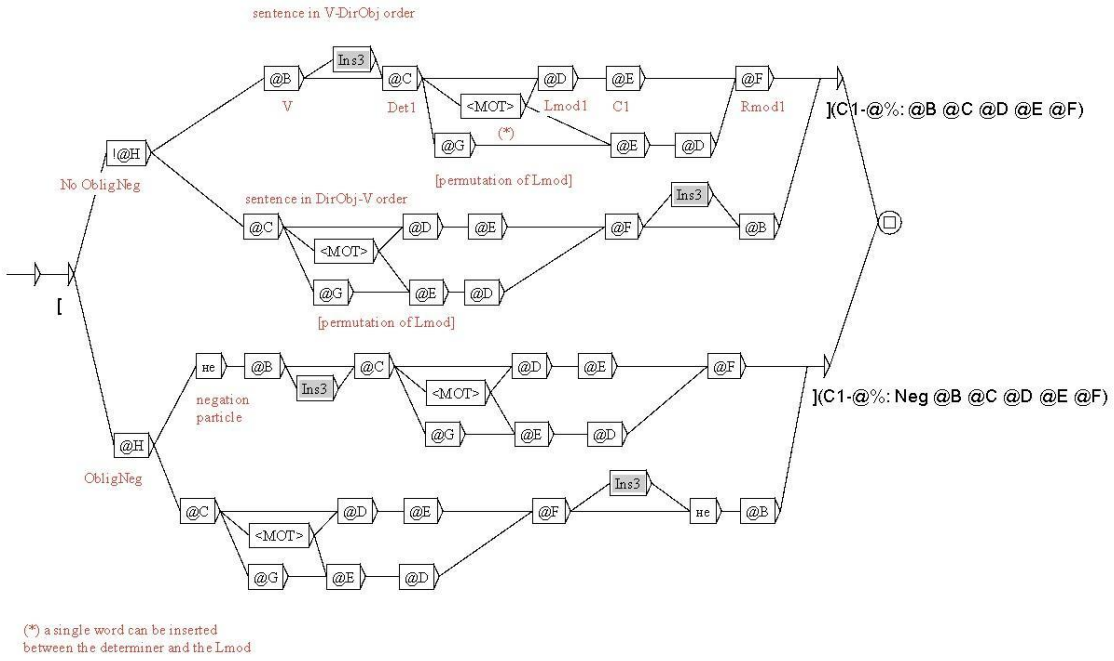


Fig. 1. Reference graph for class C1, e.g. Бить баклуши, *N_0 beat/V spoons/C_1-Acc* 'twiddle one's thumbs'

The system reads each line of the matrix in turn, and then builds a graph for each expression, replacing the variables by the content of the corresponding cell in the matrix. The graph in Fig. 1 reads as follows: Variable @H corresponds to the binary property 'obligatory negation', e.g. Бровью не повести (brov'u ne povesti) *N_0 not_stir*/V *eyebrow*/N-instrum 'do not display emotions toward smth.' This property is encoded by '+' and '-'; the graph is split in two paths, the upper path for the expressions that do not allow for this property (the symbol '!' indicates negation) and the lower path for those that do. The negative particle не (ne) explicitly appears in the lower path. When the system reads variable @H, it follows and builds a graph for the remainder of the path if this variable, at the given line in the matrix, is marked with '+' (the reverse happens if it is marked with negation: !@H); otherwise, it collapses the graph at that point.

Next, the sequence of elements constituting the idiom are represented by the variables @B for the verb, @C for the determiner, @D for the left modifier (LMod), @E for the frozen noun of the direct object (C_1), and @F for the right modifier (RMod); an insertion window from 0 to 3 words is represented by auxiliary sub-graph Ins3 (grey box).

The transformation consisting in the alternative ordering of the left modifier is encoded by the property represented by variable @G, so that in this case the intervening elements are permuted. Finally, a single word may be inserted between the determiner and the left modifier, which is represented by the node <MOT>.

The transducer's *output* is represented below the nodes. This consists in the square brackets used as delimiters for the matching strings, and the codes for the class and the line number of each idiom (provided by variable @%, indicating the current line). In this output, we also added (inside brackets) the main elements of the idiom.

A similar procedure was carried out for the remaining classes. The system generates a sub-graph for every line of the matrix Fig. 2 shows two sub-graphs from class CP1, one

with obligatory negation. All the sub-graphs of the same class are automatically gathered in a resulting graph, which can then be applied to texts.

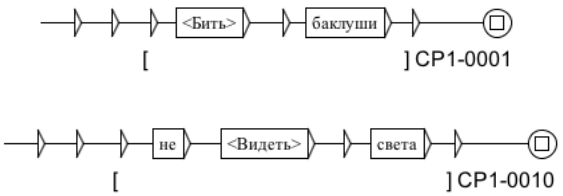


Fig. 2. Two sub-graph from class CP1, e.g. *Влететь в копеечку* (vletet' v kopeechky) N_0 *fly*/*V in*/*Prep penny*/C_1-*acc* 'to cost smb. a pretty penny'

3. EVALUATION

In order to do some preliminary evaluation of the Lexicon-Grammar and the corresponding finite-state transducers built for the automatic identification of idioms in texts, we first apply them to the two lists of examples that were produced for each idiom. Two formats of examples were produced: (a) 'lexical entry' examples, where there is neither a subject nor other free arguments, just the verb in the infinitive and its frozen complements, like one would find the idiom in a dictionary; and (b) 'artificial examples', produced as described above (section §2.1), with the verb in an inflected form, a proper noun as a human subject, etc. The first list of examples is meant to test the FSTs while ignoring inflection, agreement and word order issues. To a limited extent, the second list of examples highlights the impact of these issues to the language-processing task. Table 4 shows the results of these two experiments:

examples (N=312)	**matches**	**precision**
lexical entries	278	0.89
artificial examples	248	0.79

Table 4. Evaluation of class C1

Precision of the results is relatively high. Actually, most word combinations in the dictionary of frozen sentences are unique, therefore, unequivocal combinations. However, some matches were not found and defined as idioms. One of the reasons for not recognizing idioms in the text is the lack of a verb or a noun (frozen component), or one of their inflected forms, in the dictionary. For instance, in the example *лежать лежнем* 'lezhat lezhnem" (N_0 lie_V $ledger_{N\text{-}Instrum}$), which means 'to be in poor health', (the inflected form of) the noun *лежень* lezhen (ledger) is not present in the dictionary, and as the Lexicon-Grammar encoded C_1 by its lemma < *лежень* >, the system fails to recognize the inflected form *лежнем*. A solution is to consider the word forms only, but that cannot always be the case, as some constants do inflect, while the verb is always inflected.

Another reason for not recognizing idioms is the passive voice, which is expressed by the suffix *ся* 'sya' of verbs. These verb forms have been attributed to different lemmas from the active forms, in the dictionary produced with the system, and therefore they were not detected. In order to make such verbs not to have several lemmas, we made some changes

in the dictionary. We extracted all the verbs that ended on the suffix *ся* (sya), added lemma without this suffix and appended this grammatical feature to those entries.

4. CONCLUSIONS AND FUTURE WORK

To the best of our knowledge, there is no available machine-readable lexicon of Russian verbal idioms. Idioms represent a significant part of the lexicon units of any language. The identification of frozen sentences is a necessary part of the lexical analysis of texts for natural language processing, as they constitute units on meaning.

In this paper, we have presented the current state of an on-going project of building a Lexicon-Grammar of Russian verbal idioms, containing around 1,000 entries, described the information there encoded, and the finite-state tools built to identify those expressions in texts. The preliminary tests of the examples have shown that this is no small task, as a number of other language-processing issues have to be dealt with, besides the idioms' linguistic description. Among them, there is the degree of lexical coverage of the base dictionary and the quality and granularity of the information there represented, something outside the scope of the paper. Even so, adaptation and extension of the current, freely distributed, machine-readable dictionary of the Russian language is underway, like the treatment given to passive verb forms. On the other hand, refining of the reference graphs will allow to deal with several syntactic issues not yet addressed, like the interaction of obligatory negation with passive or the intrinsically reflexive constructions. The free-order syntax of sentence constituents in Russian also produces highly complex reference graphs that require a systematic verification of all possible combinatory variants. The paper also produced two annotated sub-corpora, extracted from the Russian National Corpus, where verbal idioms were manually delimited, including literal uses of ambiguous expressions. These resources can now be used as a workbench to help develop and test several procedures for automatic detection and delimitation of Russian verbal idioms.

Acknowledgements

Research for this paper was partially funded by a European Union Erasmus Mundus (EMA2) scholarship through the BMU-MID program, and by Portuguese national funds through Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2014.

References

BAPTISTA, J. 2004. Compositional vs. Frozen Sequences. Laporte, Eric; Ting Au-Chen, (eds).Proceedings of the Lexicon-Grammar Workshop. Beijing 14-18 de Outubro de 2004. *Journal of Applied Linguistics*, Special Issue on Lexicon-Grammar. Papers presented at the Lexicon-Grammar Workshop, pp. 81-93 (Chinese version).

BAPTISTA, J., CORREIA A. AND FERNANDES G. 2004. Frozen Sentences of Portuguese: Formal Descriptions for NLP. *Workshop on Multiword Expressions: Integrating Processing, International Conference of the European Chapter of the Association for Computational Linguistics*, Barcelona (Spain), July 26, 2004. ACL: Barcelona, pp. 72-79.

CHORNOBAY, S., BAPTISTA, J. 2014. Semantic Peculiarities of Portuguese and Russian Idioms within the Conceptual Domain "Death". International Scientific Conference Modern Philology:

Paradigms, trends, problems (Міжнародна наукова конференція "Сучасна філологія: парадигми, напрямки, проблеми"), October 9, 2014, Kyiv, Kyiv National Taras Shevchenko University.

COWIE, A. 1998. *Phraseology. Theory, analysis, and applications*. Oxford: Oxford University Press.

FEDOSOV, I. and LAPITSKY, A. 2003. *Phraseological dictionary of the Russian language*. (Федосов, И. и Лапицкий, А. *Фразеологический словарь русского языка*). Moscow: Unves.

FOTOPOULOU, A. 1993. Une classification des phrases a complèments figés en grec moderne: étude morphosyntaxique des phrases figées . Ph.D. thesis, Université Paris VIII.

GROSS, M. 1982. Une classification des phrase "figées" du français. *Revue Québécoise de Linguistique* 11-2: pp. 151-185.

GROSS, M.1996. Lexicon-Grammar. *Concise Encyclopedia of Syntactic Theories*. Cambridge. Pergamon. pp.244-258.

LAMIROY, B., KLEIN, J.-R. 2010. Lamiroy, Béatrice, and Jean-René Klein. *Les expressions verbales figées de la francophonie: Belgique, France, Québec et Suisse*. Editions OPHRYS.

MOLOTKOV, A. 1986. *Phraseological dictionary of the Russian language* (Молотков, А., *Фразеологический словарь русского языка*). Moscow: ACT.

PAUMIER, S. 2003. De la reconnaissance des formes linguistiques à l'analyse syntaxique. PhD thesis, Université de Marne-la-Vallée, 2003.

PAUMIER, S. 2014. *Unitex 3.0 - User's Manual*. Paris: Université Paris-Est Marne-la-Vallée.

VALE, O. 2001. *Expressões Cristalizadas do Português do Brasil: Uma proposta de tipologia*. Araraquara, SP (Brasil): Universidade Estadual Paulista.

VIETRI, S. 2015. *Idiomatic Constructions in Italian. A Lexicon-Grammar Approach*. Amsterdam: John Benjamins.

APPENDIX

In this table, for the 137 entries of the Lexicon-Grammar of Russian Verbal Idioms that are formed with the 10 most frequent verbs (excluding some support/auxiliary verbs), we indicate the 117 that were found, their transliteration and a gloss, along with (last two columns): (a) the total number of matches extracted from the Russian National Corpus with the queries consisting of the *verb* and the head noun of the first frozen complement (C_1), allowing for a window of 0 up to 3 intervening words (V-C_1) – these matched sentences constitute *Corpus 2*; and (b) the number of idioms found therein; if the total of V-C_1 matches is larger than 50, only a random sample of 50 matches was used; if they are less than 50, the entire V-C_1 list of matches was used.

Idiom (V-C_1)	**Transliteration**	**Gloss**	**# Corpus 2**	**idioms/50**
Держать в курсе	derzhat' v kurse	keep informed	324	26
Держать в уме	derzhat' v ume	carry smth. in one's head / mind	122	44
Держать в черном теле	derzhat' v chernom tele	keep smb on short rations	88	17
Держать верх	derzhat' verkh	gain the upper hand	16	10
Держать вожжи в руках	derzhat' vozhi v rukah	concentrate power or leadership in one's hands, taking full control of sth	1	1
Держать дверь открытой	derzhat' dver' otkritoi	be hospitable	90	6

Idiom (*V-C*$_1$)	**Transliteration**	**Gloss**	**# Corpus 2**	**idioms/50**
Держать камень за пазухой	derzhat' kamen' za pazukhoi	hold a stone inside one's shirt	43	19
Держать кого-то в ежовых рукавицах	derzhat v ezhovih rukavitsakh	keep or assume strict control over s.o.	35	28
Держать кого-то на прицеле	derzhat' na pritsele	keep in mind something or someone	62	12
Держать кого-то под колпаком	derzhat' pod kolpakom	keep a check on smb	8	5
Держать марку	derzhat' marku	maintain one's reputation	64	46
Держать на почтительном расстоянии	derzhat' na pochtitelnom rasstoyanii	keep at a respectful distance	49	30
Держать на привязи	derzhat' na privyazi	restrict s.o. s independence, not allow s.o. to act on his own initiative	66	38
Держать нос по ветру	derzhat' nos po vetru	trim the sails to the wind	83	19
Держать пари	derzhat' pari	to bet on	250	50
Держать под каблуком	derzhat' pod kablukom	keep under thumb	3	2
Держать под своим крылом	derzhat' pod svoim krilom	have smb. under one's wing, take care of smb.	20	1
Держать порох сухим	derzhat' poroh sukhim	be prepared to defend one's country (cause etc) at any moment	7	7
Держать руки по швам	derzhat' ruki po shvam	stand to attention	59	48
Держать себя	derzhat' sebya	act in a certain manner	3633	20
Держать себя в рамках	derzhat' sebya v ramkakh	behave in a restrained, correct manner	20	15
Держать себя в руках	derzhat' sebya v rukakh	To hold/keep oneself in check; to keep oneself in hand.	7505	1
Держать себя в узде	derzhat' sebya v uzde	to refrain from emotional outbursts	95	47
Держать ухо востро	derzhat' ukho vostro	have an ear to the ground	283	29
Держать хвост трубой	derzhat' khvost truboi	keep cheerful, keep one's chin / spirits up	16	8
Держать шаг	derzhat' shag	march in step	22	14
Держать язык за зубами	derzhat' yazik za zubami	keep one's tongue between one's teeth	216	42
Держать язык на привязи	derzhat' yazik na privyazi	to remain silent	66	9
Держаться за юбку	derzhatsya za yubku	be completely under the control of, and fully dependent on, some woman (one's wife, mother etc)	8	0

Idiom (*V-C*$_1$)	**Transliteration**	**Gloss**	**# Corpus 2**	**idioms/50**
Идти прахом	idti prakhom	go to rack and ruin	26	25
Идти прямой дорогой	idti pryamoi dorogoi	take the shortest route (to)	1624	1
Идти своей дорогой	idti svoei dorogoi	go one's own way	1624	3
Идти против течения	idti protiv techeniya	to go against the tide/current/stream	187	22
Идти в разрез	idti v razrez	act in defiance to someone or something	66	48
Идти в ногу	idti v nogu	to progress at a pace equal to that of (another person, branch of technology etc, or the times in general)	584	22
Идти впрок	idti vprok	to be beneficial to s.o.	50	50
Идти врозь	idti vroz'	go apart	14	14
Идти (кому-то) навстречу	idti navstrechu	to satisfy s.o.'s needs, requests, desires	1880	20
Идти насмарку	idti nasmarku	to yield no positive result, end in nothing	30	29
Идти в гору	idti v goru	to improve one's status or job, gain influence, importance, succeed in one's career	574	12
Идти в дело	idti v delo	to be or start to be used	1211	8
Идти ко дну	idti ko dnu	be almost ruined	196	15
Идти на лад	idti na lad	take a turn for the better	81	49
Идти на поводу	idti na povodu	to submit fully to s.o., not act on one's own initiative	174	41
Идти на попятную	idti na popyatnuyu	to retreat from a decision made earlier, go back on an agreement	20	18
Идти по линии наименьшего сопротивления	idti po linii naimenshogo soprotivleniya	to choose the easiest course of action, avoiding difficulties, trouble	334	1
Идти по миру	idti po miru	to lead a beggarly life	288	4
Идти под венец	idti pod venets	to marry	40	36
Идти под гору	idti pod goru	to deteriorate sharply	574	1
Не идти на ум	ne idti na um	the thought of sth. does not occupy s.o.'s mind	102	25
Идти на удочку	idti na udochku	to end up being deceived, duped	8	4
Идти по стопам	idti po stopam	to follow after s.o. in doing sth., looking to him as an example	147	42
Играть первую скрипку	igrat' pervuyu skripku	to occupy the leading, predominant position in sth	449	9

Idiom (*V-C*$_1$)	**Transliteration**	**Gloss**	**# Corpus 2**	**idioms/50**
Играть глазами	igrat' glazami	to glance at s.o. in a flirtatious manner, trying to gain his or her interest	139	15
Играть жизнью и смертью	igrat' zhizn'yu I smert'yu	to put o.s. in a life-threatening situation, disregarding danger	238	1
Играть роль	igrat' rol'	to have a certain (as specified) meaning, significance, influence	6704	47
Играть словами	igrat' slovami	to speak evasively, using imprecise, ambiguous language	173	20
Играть на руку	igrat' na ruku	help s.o. or further sth. by one's actions, often without being aware of doing so	264	21
Играть в бирюльки	igrat' v birul'ki	to occupy o.s. with trifles	9	9
Играть в загадки	igrat' v zagadki	to speak evasively, enigmatically, in a roundabout way	7	4
Играть в молчанку	igrat' v molchanku	to keep silent, avoid conversation	17	16
Играть в прятки	igrat' v pryatki	to hide the truth by being evasive	127	18
Играть в струнах души	igrat' v strunakh dushi	to touch smb's feelings	43	5
Играть с огнём	igrat' s ognem	To play with edge-tools; to play with fire.	12	1
Играть на нервах	igrat' na nervakh	to irritate s.o	30	23
Бить баклуши	bit' baklushi	to be idle, do nothing	61	61
Бить копытом	bit' kopitom	be hot to trot	142	5
Бить тревогу	bit' trevogu	to draw attention to impending danger, trouble etc, calling for measures to suppress or fight it	196	21
Бить фонтаном	bit' fontanom	emerge rapidly	167	14
Бить наповал	bit' napoval	act in a fashion that guarantees success, rules out any possibility of failure	27	12
Бить по глазам	bit' po glazam	strike the eye	361	41
Бить по карману	bit' po karmanu	cost a pretty penny	24	14
Бить кого-то смертным боем	bit' smertnim boem	beat smb. ruthlessly	59	12
Бить в набат	bit' v nabat	to draw general attention persistently to sth. alarming, to impending danger	86	31
Бить в цель	bit' v tsel'	to be effective, attain the desired result	58	34
Бить через край	bit' cherez krai	to manifest itself, come forth with great force, in abundance	86	50
Волком смотреть	volkom smotret'	look angrily, in a hostile way	51	25

Idiom (*V-C*$_1$)	**Transliteration**	**Gloss**	**# Corpus 2**	**idioms/50**
Смотреть бирюком	smotret' birukom	to look gloomy, morose	2	2
Смотреть женихом	smotret' zhenihom	look very happy, content	26	0
Смотреть зверем	smotret' zverem	be gloomy, morose	45	6
Смотреть именинником	smotret' imeninnikom	look very happy	5	5
Смотреть правде в (глаза)	smotret' pravde v glaza	see things or evaluate facts as they really are, look at things clearly	120	34
Смотреть смерти в глаза	smotret' smerti v glaza	be exposed to mortal danger	89	46
Смотреть сквозь пальцы	smotret' skvoz' paltsi	intentionally not to react to sth., as if accepting it by allowing it to go on	296	29
Смотреть со своей колокольни	smotret' so svoei kolokol'ni	to make one-sided judgments about s.o. or sth. based entirely on one's own limited perspective	18	11
Класть голову	klast' golovu	sacrifice one's life for smb., smth.	260	3
Класть оружие	klast' oruzhie	stop fighting for sth., abandon some course of action, admit defeat	28	3
Класть деньги (в кубышку, на бочку)	klast' den'gi (v kubishku, na bochku)	accumulate money (keeping it at home, not investing it)	78	0
Класть зубы на полку	klast' zubi na polku	go hungry; tighten one's belt	16	9
Класть душу	klast' dushu	To put one's heart and soul into something	21	7
Класть на бумагу	klast' na bumagu	to be expressed in written form	37	31
Класть на лопатки	klast' na lopatki	coll to win a victory over s.o. (in an argument, contest etc)	12	7
Класть на музыку	klast' na musiku	to write music for some verses or text	7	6
Класть под сукно	klast' pod sukno	to postpone making a decision on some matter	8	5
Лезть из кожи вон	lezt' iz kozhi von	To lean/bend over backwards; to go out of one's way; to go all out.	90	23
Лезть в глаза	lezt' v glaza	to attract attention to o.s. or itself, be noticeable	188	34
Лезть в голову	lezt' v golovu	creep into one's head	312	41
Лезть в душу	lezt' v dushu	(try to) worm oneself into smb's confidence	62	38
Лезть в бутылку	lezt' v butilku	To fly off the handle; to blow one's top.	25	18
Лежать бревном	lezhat' brevnom	lie like a log	87	46

Idiom (*V-C$_1$*)	**Transliteration**	**Gloss**	**# Corpus 2**	**idioms/50**
Лежать мертвым грузом	lezhat' mertvim gruzom	to go unused, be without application	36	16
Лежать пластом	lezhat' plastom	(usu. of a person who is sick or very tired) to lie completely immobile	154	44
Лежать на печи	lezhat' na pechi	to do nothing, be idle	103	27
Лежать на поверхности	lezhat' na poverhnosti	be obvious, evident	245	12
Лежать под сукном	lezhat' pod suknom	(of an application, request, complaint etc) not to be given any attention, not be processed	17	7
Выйти сухим из воды	viiti sukhim iz vodi	to escape well-deserved punishment, remain unpunished, uncompromised	364	23
Выйти в люди	viiti v ludi	make one's way (in life)	1475	45
Выйти из возраста	viiti iz vozrasta	be past the age (for)	82	33
Выйти из пеленок	viiti iz pelenok	be no longer a kid / baby	20	17
Выйти из-под кисти	viiti iz-pod kisti	be painted (by)	9	7
Выйти из-под пера	viiti iz-pod pera	come from smb.'s pen	132	26
Жить барином	zhit' barinom	live a life of ease	96	18
Жить бирюком	zhit' biryukom	to be unsociable, live in seclusion	6	4
Жить минутой	zhit' minutoi	live in the moment	108	4
Жить нараспашку	zhit' naraspashku	live openly and widely	10	9
Жить чужим (своим)умом	zhit' chuzhim (svoim)umom	let others think for oneself	139	24
Жить на вулкане	zhit' na vulkane	to be living on the edge of a volcano; to be sitting on a volcano.	27	21

PAREMIOLOGÍA BASADA EN CORPUS WACKY: ENFOQUE (INTRA- E INTER-) LINGÜÍSTICO Y CONCEPTUAL

Vincenzo Lambertini
Università di Bologna
Dipartimento di Interpretazione e Traduzione,
sede di Forlì
vincenzo.lambertini2@unibo.it

Resumen

Este estudio, que forma parte de un proyecto de investigación doctoral llevado a cabo en el Departamento de Interpretación y Traducción (DIT) de Forlì (Universidad de Bolonia), se propone categorizar de manera conceptual los refranes italianos de uso común, intentando facilitar una herramienta de gran utilidad para intérpretes, traductores y hablantes que quieran expresar determinados conceptos utilizando refranes.

Los obstáculos encontrados en esta investigación han sido muchos, ya que los estudios lingüísticos sobre paremias escasean.

No obstante, gracias al corpus italiano itWaC, se pudo hallar una metodología de identificación automática de paremias, además de llevar a cabo un estudio lingüístico sobre refranes y sus modificaciones.

Las etapas futuras consistirán en realizar una categorización conceptual de los refranes y aplicar los mismos principios de identificación automática de paremias y análisis conceptual a refranes de otros idiomas, para realizar un recurso multilingüe, utilizable de manera semasiológica para recuperar refranes en diferentes lenguas.

1. INTRODUCCIÓN

Este artículo está encaminado a describir la primera fase de un estudio de investigación que se propone categorizar de manera conceptual los refranes italianos de uso común, con el fin de facilitar una herramienta de gran utilidad para intérpretes, traductores y hablantes que quieran expresar determinados conceptos utilizando refranes.

Para darse cuenta de las dificultades relacionadas con el estudio de los refranes y su categorización, es necesario dar un paso atrás y centrarse en aspectos teóricos de suma importancia.

En primer lugar, cabe destacar la cuestión de la fijación de los refranes: según numerosos investigadores, entre los cuales hemos de recordar a Jean-Claude Anscombre, los refranes "no son expresiones fijas ni tampoco giros idiomáticos" (Asncombre, 1997, p.52).

En segundo lugar, la falta de fijación implica, en la mayoría de los casos, que el sentido de un enunciado es composicional, es decir se deduce del significado de sus partes y de las reglas con las que se combinan (Casadei, 1996; véase también Katz, 1973). Es evidente que los refranes no son frases idiomáticas, pero tampoco son frases composicionales ni siquiera

funcionan semánticamente como secuencias libres. Es suficiente poner un ejemplo de correspondencia interlingüística de refranes (español, francés e italiano) para demostrar su funcionamiento no composicional.

(1a) Hablando del rey de Roma por la puerta asoma

(1b) *Quand on parle du loup on en voit la queue* (trad. lit.: "Cuando se habla del lobo se le ve la cola")

(1c) *Si parla del diavolo e spuntano le corna* (trad. lit.: "Se habla del diablo y aparecen los cuernos")

Como se puede observar, los tres refranes tienen sentidos literales muy distintos. No obstante, al considerar sus definiciones, se aprecia que los tres refranes tienen el mismo sentido, puesto que apuntan a la misma situación genérica, que podríamos resumir citando la explicación del Refranero Multilingüe (2015): "se emplea cuando se presenta inesperadamente la persona de la que se está hablando".

Como se deriva de lo argumentado hasta aquí, partimos de la idea de que los refranes tienen cierto grado de fijación y no significan de inmediato lo que pretenden comunicar composicionalmente. De hecho, transmiten un mensaje metafórico, y más precisamente genérico, que tiene rasgos invariados en distintos contextos.

En el presente artículo, se describen en primer lugar las dificultades con las que nos encontramos al intentar identificar automáticamente refranes. A continuación, se explica el modo cómo logramos identificar refranes italianos y sus versiones modificadas. Por último, se exponen brevemente las conclusiones a las que se ha llegado por el momento y los proyectos de investigación futuros.

2. IDENTIFICACIÓN AUTOMÁTICA DE REFRANES

En muchas ocasiones, los estudios sobre refranes prescinden de datos reales, descuidando las características lingüísticas (semánticas, sintácticas y morfológicas) y concentrándose más en lo folklórico, histórico y cultural. El límite de estos planteamientos es que no consideran los refranes elementos empleados en el habla contemporánea, mientras que en muchos ámbitos, como en el lenguaje publicitario o periodístico, parecen imprescindibles. Por esta razón, es preciso analizar con detenimiento los refranes en corpus lingüísticos.

No cabe duda de que no es sencillo encontrar automáticamente refranes en corpus de grandes dimensiones, pero tampoco es imposible. En nuestro trabajo, decidimos tomar como referencia las investigaciones sobre metáforas[37], pues comparten diversos aspectos con los refranes.

En principio, hay varios métodos para identificar automáticamente metáforas en corpus. Por ejemplo, se puede analizar manualmente un corpus y decidir si una palabra se utiliza de manera metafórica. Sin embargo, este método no se puede llevar a cabo con corpus de grandes entidades, imprescindibles para estudiar refranes, ya que su frecuencia de uso no es muy elevada. Otro sistema podría consistir en buscar palabras o unidades fraseológicas utilizando lemas precisos. El límite de este método es que el investigador interfiere demasiado en la búsqueda y no permite al corpus producir autónomamente los resultados,

[37] Por ejemplo, véase Deignan (2010).

lo que en cambio debería ocurrir de acuerdo con los principios clave de la lingüística de corpus (Tognini-Bonelli, 2001).

Para intentar no influir en la búsqueda de refranes, decidimos tomar una lista de más de 25.000 refranes italianos sacados del *Dizionario dei proverbi italiani* de Carlo Lapucci (2006), una de las obras más completas sobre el acervo paremiológico italiano, que además numera cada uno de los refranes. Generamos, pues, una lista de 500 números casuales y tomamos en consideración 500 refranes correspondientes a esos números casuales. Al final, descubrimos que solo un 4,16% de los 500 refranes detectados casualmente estaba presente en nuestro corpus, lo que indicó que esta metodología no podía compaginarse con nuestro estudio.

Según los estudios sobre metáforas, otra manera de encontrar metáforas en corpus consiste en la búsqueda de los llamados "marcadores de metáforas", a saber aquellas expresiones que introducen una metáfora (Deignan, 2010, p.17). Se trata de locuciones como "una suerte de", "por así decirlo", etc. Este método no garantiza la recuperación automática de todas y cada una de las metáforas presentes en un corpus, pero sí permite identificar metáforas sin que el investigador interfiera en este proceso.

Es posible aplicar estos principios al estudio sobre refranes, dado que, como subraya Shapira (2000, pp.89-90), las paremias pueden ser introducidas por expresiones que señalan su carácter paremiológico. Sin embargo, una vez más, los estudios lingüísticos sobre refranes en italiano escasean, así que no hay datos ciertos sobre los "marcadores de refranes" italianos más frecuentes. Es más, al poner en práctica estos principios se notó que con expresiones introductorias muy generales (por ejemplo: "*come si suol dire*", "*come si dice*", etc.) no solo se obtienen refranes, sino también otros tipos de fraseologismos y frases sentenciosas, como expresiones idiomáticas, máximas, aforismas, dichos, etc.

Para superar esta dificultad, se decidió buscar la palabra "*proverbio*"[38] que está contenida con una probabilidad elevada en expresiones como: "*come dice il proverbio*"; "*c'è un proverbio che dice*"; "*come recita il proverbio*"; etc. Se trata, pues, de la única parte que está presente con elevada frecuencia y probabilidad en los marcadores de refranes.

3. TIPOLOGÍA DEL CORPUS

El corpus itWaC (Baroni, et al., 2008) forma parte de una serie de corpus de varios idiomas realizados a partir de textos recuperados en la Web reunidos bajo la etiqueta "Corpus WaCky", lo que es acrónimo de "*The Web-As-Corpus Kool Yinitiative*". El corpus italiano itWaC se compone de dos mil millones de palabras y está formado por textos recuperados en la Web con dominio .it, utilizando como *seeds* las palabras de frecuencia mediana del corpus "*La repubblica*" y un vocabulario básico de lengua italiana.

Además de ser un corpus actual que abarca varias tipologías de textos, que van desde los foros hasta los periódicos electrónicos, este corpus está etiquetado con información morfosintáctica (o *POS tagging*), lo que resultará muy útil a la hora de buscar refranes, puesto que, como se verá más adelante, estos pueden variar mucho.

Finalmente, este corpus puede ser consultado gracias al programa gratuito NoSketch Engine que está disponible en la dirección: http://nl.ijs.si/noske /wacs.cgi/first_form.

[38] "*Proverbio*" corresponde al español "refrán" y no a "proverbio". La diferencia entre el "*proverbio*" italiano y el "proverbio" español es que el segundo indica un tipo de paremia culta (Sevilla Múñoz, 2000, p.101), mientras que el primero no tiene un sentido tan restrictivo.

Además, es posible abrir el corpus itWaC directamente desde esta página, seleccionándolo en el menú desplegable "Corpus".

4. METODOLOGÍA DE BÚSQUEDA DE REFRANES EN ITWAC

Al insertar la palabra "*proverbio*" en la interfaz de consulta del corpus itWaC, se obtienen 3706 resultados que corresponden a las concordancias de la palabra "*proverbio*". Esto significa que en el contexto de cada uno de estos ejemplos se habla de refranes y muy probablemente están presentes refranes. Sin embargo, no es fácil validar de manera electrónica los resultados, ya que el procedimiento que tendría que realizar la máquina sería demasiado complejo. Por esta razón, fue necesario validar cada uno de los resultados manualmente, para distinguir entre refranes utilizados en contexto y refranes de los que se habla sin que estén presentes, además de otros casos que no son relevantes para este estudio.

En la fase de validación, se agruparon los resultados obtenidos en seis clases, dependiendo de la presencia / ausencia de los refranes y también de su tipología. De este modo, se observó que el 92% de los resultados analizados contenía refranes y el 8% restante no presentaba ningún refrán. Es más, notamos que un 64% consistía en refranes utilizados en contexto, dividido en un 28% de refranes italianos y un 36% de refranes no italianos (citados en lengua original o traducidos literalmente al italiano), lo que es muy llamativo y necesitaría ser objeto de estudios más específicos. El 8% de los resultados se refería a refranes pertenecientes a obras literarias italianas anteriores al siglo XX y que ya son en desuso, y otro 7% simplemente presentaba reflexiones sobre refranes y costumbres del pasado. Finalmente, un 13% estaba compuesto por repeticiones generadas por el programa de búsqueda.

Tras centrarnos en los refranes italianos, pudimos detectar 690 refranes. Algunos de estos ocurren más de una vez, lo que queda demostrado por el hecho de que en total los refranes italianos hallados son 1037. Esto no quiere decir que se repitan los mismos resultados, sino que se utilizan los mismos refranes en diferentes textos. Por esto, decidimos realizar una lista de todos los refranes que identificamos con las respectivas ocurrencias para lograr dos objetivos principales. El primero consistía en buscar los refranes encontrados en el propio corpus, para averiguar si estos se utilizaban también en otras ocasiones donde no aparecía la palabra "*proverbio*". El segundo objetivo estaba encaminado a detectar la frecuencia de uso de los refranes hallados, además de su contexto, para proponer una lista de frecuencia de uso de los refranes italianos que aún no habíamos encontrado.

También en este caso los resultados son muy llamativos. Por lo que respecta a los refranes detectados al principio utilizando la palabra "*proverbio*", se nota que una mayoría aplastante (el 77%) tiene una sola ocurrencia. El 11% tiene 2 ocurrencias, el 5% tiene 3, el 4% tiene 4, un 1% tiene 5 ocurrencias, otro 1% tiene 6 y finalmente otro 1% tiene entre 8 y 7 ocurrencias.

Tras obtener la lista de los 690 refranes, fue posible buscar los mismos refranes en el corpus para averiguar su funcionamiento en el contexto. Este estudio se limitó únicamente a los refranes de ocurrencias entre 8 y 2, puesto que, como ya se ha subrayado en este artículo, el objetivo principal ha sido encontrar una metodología de investigación que se pueda aplicar a otros tipos de estudios en este ámbito. De hecho, una vez determinadas las pautas adecuadas para llevar a cabo investigaciones parecidas sobre refranes, solo haría falta

aplicarlas a refranes encontrados en corpus distintos y también a refranes hallados en corpus de otros idiomas (con los ajustes necesarios para respetar las características de las diferentes lenguas).

Antes de presentar los resultados detallados de esta consulta, sería oportuno hacer hincapié en datos estadísticos sobre los mismos. Como sugiere Shapira (2000, pp.89-92) sobre el uso de refranes en lengua francesa, la mayoría de las veces los refranes no se introducen con un marcador de refrán, sino que se citan directamente. Los números que atañen a los refranes más frecuentes hablan muy claro: el refrán que más se utiliza en el corpus de referencia es "*L'unione fa la forza*" (trad. lit.: "La unión hace la fuerza"). De hecho ocurre en su forma estándar y en diversas versiones modificadas 579 veces, mientras que estaba presente en la lista de refranes obtenida utilizando la palabra "*proverbio*" tan solo 5 veces. El segundo refrán más frecuente es "*Il buongiorno si vede dal mattino*" (trad. lit.: "El buen día se ve desde la mañana"), que aparece en el corpus 447 veces en su forma estándar o en versiones modificadas, y que tenía 4 ocurrencias al buscarlos utilizando la palabra "*proverbio*". Lo mismo puede decirse sobre el refrán "*Non c'è fumo senza arrosto*" (trad. lit.: "No hay humo sin asado"), que tenía 4 ocurrencias y que, en cambio, se repite en total 382 veces.

Como ya se ha adelantado, la mayoría de los refranes se inserta directamente en el discurso sin ninguna fórmula introductoria. Sin embargo, otras observaciones subrayan su flexibilidad de uso. De hecho, una parte importante de la literatura paremiológica indica que generalmente las paremias se modifican. Pensemos por ejemplo en el estudio de Shapira (2000) que habla de *déproverbialisation* y *détournement* o, en español, desautomatización (García Yelo, 2012). Por consiguiente, nuestra atención se centró en las modificaciones de los refranes, para encontrar en el corpus no solo los refranes en su forma estándar[39], sino también sus diferentes variantes y sus versiones modificadas.

El corpus, con su clasificación y etiquetado por partes del discurso (o *POS tagging*), posibilita encontrar secuencias buscando tanto lemas específicos como partes del discurso, lo que permite hallar refranes modificados. Pero antes establecimos una serie de pautas para que la búsqueda de refranes modificados fuera igual y comparable para cualquier paremia de nuestra lista. Una vez más, hemos de subrayar que no existen estudios específicos sobre cómo se modifican los refranes y qué se puede modificar en los refranes.

Por esto, la primera regla que nos dimos fue tomar como referencia las modificaciones y variantes de los refranes que encontramos en el propio corpus gracias a la búsqueda de la palabra "*proverbio*". Por ejemplo, el refrán "*Chi ben comincia è a metà dell'opera*" (trad. lit.: "Quien bien empieza está a mitad de la obra") es la única variante aceptada por *Il grande dizionario dei proverbi italiani* (2006). No obstante, en el corpus analizado se notan otras variantes del mismo refrán, como "*Chi ben* ***inizia*** *è* ***già*** *a metà dell'opera*"; "*Chi ben comincia è* ***alla*** *metà dell'opera*"; "*Chi ben* ***principia*** *è* ***alla*** *metà dell'opera*"; "*Chi ben* ***incomincia*** *è a metà dell'opera"* (en negrita las palabras diferentes con respecto a la variante aceptada por el diccionario citado). Otro refrán que no siempre respeta las indicaciones del diccionario es "*Se la montagna non va da Maometto, Maometto va alla montagna*"[40] (trad. lit.: "Si la montaña no

[39] En este estudio, no se entiende por "forma estándar" la variante presentada por los diccionarios u otros recursos dedicados a los refranes, sino la variante encontrada en el corpus con la primera búsqueda, es decir buscando la palabra "*proverbio*".

[40] Esta es la variante aceptada por otro diccionario, el de Lapucci (2006), ya que este refrán no está presente en *Il grande dizionario dei proverbi italiani* (2006). Esto es muy llamativo, puesto que se trata de uno de los refranes más frecuentes en nuestro corpus, lo que indica su elevada frecuencia de uso en el italiano contemporáneo. Se podría pensar que los recursos que deberían dar cuenta de la presencia y uso de los refranes, además de su sentido, no siempre tiene la prioridad de describir del habla corriente sino más bien recopilar refranes, sin hacer distinción entre los realmente empleados y los anticuados.

va a Maoma, Maoma va a la montaña"). En nuestro corpus, los autores de los textos donde aparece dicho refrán se confunden (no se sabe si de manera voluntaria o involuntaria) y escriben: "*Se Maometto non va alla montagna la montagna va da Maometto*" (la coma entre las dos oraciones ya no está presente en los textos originales). En el corpus, se hallaron otros ejemplos de modificaciones: "*Se Maometto non va alla montagna,* ***è*** *la montagna che va da Maometto*" y "*Se la montagna non va da Maometto,* ***è*** *Maometto* ***ad andare*** *alla montagna*".

El análisis de las modificaciones ya presentes en un corpus lleva al investigador a intentar averiguar cómo se modifican los refranes y cuáles son los elementos que pueden variar. El problema surge para los refranes que solo aparecen una vez al buscarlos utilizando los marcadores de refranes. Por esa razón, intentamos detectar una metodología de búsqueda de refranes modificados basada en los siguientes principios:

En primer lugar, para cada uno de los refranes se insertó, en la interfaz de consulta del corpus, solo la primera parte y, posteriormente, solo la segunda. Para decidir cómo dividir los refranes nos inspiramos en los estudios sobre refranes bimembres (Honeck, 1997, pp. 51-57). Si los refranes estudiados no eran bimembres, intentamos insertar solo las dos primeras palabras y, si no obtuvimos resultados satisfactorios, fuimos introduciendo una palabra más de cada vez, hasta encontrar resultados significativos. Esto sirve en general para encontrar refranes citados solo parcialmente, lo que es muy frecuente y forma parte de lo que Shapira (2000) define *déproverbialisation*.

En segundo lugar, se decidió sustituir las palabras de los refranes por etiquetas morfosintácticas de esta manera: el objetivo, y a la vez el método para seguir, era encontrar la combinación entre palabras y etiquetas morfosintácticas que permitiera hallar más refranes que enunciados libres utilizando el menor número de lemas y el mayor número de etiquetas morfosintácticas. Se tome por ejemplo el refrán (2a).

> (2a) *Una mela al giorno toglie il medico di torno* (trad. lit.: "Una manzana cada día quita al médico de encima")

Si se utiliza una secuencia de búsqueda formada enteramente por etiquetas morfosintácticas como (2b), se encuentran refranes, pero también una gran cantidad de enunciados que no tienen nada que ver con (2a).

> (2b) [artículo] + [sustantivo] + [preposición compuesta] + [sustantivo] + [verbo] + [artículo] + [sustantivo] + [preposición simple] + [sustantivo]

Si se empieza a sustituir las etiquetas morfosintácticas con los lemas presentes en el refrán, se nota que los lemas necesarios son "*al*", "*giorno*", "*dì*", que podrían incluso constituir la estructura más fija e invariable del refrán. Asimismo, se observa que la combinación constituida por el menor número de palabras y el mayor número de etiquetas morfosintácticas, que produce resultados relevantes, es (2c).

> (2c) [artículo] + [sustantivo] + "al" + "giorno" + [verbo] + [artículo] + [sustantivo] + "di"

Como se ha notado, el hecho de encontrar refranes gracias a su estructura mínima invariable podría ser muy útil a la hora de producir nuevas reflexiones sobre qué pieza léxica se puede modificar en los refranes y cuál no se debe modificar.

5. CONCLUSIONES

Después de una introducción general sobre los asuntos más relevantes en el terreno de la lingüística de los refranes, aún poco estudiados desde el punto de vista lingüístico, la

investigación que nos está ocupando se propone realizar un estudio lingüístico sobre refranes, a través de la observación de datos reales sacados de itWaC, un corpus de lengua general italiana de dos mil millones de palabras.

El primer avance de este estudio consiste precisamente en llevar a cabo una investigación paremiológica enteramente basada en datos reales: los refranes han sido detectados de forma casi automática, buscando lo que hemos definido "marcadores de refranes". En realidad, el único elemento que desarrolla este papel y es capaz de producir la cantidad más elevada de refranes, junto con el porcentaje menor de resultados no relevantes, es la palabra italiana "*proverbio*" (refrán). De todos modos, sería importante estudiar todas las expresiones que suelen acompañar a los refranes, para dar pasos adelante en la identificación automática de paremias.

Por otra parte, gracias al etiquetado morfosintáctico del corpus, no solo es posible elaborar una lista de frecuencia de refranes, sino también buscar refranes en su variante estándar y en sus versiones modificadas.

Los datos demuestran que los refranes, si bien pueden modificarse, se cimientan en estructuras básicas y fijas, que permiten identificar la forma estándar de las paremias modificadas. Una investigación de este tipo lleva asimismo a detectar estas estructuras estables, desde un punto de vista tanto léxico como sintáctico. La etapa siguiente, pues, será buscar estructuras fijas desde la óptica semántica, lo que permitiría categorizar los refranes según su sentido "idiomático" y no "literal". Todo esto posibilitaría la búsqueda y recuperación de refranes categorizados conceptualmente en distintos idiomas, superando los límites de los recursos actuales en el ámbito paremiológico. En otros términos, se podría pasar desde la perspectiva actual de tipo semasiológico hasta una perspectiva futura de carácter onomasiológico.

Aún queda mucho por investigar, pero estamos seguros de que el camino emprendido es capaz de llevar a resultados novedosos no solo en el marco de los refranes sino también en otras disciplinas, como la fraseología, la lexicografía y la terminología.

Bibliografía

ANSCOMBRE, J.-C., 1997. Reflexiones críticas sobre la naturaleza y el funcionamiento de las paremias. *Paremia*, 6, pp.43-54.

ANSCOMBRE, J.-C., 2005. Les proverbes : un figement du deuxième type ? *Linx*, 53, [online] Available at: http://linx.revues.org/255 [Accessed 11 October 2012].

BARONI, M., BERNARDINI, S., FERRARESI, A. AND ZANCHETTA, E., 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3), [online] Available at: http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf [Accessed 9 October 2014].

BARSANTI VIGO, M.J., 2006. Problemática en torno al refrán y otras categorías paremiales. In: M. Alonso Ramos, ed. 2006. *Diccionarios y fraseología. Anexos de Revista de Lexicografía*, 3, pp.197-206.

BIDAUD, F., 2002. *Structures figées de la conversation. Analyse contrastive français-italien*. Bern: Lang.

BOGGIONE, V. AND MASSOBRIO, L., 2007. *Dizionario dei proverbi*. Torino: UTET.

CAMPOS, J. G. AND BARELLA, A., 1995. *Diccionario de refranes*. Madrid: Espasa Calpe.

CANELLADA, M. J. AND PALLARES, B., 2001. *Refranero español. Refranes, clasificación, significación y uso*. Madrid: Castalia.

CASADEI, F., 1996. *Metafore ed espressioni idiomatiche: uno studio semantico sull'italiano*. Roma: Bulzoni.

CORREAS, G., 1992. *Vocabulario de refranes y frases proverbiales y otras fórmulas comunes de la lengua castellana en que van todos los impresos antes y otra gran copa que juntó el Maestro Gonzalo Correas*. Madrid: Visor.

CRISTILLI, C., 1989. Il proverbio come esempio di testualità popolare. In: C. Vallini, ed. 1989. *La pratica e la grammatica. Viaggio nella linguistica del proverbio.* Napoli: Istituto Universitario Orientale, Dipartimento di studi letterari e linguistici dell'Occidente. pp.177-206.

DEIGNAN, A., 2009. Searching for Metaphorical Patterns in Corpora. In: P. Baker, ed. 2009. *Contemporay Corpus Linguistics*. London, New York: Continuum. pp.9-31.

DEPECKER, L., 2011. Comment aborder le concept d'un point de vue linguistique ? In : J.-J. Briu, 2011. *Terminologie (I) : analyser des termes et des concepts.* Bern: Lang. pp.17-32.

DUNETON, C. AND CLAVAL, S., 1990. *Le bouquet des expressions imagées : encyclopédie thématique des locutions figurées de la langue française*. Paris: Seuil.

GARCÍA YELO, M., 2012. El proceso de desautomatización de paremias españolas en las redes sociales. *Unidades fraseológicas y TIC*. [online] Serie «Monografías», n°2, pp.111-124. Available at: http://cvc.cervantes.es/lengua/biblioteca_fraseologica/n2_gonzalez/_default.htm [Accessed 20 August 2013].

GONZÁLEZ REY, M.I. ed., 2012. *Unidades fraseológicas y TIC*. [online] Madrid: Instituto Cervantes, Biblioteca fraseológica y paremiológica, Serie «Monografías», n°2. Available at: http://cvc.cervantes.es/lengua/biblioteca_fraseologica/n2_gonzalez/default.htm [Accessed 20 August 2013].

GROSS, G., 1996. *Les expressions figées en français.* Gap: Ophrys.

GUAZZOTTI, P. AND ODDERA, M.F., 2010. *Il grande dizionario dei proverbi italiani.* Bologna: Zanichelli.

HERBST, T., FAULHABER, S. AND UHRIG, P. eds., 2011. *The Phraseological View of Language. A Tribute to John Sinclair*. Berlin: De Gruyter Mouton.

HONECK, R.P., 1997. *A proverb in mind: the cognitive science of proverbial wit and wisdom*. Mahwah (N. J.): Lawrence Erlbaum associates.

JUNCEDA, L., 1991. *Del dicho al hecho*. Barcelona: Obelisco.

KATZ, J.J., 1973. Compositionality, idiomaticity and lexical substitution. In: S. R. Anderson, P. Kiparsky, eds. 1973. *A Festschrift for Morris Halle*. New York: Holt, Rinehart and Winston. pp.357-376.

KLEIBER, G., 1990. *La sémantique du prototype. Catégories et sens lexical*. Paris: PUF.

KLEIBER, G., 1999. Les proverbes : des dénominations d'un type « très très spécial ». *Langue française*, 123, pp.52-69.

KLEIBER, G., 2000. Sur le sens des proverbes. *Langages*, 139, pp.39-58.

LAKOFF, G. AND JOHNSON, M., 1980. *Metaphors We Live By*. Chicago: University of Chicago.

LAPUCCI, C., 1993. *Dizionario dei modi di dire della lingua italiana*. Milano: Garzanti.

NORRICK, N.R., 1985. *How proverbs mean: semantic studies in English proverbs*. Berlin: Mouton.

PERRIN, L., 2000. Remarques sur la dimension générique et sur la dimension dénominative des proverbes. *Langages*, 139, pp.69-80.

PHILIP, G., 2011. *Colouring Meaning. Collocation and connotation in figurative language.* Amsterdam: John Benjamins.

PITTANO, G., 1996. *Frase fatta capo ha: dizionario dei modi di dire, proverbi e locuzioni*. Bologna: Zanichelli.

QUARTU, B.M., 2001. *Dizionario dei modi di dire della lingua italiana: 10.000 modi di dire ed estensioni figurate in ordine alfabetico per lemmi portanti e campi di significato*. Milano: Rizzoli.

QUIROGA, P., 2006. *Fraseología italo-española. Aspectos de lingüística aplicada y contrastiva*. Granada: Granada lingüística.

SEVILLA MUÑOZ, J. AND ZURDO RUIZ-AYÚCAR, M.I.T. eds., 2009. *Refranero multilingüe*. Madrid. Instituto Cervantes (Centro Virtual Cervantes). [online] Available at: **¡Error! Referencia de hipervínculo no válida.** [Accessed 20 March 2015].

SEVILLA MÚÑOZ, J., 2000. Les proverbes et phrases proverbiales français, et leurs équivalences en espagnol. *Langages*, 139, pp.98-109.

SHAPIRA, C., 2000. Proverbe, proverbialisation et déproverbialisation. *Langages*, 139, pp.81-97.

SIMPSON, J., 1992. *The Concise Oxford Dictionary of Proverbs*. Oxford: Oxford University Press.

SINCLAIR, J., 2003. *Reading Concordances. An Introduction*. London: Pearson Longman.

TOGNINI-BONELLI, E., 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamin.

TURRINI, G., ALBERTI, C., SANTULLO M.L. AND ZANCHI, G., 1995. *Capire l'antifona*. Bologna: Zanichelli.

VISETTI, Y.M. AND CADIOT, P., 2006. *Motifs et proverbes. Essai de sémantique proverbiale.* Paris: Presses Universitaires de France.

XATARA, C.M., 2009. *Dictionnaire d'expressions idiomatiques: Français – Portugais –Français.* [online] Available at: http://www.cnrtl.fr/dictionnaires/expressions_idiomatiques/ [Accessed 10 March 2015].

CONTRASTIVE ANALYSIS OF PHRASEOLOGICAL UNITS WITH SPECIFIC ANIMAL CONSTITUENTS IN ENGLISH, SPANISH AND GERMAN

Marta Morer Murcia
marta.morer@um.es

Abstract

In the following paper we present a contrastive analysis in which three different languages (English, Spanish, and German) are contrasted in order to examine the degree of equivalence that their phraseological units (PUs) with specific animal constituents present. More precisely, the analysis uses this comparison to classify the languages in different groups based on semantic connotations and several idiosyncrasies of the PUs following Corpas Pastor's (2003) degrees of equivalence. The PUs taken into consideration to build the corpus of the current paper refer to those idioms and routine formulas that have the lexical element of specific animals. The animal constituents that we have selected are *cat, dog, horse* and *monkey*. After the analysis of these specific English PUs, the paper shows the type of equivalence that they have in the other two languages and the classification of these units depending on their degrees of equivalence.

1. INTRODUCTION

Due to its importance in the study of language as well as the difficulties it frequently presents in translation and second language teaching, phraseology plays a significant role in the field of linguistics. Some PUs can present special elements which, in some cases, make these tasks easy and, in others, complex; this is the case with PUs referring to specific animals. As a result, this topic requires further study. Furthermore, we must consider the fact that there are several contrastive studies which have analysed the PUs of two languages and that frequently the languages compared are English and another language, German and another language, or Spanish and another language. There are not many phraseological studies in which several languages are compared in order to look for the equivalences of the PUs between them. That is the aim of the present study: to contrast three different languages with the purpose of collecting data about their relative degrees of equivalence of selected PUs.

2. CLASSIFICATION OF PUs

Corpas Pastor (1997) presents criteria in order to distinguish different types of PUs and their further classification. She establishes this proposal, due to the fact that none of the criteria existing before

> serves on its own to vertebrate a global classification of the Spanish phraseological system. Because of this reason, we propose to combine the statement's criteria- and, consequently, the speech's act- with the one of fixing in the norm, the system or the speech.
> Both criteria provide the base for establishing a first level of classification of the PUs in three different spheres.
> (1997: 50)

With this proposal, the units are divided into two groups: First, we find the group of units that do not constitute complete statements alone, and make reference to syntagmas that are combined with other linguistic symbols in order to establish acts of communication. The second group consists of those that constitute statements themselves; these units are called idiomatic expressions.

Gläser (1986) organises propositional PUs in two sub-groups: partial and complete propositional. The author presents several classifications in different sub-divisions of the propositional groups, but this part is only relevant for the study and the explanation of *Routineformeln* (routine formulas).

3. FEATURES OF PHRASEOLOGICAL UNITS

Bearing the information above in mind, it is important to emphasize that the present study only focuses on the analysis of a certain subset of units. As a result, the following lines explain the main PUs, as idioms, required in order to facilitate the understanding and further analysis of this study.

> As Corpas Pastor (1997: 88) stated, idioms are:
> "unidades fraseológicas del sistema de la lengua con los siguientes rasgos distintivos: fijación interna, unidad de significado y fijación interna pasemática. Estas unidades no constituyen enunciados completos y, generalmente, funcionan como elementos oracionales."
>
> Another linguist, Ruiz Gurillo (2001:31) defines them as:
> "[...]sintagmas fijos que en ciertos casos presentan idiomaticidad"

According to one of the several meanings that the Oxford English Dictionary gives to the noun *idiom,* the one that is linked to the studies made about English idiomaticity is the following:

> "A form of expression, grammatical construction, phrase etc, peculiar to a language; a peculiarity of phraseology approved by the usage of a language and often having a significance other than its grammatical or logical one."

The idiomaticity is considered one of the most important characteristics that define a PU, as it differentiates idioms from other statements. Corpas Pastor (1996), states that the term

idiomaticity refers to that semantic lexicalization or specialization in its highest level. In the same way, Corpas Pastor adds that "Phraseological Units can have two types of denotative meaning: denotative literal meaning and denotative figurative meaning, namely, idiomatic" (1996:27). It is important to underline that not all PUs are idiomatic, which makes it a potential characteristic.

4. ONOMASIOLOGICAL APPROACHES

Due to the peculiarities that some PUs present, many authors have focused their studies on onomasiological approaches. These studies are based on the specific lexical elements that constitute the PUs, as many of these PUs can be grouped depending on their lexical constituents. There is a wide variety of onomasiological approaches taken, such as studies focusing on constituents which designate color, parts of the human body or names of people.

> According to Fiedler (1999), these specific constituents found in PUs "verified supposed universal phraseological properties", such as those ones related with the parts of the human body. PUs under this classification are called *somatisms,* as the author explains. Moreover, regarding PUs with animal constituents, it is stated that these specific constituents are conditioned by the culture depending on each country and language.

5 . AIMS

The target of the current paper is to find the degree of equivalence among the idioms that have the specific animal constituent of *cat, dog, horse* and *monkey* in English, Spanish and German. In addition to this general aim, the specific aims of the work are to ascertain whether there is a higher degree of equivalence in the PUs between English and German (as they are Germanic languages and they are closely related) or whether this is more prevalent between English and Spanish. After collecting the corpus of idioms for its further analysis, we classify them following Corpas' classification, depending on the degree of equivalence that they share between the languages. Therefore, we present:

1. Degrees of equivalence in the three languages English, Spanish and German.
2. Comparison of degrees of equivalence between English and Spanish, and between English and German.

6. METHODOLOGY

In accordance with the OED, a list comprising the idioms in English has been created. The criteria that are used in order to select them were: first, that they are idioms with the constituent of an animal; second, only those with the animal constituents *cat, dog, horse* and *monkey* were chosen, as cited above. Plurals of these constituents were not taken into consideration when collecting the data. As a result, only the PUs formed by singular forms of *cat, dog, horse* or *monkey* were selected for analysis.

We have worked with them taking into account the animal used and we have grouped them in three different columns, one for each language, although they do not always keep the same equivalences between languages. For the abovementioned reasons, Corpas Pastor (2003)'s classification referring to idioms' different degrees of equivalence is used. The contrastive analysis follows these types of equivalences: full, partial or zero.

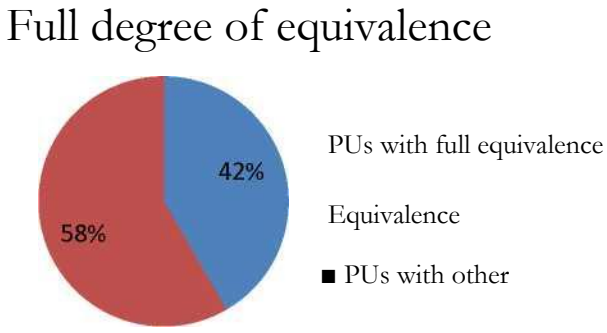
7. RESULTS

After the study of the PUs in the three different languages and their further classification, different sections are presented below. First, each of the degrees of equivalence is separated in a different section, distinguishing between full, partial and zero degree of equivalence of the PUs. Furthermore, within these sections, the PUs are classified in other subsections depending on the languages that are contrasted. Therefore, we find also three groups: the comparison of PUs in English, Spanish and German; English and Spanish; and English and German. The contrastive analysis is explained and divided into three different sections depending on the degree of equivalence that PUs present between the languages. Firstly, the contrasts between the three languages are explained in each of the sections. Then, the results of the German and Spanish PUs are separately exposed in relation to the equivalences that each share with the English PUs.

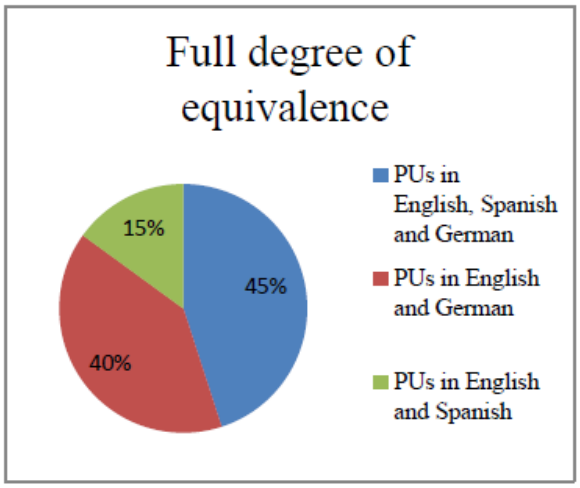
7.1 Full degree of equivalence

Concerning the important role that the specific animal constituents play in the classification as well as in the analysis of the corpus, it is important to underline that the classification of full equivalent PUs have been precise; only those PUs that had the same animal constituent on each language are taken as completely equivalent.

A number of selected samples of those PUs that present complete equivalence taken from the overall corpus are shown in this part. Before focusing on the specifics examples of PUs found with full equivalence, it is necessary to observe in detail the percentages exposed in the Graphic 1. It shows the amount of full equivalences out of the total quantity of PUs selected;
20 from 48 total PUs. We can assume that, although some PUs share a complete equivalence, the majority of them are more likely to possess partial or zero equivalence when translating them to other languages.



Graphic 1. Full degree of equivalence a



Graphic 2. Full degree of equivalence b

In addition to the global percentages, a further graph illustrates the concrete percentages of the three languages in which the PUs are contrasted. In Graph 2 we can clearly observe that PUs in the three languages are more frequently completely equivalent than in English and Spanish or in English and German. Nevertheless, Graph 2 also shows that English and German have a remarkable number of equivalences, 8 out of 20, and that only the 15% of these English PUs remain fully equivalent regarding Spanish.

Full degree of equivalence in the three languages English, Spanish and German Some typical examples in which the three languages have the same idioms as well as the same meaning are found in Table 1. Only those PUs presenting a common animal in all languages are considered. For instance, the English idiom *A cat has nine lives* presents full equivalence to the Spanish and German idioms, as in both languages the constituent *cat* appears. Furthermore, although the meaning remains, when translating the English idiom into Spanish and German literally, we observe one difference. As shown in Table 1, while in English number nine is used to express the lives of a cat, in the other two languages the number in question is seven: *Siete vidas tiene un gato* and *Eine Katze hat sieben Leben.*

Using English idioms as the starting point, other features are found among these PU. Full equivalences are found in the two other contrasted languages, including meaning, form and lexicon such as a dog's life, vida de perros or *ein Hundeleben.*

English	Spanish	German
A cat has nine lives	Siete vidas tiene un gato	Eine Katze hat sieben Leben
A dog's life	Vida de perros	Ein Hundeleben

Table 1. Full equivalence in English, Spanish and German

Full degree of equivalence between English and German

Through the contrasted analysis, English and German show a high number of full equivalences regarding PUs; 40% of the total amount of full equivalences. In order to see the similitude between both languages, Table 2 contains two samples from the corpus. The English idiom You can lead a horse to water, but you can't make him drink keeps the same meaning, lexicon and syntax as the German idiom, and presents a literal translation. In the second example, we find the same characteristics but the constituent *dog* in the PUs. Due to their common linguistic origins both languages share high percentages of full equivalences, and therefore do not present any kind of complexity when contrasting their PUs.

English	German
You can lead a horse to water, but you can't make him drink	Man kann ein Pferd zum Brunnen führen, aber trinken muss es selbst
Love me, love my dog	Wer mich liebt, der liebt auch mein Hund

Table 2. Full equivalence in English and German

Full degree of equivalence between English and Spanish

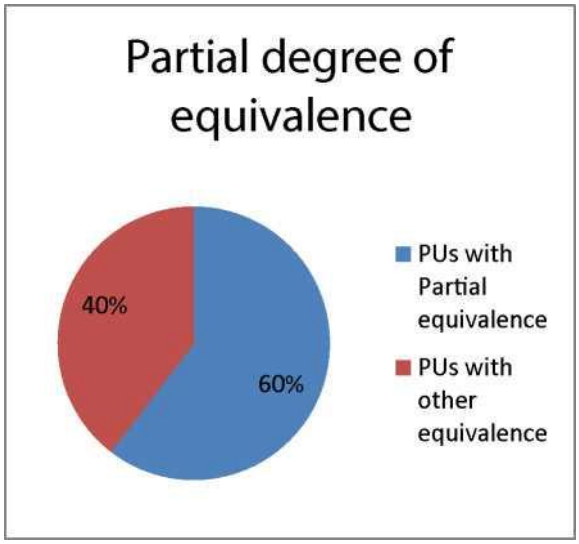
The comparison between English and Spanish concerning full equivalences show opposite results than the ones obtained between English and German. Only three from the total number of English PUs analysed are fully equivalent in Spanish; they represent the 15% of the total amount of full equivalences. It is interesting the fact that these three PUs contain all the same animal constituent: *dog*. It may be because it is a common and typical animal, or because both languages use considerably this constituent in the communicative frame. Table 3 shows the PU You can't teach an old dog new tricks, which in Spanish is *Perro viejo todo sonpulgas*. In this example we do not find a literal translation but still both languages contain the constituent *dog* and the metaphorical meaning remains the same; thus, the PUs present full equivalence. In the second example, both PUs remain fully equivalent in all aspects, the literal translation is completely similar as well as the meaning: *Work like a dog* and *Trabajar como un perro*.

English	Spanish
Work like a dog	Trabajar como un perro
You can't teach an old dog new tricks	Perro viejo todo son pulgas

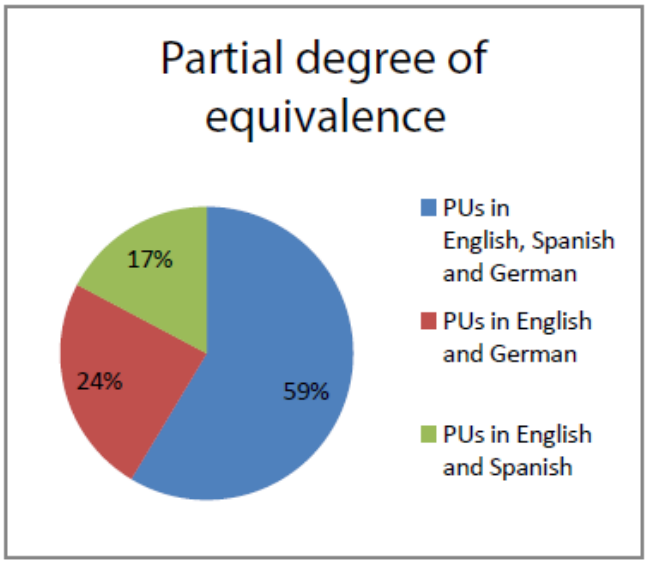
Table 3. Full equivalence in English and Spanish

7.2 Partial degree of equivalence

Some of the PUs have differences in the animal constituents of the PUs, as different animals in each language are used to express the same metaphorical meaning of the idiom. However, it is crucial to remark that not all the PUs in this section have the same features, since some PUs do not have any animal constituent yet keep the metaphorical meaning between the compared languages. The number of PUs found presenting partial equivalence are 29 out of the total amount of the corpus; 48 PUs. The number of PUs with partial equivalence varies depending on the languages that are contrasted. Thus, 17 PUs share partial equivalence in English, Spanish and German, 7 in English and German and 5 in English and Spanish. Graph 4 shows the biggest amount of partial equivalence, which appear when comparing the three have 24% of partial equivalence and the lowest number is represented between English and Spanish, only 17%.



Graphic 3. Partial degree of equivalence a languages: 59%; English and German PUs



Graphic 4. Partial degree of equivalence b

After the global results of this section, three separate sections are presented in order to explain several examples from the partial equivalences between the three languages, English and German, and English and Spanish.

Partial degree of equivalence in the three languages English, Spanish and German

According to the analysis of the corpus, different types of partial equivalences are found, and they present several characteristics which it is important to mention. Some of them do not present the specific animal constituents but still maintain the same meaning based on their communicative values in the other languages. In our research, we find PUs such as the ones provided in the Table 4, *I'm so hungry I could eat a horse,* where the PUs keep the same meaning across the three languages, but the animal constituent changes from one language to another. While in English appears the constituent *horse,* the Spanish equivalence is *tengo tanta hambre que me comeria una vaca,* and has the constituent *vaca;* in German the animal constituent is *Schwein: Ich habe so einen Hunger, Ich könnte ein halbes Schwein auf Toast essen* is used. Three different animals but still the same functional value. As it was previously mentioned, partial degree of equivalence concerning idioms is characterized by the similitude that languages share, not in the literal meaning but in the metaphorical one.

English	Spanish	German
I'm so hungry I could eat a horse	Tengo tanta hambre que me comería una vaca	Ich habe so einen Hunger, Ich könnte ein halbes Schwein auf Toast essen

Table 4. Partial equivalence in English, Spanish and German

Partial degree of equivalence between English and German

Partial equivalence is found moderately between English and German PUs, making up 24% of the partial equivalences of the analysis. Although the majority of idioms share the same meaning, the animal constituent varies. In the examples shown in Table 5 we see that while in English the constituent *dog* in *Work like a dog* appears, in German there is no specific animal in *wie ein Tier schuften,* the metaphorical meaning remains but the constituent changes.

English	German
Work like a dog	Wie ein Irrer arbeiten/ wie ein Tier schuften

Table 5. Partial equivalence in English and German

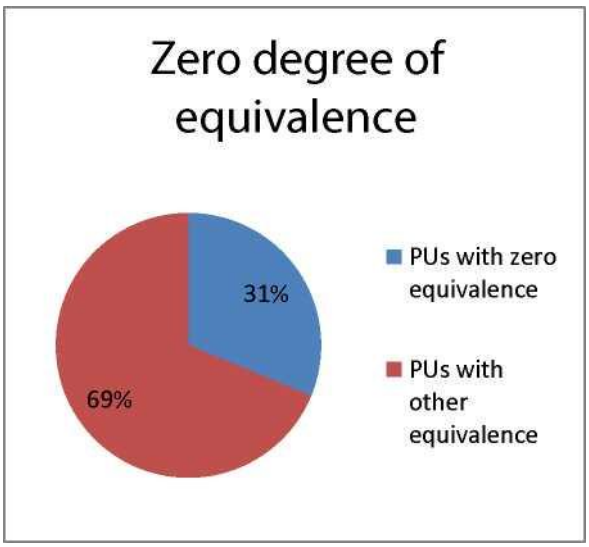
Partial degree of equivalence between English and Spanish

In comparison with English and German, partial degrees of equivalence between English and Spanish do not have the same metaphorical translation or present an animal constituent in both PUs. Their PUs show a different lexicon, although in each language they keep the same meaning. When we translate literally the dog's bollocks into Spanish, we get a nonsense sentence, and vice versa: *A buenas horas, mangas verdes* does not have any metaphorical nor literal meaning in English but still, both idioms in each language share a partial equivalence as their literal meaning coincide and they are applied in the same type of circumstances.

English	Spanish
The dog's bollocks	Ser la leche
Close the stable after the horse has bolted	A buenas horas, mangas verdes

Table 6. Partial equivalence in English and Spanish

7.3 Zero degree of equivalence

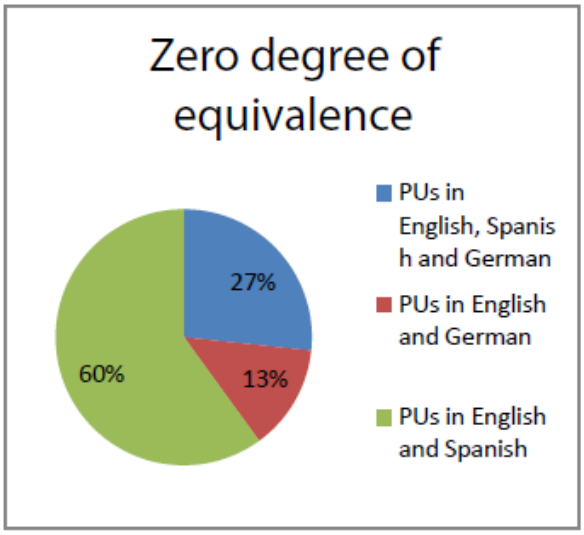


Graphic 5. Zero degree of equivalence

According to Corpas Pastor (2003), zero equivalence includes the idioms of one language that do not present a translation equivalent in another language. Therefore, it is hard to make assumptions of the possible translation that the English expressions could have in Spanish and German; as a result, we leave blank spaces as they are an issue for further study. These PUs do not present phraseological equivalences and their possible translation would be the equivalent of a paraphrase.

This type of equivalence is not very likely to occur among the PUs selected. Only 31% of the total number have zero equivalence in the other languages, and it is unusual that English PUs doe not find any kind of translation in other language in this regard.

Additionally, Graph 6 shows the percentages of PUs presenting zero degree of equivalence in English, Spanish and German, where 27% of the cases were found. Also, it presents PUs in English and German, with a small number of results (13%). Finally, the thir[d] percentage confirms that more than half of the PUs compared between English and Spanish (60%) present zero equivalence.



Zero degree of equivalence in the three languages English, German and Spanish PUs such as *get on your high horse* or *make a monkey out of somebody* describe certain attitudes people have. Furthermore, in relation with the Spanish language, some free interpretation of these PUs could be proposed in form of paraphrases due to these particular PUs not existing as such, for instance, *sentir superioridad* or *hacer que alguien parezca estúpido.* In certain situations paraphrases are the only way out to translate PUs among languages, in these occasions zero equivalence exists. Finally, German interpretations of the possible paraphrases regarding the mentioned degree of equivalence are presented too. These are *Überlegenheitsgefühl* or *es lässt jemanden dumm darstehen.* It should be noted that, all the possible interpretations of the English idioms into Spanish and German are not considered PUs, but paraphrases.

English	Spanish	German
Get on your high horse	Sentir superioridad	Überlegenheitsgefühl
Make a monkey out of somebody	Hacer parecer a alguien estúpido	Es last jemanden in dumm darstehen

Table 7. Zero equivalence in English, Spanish and German

Zero degree of equivalence between English and German

The equivalences found between English and German, concerning zero degree make up, as stated above, 13% of the total PUs. In Table 8 we give a free interpretation in German as there is no equivalent for the English idiom. For instance, *Wundervolle Person* as a paraphrase for the English idiom The dog's bollocks.

English	German
The dog's bollocks	Wundervolle Person

Table 8. Zero equivalence in English and German

Zero degree of equivalence between English and Spanish

At this point is important to point out that there is a high degree of zero equivalence between English and Spanish PUs, representing 60% out of the total English PUs with zero equivalence analysed. Thus, from the samples in Table 9, we give some possible paraphrases. For example, the English PU *Love me, love my dog* could be paraphrased in Spanish as *querer a alguien aceptando cualquier consecuencia.*

English	Spanish
Love me, love my dog	Querer a alguien aceptando cualquier consecuencia

Table 9. Zero equivalence in English and Spanish

6. CONCLUSION

The corpus created for the study of the PUs provides useful data which, narrowed by the semantic field of animals and considering four specific animal constituents, functions as an introductory line for further study. The study of these three different languages gives us a different perspective about the complexity of their equivalent translation. Depending on the languages compared, the number of equivalences in PUs varies, thus, several conclusions can be drawn.

Firstly, the study has demonstrated that the complete degree of equivalence between the three languages compared is medium; English, Spanish and German are not likely to be fully equivalent when dealing with PUs with specific animal constituents. Secondly, it is more usual to find a partial equivalence between the three languages, as it is unlikely that the PUs of three different languages coincide when the filter is narrowed only by certain animals. Almost all the PUs in English can be found in other languages; even though the translation is not exactly the same, the meaning remains and there is and idiomatic correspondence. Lastly, as the results demonstrate, zero equivalence between the three languages is low and in those particular cases the PUs can be interpreted in the other two languages through paraphrase. Through these conclusions the principal aim of the study has been achieved.

In addition, another specific aim is presented: to prove whether the equivalences of the PUs with specific animal constituents are higher between English and German or, conversely, between English and Spanish. Differing conclusions regarding the contrast between English and Spanish PUs can be drawn. Full equivalences between both languages are extremely low. Also, the instances of partial equivalences remain low between both languages, the 17%. Finally, English and Spanish have the highest amount of zero equivalences, 60% and this fact implies that these two languages are quite different and that equivalent translation between them remains hard. This fact leads us to suggest that, in the study of English and Spanish PUs with specific animal constituents, paraphrasing is simpler than literal translation.

From the results of the contrasts between English and German, the following conclusions can be drawn: the degree of full equivalence between both languages is very high, 40%. Moreover, we have observed that, when the languages do not have total equivalence, they share a high degree of partial equivalence, as almost all the English PUs have a partial similitude. The degree of zero equivalence among the languages is low almost to the point of insignificance.

As a result, the study's specific aim has also been achieved through the contrasted analysis. The degree of full equivalences is higher between English and German: thus, they are fully equivalent languages in this regard and the fact that they are both languages with common origin may seem favorable for the high degree of full equivalences between them. Finally, our results point out that the degree of zero equivalence is higher between Spanish and English; therefore, both languages are not equivalent and have not common features regarding animal PUs, may be due to cultural or historical reasons between both languages.

PUs are an essential element in language, and a good knowledge of them seems to be essential for solid understanding and effective translation of the languages. Consequently, theoretical and practical study of PUs should form an essential part of language learning. Finally, the current contrastive study provides certain knowledge about those specific PUs containing concrete animal constituents used in each language that belong to the lexicon of each speaker and culture; thus, it is a useful resource to draw on when determining the methodology of further contrastive studies of the languages in question

References

ALEXANDER, R. 1984. *Fixed expressions in English: reference books and the teacher.* ELT Journal, Vol. 38/2. Oxford: Oxford University Press.

BURGER, H. 1998. *Phraseologie. Eine Einführung am Beispiel des Deutsches.* Berlin: Erichschmidt.

CORPAS PASTOR, G. 1997. *Manual de fraseología española.* Madrid: Editorial Gredos.

CORPAS PASTOR, G. 2003. *Diez años de investigación en fraseología, análisis sintáctico- semánticos, contrastivosy traductológicos*. Madrid: Iberoamericana

DICCIONARIO DE MARÍA MOLINER. 1 998. 2nd Ed. Madrid: Editiorial Gredos

DICCIONARIO PRÁCTICO DE LOCUCIONES Y FRASES HECHAS. 1998. Madrid: Editorial Everest.

DROSDOWSKI, B. and SCHOLZE-STUBENRECHT, W. 1992. *DUDEN. Redewendungen und sprichwörtliche Redensarten. Band 11*. Mannheim: Dudenverlag.

FERNANDO, C. and FLAVEL, R. 1981. *On Idiom. Critical views and perspectives.* Vol.5 Exeter: University of Exeter

GELBRECHT, A. 2011. *Phraseology in Intercultural Communication. A contrastive approach towards German and English phraseological units of fire and water.* München: GRIN.

GLÄSER, R. 1986. *Phraseologie der englischen Sprache.* Tübingen: Niemeyer. GRANGER, S. and MEUNIER, F. 2008. *Phraseology. An interdisciplinary perspective.* Amsterdam: John Benjamins Publishing Company

HERBST, T., FAULHABER, S. and UHRIG, P. 2011. *The phraseological view of language. A tribute to John Sinchair.* Berlin: De Gruyter Mouton.

KATZ, J. and POSTAL, P. 1963. *Semantic interpretation of idioms and sentences containing them. Quarterly Progress Report.* Massachusetts: MIT Press.

MAKKAI, A. 1972. *Idioms structure in English.* The Hague: Mouton.

Nash, R. 1973. *Reading in Spanish-English Contrastive Linguistics.* Puerto Rico: Inter American University Press.

OXFORD ENGLISH DICTIONARY ONLINE. http://www.oxforddictionaries.com, Accessed 15 th January 2015.

ZULUAGA, A. 1980. *Introducción al estudio de las expresiones fijas.* Studia Romantica et lingüística; 10. Frankfurt am Main: Peter Lang

Appendix
Section 1. Classification of the PUs with full degree of equivalence

Table A. Full degree of equivalence in English, Spanish and German

English Spanish German

Cat:

Play cat and mouse	Jugar al ratón y el gato	Mit jdm Katz und Maus spielen
While the cat's away, the mouse will play	Cuando duerme el gato los ratones bailan	Ist die Katze aus dem Haus, regt sich die Maus.
Curiosity killed the cat	La curiosidad mato al gato	Neugierde hat die Katze zu Tode gebracht
A cat has nine lives	Siete vidas tiene un gato	Eine Katze hat sieben Leben
Has the cat got your tongue?	¿Te ha comido la lengua el gato?	Hat *es* dir die Sprache verschlagen

Dog:

A dog's life	Vida de perros	ein Hundeleben
Fight like cat and dog	Llevarse como el perro y el gato	Sie sind wie Hund und Katze
Why keep a dog and bark yourself?	Perro ladrador poco mordedor	Wenn du einen Hund hältst, belle nicht selber!

Horse:

Never look a gift horse in the mouth	A caballo regalado no le mires el dentado	Einem geschenkten Gaul schaut man nicht ins Maul

Table B. Full degree of equivalence in English and German

Cat:

Let the cat out of the bag	Die Katze aus dem Sack lassen

Dog:

Go to the dogs	vor die Hunde gehen
Let sleeping dogs lie	Keine schlafenden Hunde wecken
Love me, love my dog	Wer mich liebt, der liebt auch meinen Hund
Give a dog a bad name	Schnell einen Stock, wenn man einen Hund schlagen will

Horse:

You can lead a horse to water, but you can't make him drink	Man kann ein Pferd zum Brunnen führen, aber trinken muss es selbst
Put the cart before the horse	Das Pferd beim Schwanz aufzäumen
Back to the wrong horse	Aus falsche Pferd setzen

Table C. Full degree of equivalence in English and Spanish

Dog:

Work like a dog	Trabajar como un perro
You can't teach an old dog new tricks	Perro Viejo todo son pulgas
Sick as a dog	Estar malo como un perro

Section 2. Classification of the PUs with partial degree of equivalence

Table D. Partial degree of equivalence in English, Spanish and German Cat:

Not enough room to swing the cat	No cabe ni un alfiler	In beengten Pl atzverhältni ssen
Its raining cats and dogs	Esta lloviendo a cantaros	Es regnet in Strömen
Be the cat's whiskers	Ser el ombligo del mundo	Affengeil sein
Like a cat on a hot tin roof	Estar como un flan	Wie auf glühenden Kohlen sitzen
Look like something the cat brought in	Mira quien se ha dejado caer por aquí	Aussehen wie unter die Räuber gefallen/ aussehen wie eine Vogelscheuche
To grin like a Chesire cat	Sonreir de oreja a oreja	Grinsen wie ein Honigkuchenpferd

Dog:

A dog in the manger	Ser un aguafiestas	Er ist ein Spielverderber
It was a dog's breakfast	Estar patas arriba	Es war für die Katz.
Every dog has its day	Tener su momento de Gloria	Ein blindes Huhn findet auch mal ein Korn
Put on the dog	Vestirse de punta en blanco	Sich fein herausputzen
Dressed up like a dog's dinner	Ir vestido como un mono de feria	Herausgeputzt sein wie ein Zirkuspferd

Horse:

Straight to horse's mouth	De primera mano	Aus erster Hand
Horses for courses	Siempre hay un roto para un descosido	Für jede Gelegenheit das Passende/ für jeden Topf den richtigen Deckel
Flog a dead horse	Machacar en hierro frio	Tauben Ohren predigen
I'm so hungry I could eat a horse	Tengo tanga hambre que me comería una vaca	Ich habe so einen Hunger, Ich könnte ein halbes Schwein auf Toast essen

Monkey:

Brass monkey weather	Hace un frio que pela	Saukälte/ Es ist arschkalt
Monkey around with something	Marear la perdiz con algo	Als ob man einen Gaul auf dem Rücken schleppt.

Table E. Partial degree of equivalence in English and German

Cat:

Put he cat among the pigeons — Den Bock zum Gärtner machen

Dog:

See a man about a dog	Ich muss noch auf die Toilette
Work like a dog	Wie ein Irrer arbeiten/ wie ein Tier schuften
Sick as a dog	Mir geht's richtig schlecht
You can't teach an old dog new tricks	Was Hänschen nicht lernt, lernt Hans nimmermehr

Monkey:

Throw a monkey wrench in the Works	Jemanden in die Parade fahren/ Sand ins Getriebe streuen
Be cold enough to freeze the balls off a brass monkey	Mann kann sich die Beine abfrieren von Kälte

Table F. Partial degree of equivalence in English and Spanish

Cat:

Let the cat out of the bag — Irse de la lengua/ descubrir el pastel

Dog:

The dog's bollocks	Ser la leche
Go to the dogs	Echarse a perder
Let sleeping the dogs lie	Mejor no meter el dedo en la yaga

Horse:

Close the stable after the horse has bolted — Boca cerrada no entran moscas

Section 3. Classification of the PUs with zero degree of equivalence

Table G. Zero degree of equivalence in English, Spanish and German

Dog:

The hair of the dog

Horse:

Get on your high horse

Monkey:

Make a monkey out of somebody		
Not give a monkey's		

Table H. Zero degree of equivalence in English and German

Dog:

The dog's bollocks

Horse:

Closet he stable after the horse has bolted

Table I. Zero degree of equivalence in English and Spanish

Cat:

Put the cat among the pigeons

Dog:

See a man about a dog	
Love me, love my dog	
Give a dog a bad name	

Horse:

You can lead a horse to water, but you can't make him drink	
Put the cart before the horse	
Back to the wrong horse	

Monkey:

Throw a monkey wrench in the works	
Be cold enough to freeze the balls of a brass monkey	

"*PALE AS DEATH*" OR "*PÂLE COMME LA MORT*": FROZEN SIMILES USED AS LITERARY CLICHÉS

Suzanne Mpouli
LIP6, Université Pierre et Marie Curie
Labex OBVIL, Paris IV
mpouli@acasa.lip6.fr

Jean-Gabriel Ganascia
LIP6, Université Pierre et Marie Curie
Labex OBVIL, Paris IV
jean-gabriel.ganascia@lip6.fr

Abstract

The present study is focused on the automatic identification and description of frozen similes in British and French novels written between the 19th century and the beginning of the 20th century. Two main patterns of frozen similes were considered: adjectival ground + simile marker + nominal vehicle (e.g. *happy as a lark)* and eventuality + simile marker + nominal vehicle (e.g. *sleep like a top*). All potential similes and their components were first extracted using a rule-based algorithm. Then, frozen similes were identified based on reference lists of frozen similes and on the semantic distance between the tenor and the vehicle. The results obtained tend to confirm the fact that frozen similes are not used haphazardly in literary texts. In addition, contrary to how they are often presented, frozen similes often go beyond the ground or the eventuality and the vehicle to also include the tenor.

1. INTRODUCTION

Even though literary style is mainly associated with creative writing and deviations from stereotypes, some literary critics have argued that clichés can be used in literary texts for stylistic effects (Amossy & Herschberg-Perrot, 1997). Riffaterre (1964), for example, states that a cliché can either constitute a feature of the author's style that reinforces the literary status of the text or can serve to highlight the moral as well as social behaviours of a certain group of people. According to Abrams (1999), a cliché can be defined as "an expression that deviates enough from ordinary usage to call attention to itself and has been used so often that it is felt to be hackneyed or cloying". This definition definitely echoes the definition that Abrams (1999) gives of the trope: a figure "in which words or phrases are used in a way that effects a conspicuous change in what we take to be their standard meaning". In this respect, it can be said that clichés specifically refer to word combinations that started out as being creative, but, due to their popularity and the passing of time, became phraseological units. Such word combinations include, among others, dead metaphors and cliché similes, which have been vastly studied in phraseology (Wikberg, 2008).

Similes such as (1) "*Nos actions sont comme des bouts-rimés que chacun tourne comme il lui plaît*" and its translation (2) "*Our actions are like the termination of verses, which we rhyme as we please*" are figures of speech which rely on a linguistic marker to draw a parallel between some explicit or implicit properties that at least two semantically unrelated entities have in common. Since they generally follow the same structure as comparative statements and add concreteness to an utterance by introducing common knowledge, they constitute an inherent part of everyday language. In literary works, similes actively participate in rendering depictions and portrayals real, resonant or even surprising. Moreover, they are flexible enough to be effectively combined with other rhetorical figures such as metaphor, irony, hyperbole or alliteration (Shabat Bethlehem, 1996). Furthermore, as can be seen from the examples (1) and (2), they often have identical patterns in different languages.

The present study attempts to take advantage of the semantic and syntactic similarities of English and French as far as their simile constructions are concerned in order to, first, extract and mine similes in a corpus of French and British novels written between the 19th and the beginning of the 20th century, then determine whether each simile is cliché or not. In addition, it seeks to find out which relevant elements can be used to describe these similes from a corpus-based point of view. In its first section, this paper reviews the rhetorical structure of the similes and computational approaches proposed to identify similes and to account for their creativity. The second section describes the method used to extract as well as to identify cliché similes in written texts. The third section presents the corpus of novels and discusses the results obtained.

2. COMPUTATIONAL APPROACHES TO SIMILE IDENTIFICATION

In rhetoric, a simile such as (3) "This girl is graceful like a lily" is made up of the following components:

- a tenor or the object of comparison;

- a tertium comparationis or ground: the adjective or verb which denotes the property shared by the compared entities;

- a marker which is the linguistic trigger that introduces the comparison;

- a vehicle which refers to the term that establishes the reference against which the tenor is evaluated (Fishelov, 1993).

In addition to these components, Hanks (2012) also distinguishes the eventuality, the main verb on which the simile is built.

This girl	is	graceful	like	a lily.
tenor	eventuality	ground	marker	vehicle

Figure 1. Constituents of the simile "This girl is graceful like a lily"

As far as French is concerned, the comparative clause with a verb and an adjective ellipsis of the form "A est B comme C" is generally considered the prototypical form of the simile and has been extensively discussed in rhetoric in relation to the metaphor (Cohen, 1968). If 'comme' is undoubtedly the most frequently used simile marker in French, Shabat Bethlehem (1996) observes that be it in English dictionaries or in research articles, 'like'

and 'as' are generally presented as the main or the only simile markers of the English language. This reductionist view of the simile could explain why computational approaches to simile identification has been centred on a rather small number of markers ('like', 'as', 'as ... as', 'more … than', 'less … than').

Automatic simile detection per se can be divided into partial and full simile detection. Partial simile detection has mainly concerned with retrieving specific simile patterns (Veale & Hao, 2007; Veale, Hao & Li, 2008; Wikberg, 2008). It consists either in looking for all corresponding patterns in a corpus or in restricting the search to preselected grounds and vehicles. Moreover, it relies on heuristics or human judgment to differentiate similes from literal comparisons.

In contrast, full simile detection extracts and analyses all sentences containing a simile marker in unrestricted text, identifies the different components of the simile and separates literal similes from figurative ones. As there exists a correlation in comparative statements between the syntactic function of a word and its semantic role, Niculae (2013) uses dependency parsing to extract and to identify the components of English similes with 'like'. He also proposes a method to recognise creative similes by measuring the semantic distance between the tenor and the vehicle using distributional semantics. The approach suggested, however, only concerns similes that have both a nominal tenor and a nominal vehicle and require both nouns to be present in the corpus used to retrieve distributional statistics. In addition, it does not take into account certain syntactic structures such as juxtaposed and coordinated verbs.

3. PROPOSED METHOD

The present study is restricted to nominal similes of the form adjectival ground + simile marker + nominal vehicle (e.g. *cunning as a fox)* and eventuality + simile marker + nominal vehicle (e.g. *cry like a baby*). Since similitude or dissimilitude are primarily inferred by meaning, unlike previous works on cliché similes (Bolshakov, 2003), not only traditional markers, but also other markers which can be combined with these two predefined simile patterns were included. Table 1 lists all selected markers as well as their corresponding simile structures.

The proposed approach to the detection of frozen similes in literary texts can be divided into two main parts:

- extracting and mining of similes, comparisons and pseudo-comparisons;

- filtering all the extracted elements based on their idiomaticity and the semantic distance between the vehicle and the tenor.

	English	French	Possible structures
Prepositions and adverbs	like, unlike, as, as…as, more…than, less…than, -er … than	comme, ainsi que, de même que, autant que, plus…que, tel que, moins…que, aussi…que	- Verb + marker + vehicle - Tenor + verb + marker + vehicle - Adjectival ground + marker + vehicle -Tenor + adjectival ground + marker + vehicle -Tenor + verb + adjectival ground + marker + vehicle
Prepositional phrases		à l'image de, à l'instar de, à la manière de, à l'égal de, à la manière de, à la façon de	

Table 1. Similes markers for English and French

3.1. The simile extraction and mining module

The different steps detailed in this section are derived from the rule-based algorithm for simile mining presented in Mpouli and Ganascia (2015). The extraction and mining phase comprises various preprocessing tasks, simile candidate sentence extraction and finally, the identification of the components of each extracted simile candidate. Preprocessing tasks include tokenisation, part-of-speech tagging, syntactic chunking and sentence segmentation. The first three tasks are performed with TreeTagger (Schmid 1994), a freely available multilingual tokeniser, part-of-speech tagger and chunker[41]. The sentence boundaries defined by TreeTagger constitute the first basis for sentence segmentation and are corrected with specific rules when an ellipsis, a question or an exclamation mark are not followed by a capital letter.

```
Guests, like fish, begin to smell after three days.

 <NC>  Guests (NNS) </NC> ,  <PC> like (IN) <NC> fish  (NN) </NC>
</PC>  , (,)  <VC>  begin (VVP) </VC>  <VC>  to (TO) smell  (VV)
</VC> <PC>  after (IN) <NC>  three (CD) days (NNS) </NC> . (SENT)
```

Figure 2. Example of a chunked sentence

Syntactic chunking is an intermediary stage between part-of-speech tagging and constituency or dependency parsing and produce "flat, non-overlapping segments of a sentence that constitute the basic non-recursive phrases corresponding to the major parts-of-speech found in most wide-coverage grammars" (Jurafsky & Martin, 2009). These chunks combined with handcrafted rules are essential for the two next phases. Since chunks do not give information about the grammatical function of a word, the algorithm mainly relies on syntax, dependency grammar and syntactic clues. For example, based on the syntactic order prevalent in English and in French, it can be deduced that the vehicle would be the head noun of the noun phrase that follows the marker either directly or after an appositive phrase. In addition, the algorithm takes into consideration the ambiguity inherent to some comparative constructions and, depending on the sentence structure, labels all words that can plausibly be a component of the simile. Consequently, in a

[41] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

sentence such as "[...] a spark was kindled that wanted but opportunity to blaze into a flame, pure and bright as the shrine on which it burned", the sentence is analysed as follows:

- marker: 'as'
- vehicle: 'shrine'
- grounds: 'pure', 'bright' → tenor: 'flame'
- eventuality: 'blaze' → tenor: 'spark'

Table 2 summarises the different characteristics and textual clues used to identify each potential simile component.

Constituent	**Grammatical category**	**Informative Clues**	**Governor**
Adjectival ground	Adjective, past or present participle	Not separated from the marker by a coordinating conjunction, a relative pronoun, a preposition or a noun phrase	/
Tenor – Noun that the adjectival ground modifies	Common noun	Part of the noun phrase before or after the adjective	Non-predicative adjectival ground
Tenor – postposed direct object	Common noun	Not after a preposition Follows a verb or a prepositional phrase that follows a verb	Verb
Tenor – preposed direct object	Common noun	Part of the noun phrase directly before 'que', 'that', 'which' and the subject	Verb
Tenor – objective personal pronoun (direct object)	Personal pronoun	Directly before a verb	Verb
Tenor – subjective personal or demonstrative pronoun	Personal and demonstrative pronouns	Directly before or after a conjugated verb	Verb
Tenor – subject	Common or proper noun	Before a verb and not after a preposition	Verb
Eventuality	Verb	Not separated from the marker by a colon or a semi-colon	/
Vehicle – common noun	Common noun	Head of a noun phrase Not separated from a verb that follows him by a punctuation mark, a relative pronoun subject, a subjective personal pronoun, a coordinating or subordinating conjunction	Marker

Table 2. Correlation between each type of constituent, the clues to identify it and its grammatical function

3.2. Detection of frozen similes

In order to detect frozen similes, all ground/eventuality + vehicle couples were first compared to a list of frozen similes of the corresponding language, compiled from different sources such as *Les Comparaisons du français* by Nicolas Cazelles (1996) or *Dictionnaire français/anglais des comparaisons* by Michel Parmentier (2002). Since the resulting reference lists of frozen similes were mostly based on sources that were not contemporary with the novels of the corpus and were by no means comprehensive, frequency and semantic distance were used to single out frozen similes among the remaining ground/eventuality + vehicle couples. All ground/eventuality + vehicle couples that appeared at least 5 times in novels by different authors were thus selected. Then, the semantic distance between the tenor and the vehicle in each sentence was assessed.

When both the tenor and the vehicle were nouns, their semantic categories were extracted from two machine-readable dictionaries: Wordnet (Fellbaum, 1998) and *Le Dictionnaire électronique des mots* (Dubois & Dubois-Charlier, 2010) for English and French respectively. As far as Wordnet is concerned, the number of the noun's corresponding lexicographer file, which precedes the term in every definition, was taken as its semantic category. For *Le Dictionnaire électronique des mots,* three semantic categories are provided: 'animal' and 'humain' (human being) and 'non-animé' (inanimate).

Semantic distance, however, required manual disambiguation in the following cases:

- when different semantic categories were attributed to either the tenor or the vehicle;
- when an English tenor or vehicle was classified as a unique beginner for nouns;
- when both French nouns were tagged as 'non-animé';
- when the ground/eventuality + vehicle couple only used a personal pronoun as tenor.

My cousin behaves like a child. 09972010 **18** n 04 cousin 09918248 **18** n 02 child 18 Noun denoting people	Sa voix était aussi faible qu'un souffle. - voix: type="non-anime" - souffle: type="non-anime"

Figure 3. Semantic categories extracted from Wordnet (left) and *Le Dictionnaire électronique des mots* (right)

4. RESULTS AND DISCUSSION

4. 1. Presentation of the corpus

The two corpora used for this experiment were built with digital versions of literary texts in the public domain, collected mainly from the Project Gutenberg website[42] and from the Bibliothèque électronique du Québec[43], for British and French novels respectively. Most of the novels included in the corpus were written during the 19th century so as to ensure linguistic homogeneity and because that century witnessed the novel imposing itself as a predominant literary genre. In addition, a ratio of least 3 novels per writer was

[42] www.gutenberg.org
[43] beq.ebooksgratuits.com

observed and the novels were not restricted to a specific literary genre. This method enabled to create a corpus of 1191 British novels authored by 62 writers and a corpus of 746 French penned by 57 novelists. In terms of size, the British corpus contains 152,941,750 tokens and its French counterpart, 119,914,914 tokens.

4.2. Description of detected frozen similes

The rather low number of total occurrences of the top frozen similes in each language (listed in Table 3) seems to confirm the fact that clichés in general and frozen similes in particular are far from being common in literary texts. In addition, French novelists are more inclined to use frozen similes than British ones. One striking result of this study is the fact that the same simile (*pale + marker + death* and *pâle + marker + mort*) is the most frequently used in both languages. This simile is also interesting in itself, from a stylistic point of view, as it assigns a human feature to an abstract entity. As a matter of fact, 'pale' and its synonym, 'white', appear several times in the most frequent similes. Similarly, 'death' is used in three of the top frozen English similes. Apart from the fact that 'paleness' mostly characterises humans, its context of occurrence suggests that it is generally used to stress somebody's distress or fear in reaction to a particular news or event. It also sometimes conveys the impression that the narrator has of a protagonist. Generally speaking, this simile comes in rather such sentences for maximal effect, such as in: "The old Cavalier looked pale as death, and greatly agitated".

English	French
pale + marker + death (152)	*pâle + marker + mort (283)*
cold + marker + ice (128)	*pleurer + marker + enfant (188)*
bad + marker + death (114)	*immobile + marker + statue (179)*
white + marker + death (108)	*rapide + marker + éclair (164)*
good +marker + gold (108)	*blanc + marker + neige (162)*
white + marker + sheet (102)	*aimer + marker + frère (140)*
good + marker + word (98)	*tomber + marker + massue (135)*
come+ marker + shock (80)	*tuer + marker + chien (122)*
black+ marker + night (87)	*pâle + marker + morte (121)*
white +marker +snow (83)	*beau + marker + ange (115)*
silent + marker +grave (83)	*passer +marker+ éclair (112)*
clear +marker + daylight (83)	*rapide +marker + pensée (106)*

Table 3. The 12 most frequent frozen similes in both corpora

Furthermore, several detected frozen similes, in fact, are variants of another frozen simile. Frequency could therefore help to determine the main simile that the others seek to replace. For example, 'motionless + marker + statue', 'rigid + marker + statue', 'motionless + marker + image' are all non-idiomatic variations of 'immobile + marker + statue'.

Still based on their frequency in each corpus, it is possible to propose a scale of clichédness that can further describe similes for stylistic analysis. For instance, 'white as the snow" would be a prominent literary cliché, while 'heavy + marker + lead' (42 occurrences) would be a medium literary cliché and 'harmless + marker + dove' (6 occurrences) would be a relatively rare literary cliché.

As far as compositionality is concerned, some frozen similes use various markers while others always use the same one, especially when they make use of the comparative forms "more/less… than" or when they are derived from proverbial expressions such as "blood

is thicker than water". In addition, some frozen similes tend to always be associated with the same tenor or different tenors belonging to the same semantic fields. Examples of this type of similes include 'eyes + wide + saucers', 'cheveu +noir + aile de corbeau', ' nez +recourbé + bec' and 'yeux + brûler/briller + charbon'. In this respect, it is possible to say that some literary frozen similes are restricted to specific cognitive associations, so much so that they come automatically in mind when one wants to emphasise a particular state of an entity.

5. CONCLUSION

The aim of this paper was to study frozen literary similes in a corpus of novels written in English and in French. If the greater part of the extraction of similes and of their components was done automatically, the recognition of frozen similes still relies partially on human knowledge. In this respect, it seems necessary for future work to research how to disambiguate first the tenor or the vehicle so as to find directly its corresponding semantic category. Since this study is mostly oriented towards a stylistic description of similes, another perspective would be to add information about the literary clichédness of a particular simile to a simile mining system, in order to shade new light on the perception and interpretation of frozen similes in literature as a whole.

Acknowledgment

This work was supported by French state funds managed the ANR within the Investissements d'Avenir programme under the reference ANR-11-IDEX-0004-02.

References

ABRAMS, M. H., 1999. *A Glossary of Literary Terms*. Boston: Heinle and Heinle.

AMOSSY, R. AND HERSCBERG-PERROT, A., 1997. *Stéréotype et clichés. Langue – Discours – Société*. Paris: Nathan.

BOLSHAKOV, I. A., 2003. Simile cliché phrasemes in colloquial language. *Proceedings of the First International Conference on Meaning-Text Theory*, [online] Available at: http://meaningtext.net/mtt2003/proceedings/10.Bolshakov.pdf [Accessed 30 March 2015].

COHEN, J., 1968. La comparaison poétique : Essai de systématique. *Langages*, 3 (12), pp. 43-51.

DUBOIS, J. AND DUBOIS-CHARLIER, F., 2010. La combinatoire lexico-syntaxique dans *Le Dictionnaire électronique des mots*. Les termes du domaine de la musique à titre d'illustration. *Langages*, 3 (179-180), pp. 31-56.

FELLBAUM, C., 1998. *Wordnet: An Electronic Database*. Cambridge, MA: MIT Press.

FISHELOV, D., 1993. Poetic and non-poetic simile: Structure, semantics, rhetoric. *Poetics Today*, 14 (1), pp. 1-23.

HANKS, P., 2012. The Roles and structure of comparisons, similes and metaphors in natural language. Available at:http://www.english.su.se/polopoly_fs/1.100637.1347450592!/menu/standard/file/SMF_2012_Patrick_Hanks_plenary.pdf [Accessed 9 July 2015].

JURAFSKY, D., AND MARTIN, J. H., 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition and Computational Linguistics.* New Jersey: Prentice-Hall.

MPOULI, S., AND GANASCIA, J.-G., 2015. Extraction et analyse automatique des comparaisons et des pseudo-comparaisons pour la détection des comparaisons figuratives. *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles,* pp. 621-627.

NICULAE, V., 2013. Comparison pattern matching and creative simile recognition. *Proceedings of the Joint Symposium on Semantic Processing, Textual Inference and Structure in Corpora*, pp. 110-114.

RIFFATERRE, M., 1964. Fonctions du cliché dans la prose littéraire. *Cahiers de l'Association internationale des études françaises*, 16, pp. 81-95.

SCHMID, H., 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44-49.

SHABAT BETHLEHEM, L., 1996. Simile and figurative language. *Poetics Today*, 17(2), pp. 203-240.

VEALE, T. AND HAO, Y., 2007. Learning to understand figurative language: From similes to metaphor to irony. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, pp. 683-688.

VEALE, T., HAO, Y., AND LI, G., 2008. Multilingual Harvesting of Cross-Cultural Stereotypes. *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics,* pp. 523-531.

WIKBERG, K., 2008. Phrasal similes in the BNC. In: S. Granger and F. Meunier, eds. 2008. *Phraseology: An Interdisciplinary Perspective*. Amsterdam and Philadelphia: John Benjamins, pp. 127-142.

EXTRACTING TERMS WITH EXTra

Lucia C. Passaro
CoLing Lab
Dipartimento di Filologia,
Letteratura e Linguistica
University of Pisa (Italy)
lucia.passaro@for.unipi.it

Alessandro Lenci
CoLing Lab
Dipartimento di Filologia,
Letteratura e Linguistica
University of Pisa (Italy)
alessandro.lenci@unipi.it

Keywords: Term recognition, Multiword expressions, Information extraction, Ontology population, Automatic Indexing

Abstract

The identification and extraction of terms play an important role in many areas of knowledge-based applications, such as automatic indexing, knowledge discovery and management, as well as in computational approaches to terminology and lexicography. In this paper, we present EXTra, a tool designed to extract and calculate the degree of termhood of multiword expressions as a function of the statistical distribution of their parts and of the presence of other sub-terms. This work describes EXTra's algorithm, and provides the results of its evaluation on a task of term extraction from an Italian corpus of documents belonging to the domain of Public Administration.

INTRODUCTION

In recent years, the development of robust approaches to terminology extraction is playing an important role in many areas of knowledge-based applications such as automatic indexing, knowledge discovery and knowledge management. The need for domain terminology extraction has emerged from different disciplines and to answer to various goals, such as dictionary and thesaurus construction, text indexing, machine translation, automatic summarization, etc.

A general definition of *term* is "a surface representation of a specific domain concept" (Jacquemin, 1997; Pazienza, 1999). In general, a term can be either a single word or a multiword unit. In this study we focus on the latter kind of terms. The bag-of-words model, based on single word terms, is in fact a simplified representation of the lexicon used in natural language processing (NLP) and information retrieval (IR). We assume that "multiword expressions" (i.e. complex terms) range from completely opaque idioms to

semantically compositional word combinations (Evert, 2008). Multiword terms are less ambiguous and less polysemous than single word terms, yielding a better representation of the document content. Moreover, the lion share of domain concepts are normally expressed through multiword terms, which represent a crucial component of natural language lexicons (Jackendoff, 1997).

Term Extraction is a key application in Information Extraction (IE) and IR, and a crucial component to tackle several NLP tasks, such as Ontology Learning and Ontology population, Key-words extraction and Document Indexing. The recognition of complex terms from texts is performed on the basis of different criteria. Major differences exist between algorithms that take into account only the distributional properties of terms, such as frequency and TF / IDF (Salton and McGill, 1983), and those using contextual information such as syntactic, terminological and semantic features as in Maynard and Ananiadou (2000), Frantzi and Ananiadou (1999), Maynard (2000), Dell'Orletta et al., 2014, and Bonin et al. (2010). The common trait of most of the strategies above is the identification of a set of ranked candidates from texts, and then the application of a filtering function to separate real terms from non-terms. In this latter phase, the candidates are usually sorted according to their association strength as an estimate of their degree of termhood.

We have organized this paper as follows: In section 1, we present the term extractor EXTra by describing its approach to the candidate selection step, its original weighting algorithm and its possible parameters. In section 2, we report the evaluation of EXTra to extract domain terminology from documents belonging to the Italian Public Administration (PA). Section 3 reports the results obtained from the validation of the terms in this case study, focusing on the precision and on the quality of the ranking produced by EXTra.

1. EXTRA

The term extractor EXTra takes into account the linguistic structure of multiword terms by implementing a candidate selection step that uses manually-defined *structured* PoS-patterns. Moreover, in order to tackle the complexity of term phrases, EXTra adopts a new association measure that promotes terms composed by one or more sub-terms. The intuition is that the degree of termhood of a candidate pattern is a function of the statistical distribution of its parts, and of the presence of highly weighted sub-terms. The last step of EXTra applies a filtering function to separate real terms from wrong candidates. EXTra also includes various parameters that allow the user to optimize the extracted terms with respect to the target corpus and domain. In particular, the user can specify the set of structured patterns that guide the extraction process, a list of stopwords, and the thresholds for the association measure and the n-gram frequency. In the configuration file, the user also selects the association measure used by the weighting algorithm. The association measures currently implemented in EXTra are the Pointwise Mutual Information (Church and Hanks, 1990), the Local Mutual Information (Evert, 2008), and the Log Likelihood Ratio (Dunning, 1993), as well as an identity function weighting the n-grams with their raw frequency. In order to assure the flexibility of the system, a further parameter affects the importance given to long terms by the weighting algorithm (cf. section 1.2). The input of EXTRA is a PoS-tagged and lemmatized text in a tab-delimited CONLL format. The output of EXTra consists of two files: the input file enriched with the extracted multiword terms, and a list of multiword terms ranked according to their termhood.

TokenID	Token	Lemma	PoS	Term (EXTra)
1	Registro	registro	S	_
2	Generale	Generale	SP	_
3	n.	n.	SA	_
4	961	961	N	_
5	Del	di	EA	_
6	26/09/2012	26/09/2012	N	_
1	Copia	copia	S	_
1	Determinazione	determinazione	S	_
1	n.	n.	SA	_
2	140	140	N	_
3	del	di	EA	_
4	26/09/2012	26/09/2012	N	_
1	Settore	settore	S	_
2	Servizi	Servizi	SP	_
3	Finanziari	Finanziari	SP	_
1	Oggetto	oggetto	S	_
2	:	:	FC	_
3	Affidamento	Affidamento	SP	_
4	del	di	EA	_
5	servizio	servizio	S	servizio_di_manutenzione
6	di	di	E	servizio_di_manutenzione
7	manutenzione	manutenzione	S	servizio_di_manutenzione
8	presidi	presidio	S	_
9	antincendio	antincendio	A	_
10	posti	posto	S	_
11	negli	in	EA	_
1	edifici	edificio	S	edificio_di_proprietà
2	di	di	E	edificio_di_proprietà
3	proprietà	proprietà	S	edificio_di_proprietà

Term (EXTra)	PLMI
bollettino_ufficiale_di_regione_autonomo	2059330,459
carta_libero_ad_uso_amministrativo	1363747,944
carta_libero_per_uso_amministrativo	1363747,944
originale_in_carta_libero	453971,4196
copertura_finanziaria	90748,258
attribuzione_al_dirigente_di_dotazione	78023,94
diminuzione_permanente_di_capacità_lavorativo	74245,3462
protezione_dei_dati_personali	62194,11451
trattamento_di_dati_personali	61056,1527
trattamento_di_dati_personali	61056,1527
incarico_di_dirigenza_di_settore	60877,00747
dinamica_costiero_in_unità_fisiografica	56705,82765
codice_identificativo_di_gara	52275,99217
intervento_di_assistenza_sociale	46027,35039
cambio_di_destinazione_di_uso	44029,83091
cambio_della_destinazione_di_uso	44029,83091
briglia_esistente_su_torrente_lucido	42739,01878
equiparazione_stabilito_da_legge_vigente	42473,67371
dichiarazione_sostitutivo_di_atto_notorio	42074,08231
rispetto_di_normativa_previsto	39939,49118

Figure 2. Examples of output files produced by EXTra

1.1. Candidate selection

Candidate terms are identified using manually-defined *structured PoS patterns* that represent the recursive phrase structure of terms. A *structured PoS pattern* is a bracketed list of constituents, where each constituent can be either a sequence of two *content PoS* or another bracketed constituent. This structure defines long term patterns as a composition of smaller patterns. The c*ontent PoS* are specified in the configuration file, allowing the user to exclude from the termhood computation particular classes of PoS (e.g. articles and prepositions). The following is an example of *structured PoS pattern*:

[[noun (-s), preposition (-e), noun (-s)], preposition (-ea), [noun (-s), adjective (-a)]]

It is composed by two constituents, [noun (-s), preposition (-e), noun (-s)] and [noun (-s), adjective (-a)]. This structured pattern identifies the candidate "Politica di sviluppo delle Risorse Umane" *(human resource development policy)*. Following the pattern structure and ignoring prepositions, we can isolate two embedded sub-terms: [politica-s di-e sviluppo-s] ([noun (-s), preposition (-e), noun (-s)]) and [risorse-s umane-a] ([noun (-s), Adjective(-A)]). From a computational point of view, during the candidate selection phase, EXTra first stores the statistical information of each sub-patterns (e.g., the frequencies of the embedded pairs <Politica, Sviluppo> and <Risorse, Umane>), and then stores the frequency of the aggregate pair <politica_di_sviluppo, risorse_umane>.

1.2. Weighting algorithm

The structure of the PoS patterns is also used to guide the process of statistical term weighting by following the same order of incremental composition. Following a recursive structure, the weighting algorithm assigns a termhood score to each of the embedded phrases, and then computes the global score for the complex term by combining the partial weights of its components.

EXTra's term weighting algorithm is applied recursively to the internal structure of the patterns: At the base step it measures the association strength σ of each candidate two-

word term $< w_1, w_2 >$by computing standard association measures, such as for instance Pointwise Mutual Information (PMI). The candidates whose score σ is above an empirically fixed threshold are added to the set of the terms $T=\{t_1,...,t_n\}$. In the recursive step, EXTra measures the association strength σ of any n-word candidate term $<c_1, c_2>$ by combining the association strengths of its sub-elements. The termhood of a candidate is calculated using the following formula:

$$\sigma(c_1, c_2) = S(c_1) * S(c_2)$$

If $c_i \notin T$, $S(c_i) = 1$, else $S(c_i) = (\log_2 \sigma(c_i)) / k$. As we said above, this weighting scheme formalizes the assumption that the termhood of longer terms depends on the degree of termhood of their parts. The parameter k controls the contribution of sub-terms to the weight of longer terms: The smaller the k, the higher the weight assigned to longer terms containing them.

Coming back to the previous example

[[politica (-s), di (-e), sviluppo (-s)], delle (-ea), [risorse (-s), umane (-a)]]

in the base step, EXTra measures the association strength σ of each two-word term $<w_1, w_2>$ using standard association measures. Supposing that the selected association measure is the PMI, at the base step EXTra measures the scores for the pairs <risorse-s, umane-a> and <politica-s, sviluppo> and it stores their termhood value. In the recursive step, the system calculates the score σ between the sub-candidates <politica_di_sviluppo, risorse_umane> by applying the formula:

σ (politica_di_sviluppo, risorse_umane) = S (politica_di_sviluppo) * S (risorse_umane).

Since $\sigma(< c_1, c_2 >)$as$S(c_1)$*$S(c_2)$both the sub-terms belong to the set of accepted terms T, the termhood score σ is calculated using the formula $S(c_i) = (\log_2 \sigma(c_i)) / k$.

1.3. Filtering

Candidate multiword terms are filtered by using three main filters. First of all, an optional stoplist is used to exclude the terms containing one or more words in the blacklist during indexing. Then, patterns with a frequency below a frequency threshold are discarded before computing their strength of association. Finally, the association measure filter defines the minimum strength of association that an n-gram must have to be considered as a multiword term: The candidates whose score σ is above an empirically fixed threshold are then added to the set of terms T.

2. EVALUATING EXTRA

We have evaluated EXTra on a term extraction task in the Italian Public Administration (PA) domain. This is a particularly challenging domain because of the highly heterogeneous nature of its terminology, which typically includes domain terms belonging to the multifarious fields covered by PA, ranging from the management of schools up to urban planning and health care.

As a preliminary step, we automatically collected PAWaC! (Public Administration Web as Corpus) which contains documents extracted from the Italian online "Albo pretorio" (Council notice board) of various small and medium municipalities in Tuscany. Most of these documents are "Delibere" (Town council resolutions), "Determine" (Executive resolutions), and generic administrative acts, such as bidding processes, local regulations etc. PAWaC includes 15,321 documents, for a total of 34,725,652 tokens and 17,272,068 content words (nouns, adjectives and verbs). We PoS-tagged the corpus using the PoS-Tagger described in Dell'Orletta (2009) and we identified the list of structured PoS patterns showed in Table 5. Since we were mainly interested in extracting nominal phrases, the list of the patterns only include nouns, adjectives and prepositions (Justeson et al., 1995).

PoS Structured pattern	**Example**
[noun, adjective]	delibera comunale (*municipal resolution*)
[noun, preposition, noun]	presidente del consiglio (*Prime Minister*)
[[noun, adjective], preposition, noun]	delibera comunale di giunta (*municipal council resolution*)
[[noun, adjective], preposition, [noun, adjective]]	gestione provvisoria delle risorse finanziarie (provisional management of financial resources)
[noun, preposition, [noun, adjective]]	ordine di regolarità contabile (*accounting consistency order*)
[noun, preposition, [noun, preposition, noun]]	approvazione del verbale di gara (approval of the bidding process)
[[noun, preposition, noun], preposition, [noun, adjective]]	politica di sviluppo delle risorse umane (*human resource development policy*)
[noun, preposition, [noun, preposition, [adjective, noun]]]	Attestazione del responsabile del servizio finanziario (*declaration of the financial service manager*)

Table 5. Structured PoS patterns

We have made several experiments (section 2.1.1) with EXTra, in which we used the same set of PoS patterns but with different association measures and different k values in order to assign a different boost to long terms. Although the patterns include prepositions, they are not considered in the computation of termhood, which is calculated only considering the strength of association between nouns and adjectives.

2.1.1. Experiments

We tested EXTra with three different association measures (one of which is the raw term frequency, which we used as a baseline) and three values for the parameter k, in order to control the importance assigned to long terms in the ranking. For each configuration, from left to right Table 6 shows the association measure, the value of the parameter k , the minimum n-gram frequency and the number of extracted terms.

Configuration	Association measure	*k*	Min. Freq	#Terms
Frequency.Knull.3	Frequency	-	3	65,120
PLMI.K1.3	Positive Local MI	1		58,380
PLMI.K5.3		5		58,380
PLMI.K10.3		10		58,380
PPMI.K1.15	Positive Pointwise MI	1	15	13,032
PPMI.K5.15		5		13,032
PPMI.K10.15		10		13,030

Table 6. Configuration

In the case of the baseline configuration, we do not provide any boost to long terms, hence the parameter k is not specified. In our experiments, we used the Positive LMI (PLMI) and the Positive PMI (PPMI), in which negative scores are changed to zero, and only positive ones are considered. Following Evert (2008), PMI has been calculated as $log_2 (O/E)$ and the LMI has been calculated as $O * log_2 (O/E)$, where O is the observed co-occurrence frequency and E is the expected frequency under the null hypothesis of independence (i.e. complete absence of association). For PPMI and PLMI models, we specified the following values of k: $k = 1$ (maximum boost for long terms), $k = 5$ (medium boost) and $k = 10$ (low boost). The frequency threshold has been set to 3 in the models based on frequency and PLMI, but it was increased to 15 in the PPMI ones, because of the well-known bias of this association measure towards low-frequency n-grams.

3. RESULTS

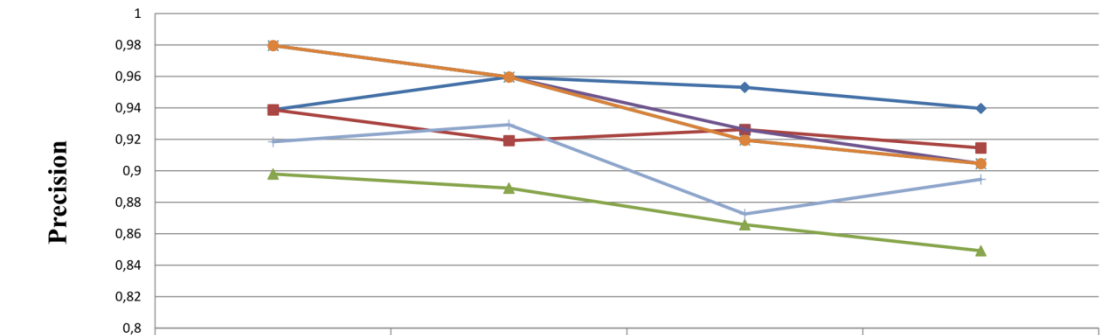
In order to evaluate the precision of EXTra, a domain expert judged the top 200 terms produced by EXTra for each configuration in Table 6. The annotator was asked to decide whether a candidate term was both a valid multiword expression and a domain-specific term in the field of PA. For example, the candidates "Ponte levatoio" (drawbridge) was discarded because it is not a domain term, while the candidate "Documento di identità in corso" was discarded because it was a truncation of the term "Documento di identità in corso di validità" (Valid identity document). Both the previous candidates were therefore labeled as False Positives (FP). On the contrary, the candidate "Esercizio finanziario" (fiscal year) complies with both requirements, and therefore was considered a True Positive (TP).

The global Precision was calculated for the top 200 evaluated terms as TP / (TP + FP). Table 7 shows the results. The baseline model (Frequency.Knull.3) obtained a precision score of 0.89. The best model is PLMI with the maximum weight for long terms ($k = 1$). The worst model, outperformed by the baseline, is PLMI with the lowest salience assigned to long terms ($k = 10$). The precision for PPMI models is stable with respect to the parameter k, scoring 0.905. On average, the models reach a precision score of ~0,9.

Configuration	k	Min. Freq	#Terms	Precision
Frequency.Knull.3	-	3	65,120	0.89
PLMI.K1.3	1	3	58,380	**0.935**
PLMI.K5.3	5	3	58,380	0.915
PLMI.K10.3	10	3	58,381	0.85
PPMI.K1.15	1	15	13,032	0.905
PPMI.K5.15	5	15	13,032	0.905
PPMI.K10.15	10	15	13,030	0.905

Table 7. Global precision

The quality of the termhood ranking produced by EXTra has been evaluated by considering the Precision@n with $1 \leq n \leq 200$. Figure 3 reports the Precision@n for $n \in \{50, 100, 150, 200\}$. We can observe that for the top 50 terms, the best performing models use PPMI, reaching a Precision of 0.98. Considering top 100 terms, the precision decreases for PPMI models, and increases for the best PLMI one. Going down in the ranking, the precision of all models decreases, as expected. If we consider the quality of the ranking (Figure 3), we can notice that for the top 50 terms, the discriminating factor lies in the type of association measure. In fact, all PPMI models reached a P@50 equal to 0.98. PLMI models, on the contrary, reached a score ranging from 0.90 and 0.94. In addition, we can observe that PLMI, but not PPMI models are influenced by the *k* parameter. The results concerning the PLMI models show a trend in which precision seems to be inversely proportional to the parameter *k*. In other words, the precision of the model decreases when we reduce the importance of sub-terms. This trend is not evident in the PPMI models, in which the precision seems to be independent of the parameter *k*.



	P@50	P@100	P@150	P@200
PLMI.K1.3	0,9388	0,9596	0,953	0,9397
PLMI.K5.3	0,9388	0,9192	0,9262	0,9146
PLMI.K10.3	0,898	0,8889	0,8658	0,8492
PPMI.K1.15	0,9796	0,9596	0,9262	0,9045
PPMI.K5.15	0,9796	0,9596	0,9195	0,9045
PPMI.K10.15	0,9796	0,9596	0,9195	0,9045
Frequency.Knull.3	0,9184	0,9293	0,8725	0,8945

Figure 3. Ranking evaluation

Presumably, the contrast between PPMI and PLMI determines the different behavior of the models. In fact, PPMI favors more idiosyncratic, low-frequency expressions, while PLMI has a greater bias towards frequent expressions (Evert 2008). This might be the reason why the weighting algorithm works better with PLMI, in which the boost given to long terms is more evident.

Error analysis shows that a great portion of the FPs depends on the fact that the candidates were correct multiword terms that did not belong to the domain of the PA. This fact prompts us to enrich EXTra with additional features to identify genuine domain terms, for instance by computing a confidence score based on the distribution of the terms in domain vs. general corpora (Penas et al., 2001; Chung et al., 2004; Basili et al., 2001).

4. CONCLUSIONS

In this paper, we have introduced EXTra, a term extractor designed in order to identify multiword terms taking into account both their linguistic structure and their internal complexity. In EXTra, the degree of termhood of a candidate pattern is a function of the statistical distribution of its parts, and of the presence of highly weighted sub-terms. EXTra only requires a PoS-tagged corpus and a set of PoS-patterns defining the phrase structure of candidate terms. Therefore, it can easily be adapted to different languages and domains in an economic and very scalable way.

The proposed methodology has been tested on the domain of Italian PA achieving very good results. However, we are aware that a better evaluation of EXTra requires us to compare the extracted terms against domain-specific terminological resources such as ontologies or thesauri, which we plan to do in the near future. Moreover, we aim at implementing additional association measures, a more efficient way of specifying the structured PoS patterns and statistical filters to single out genuine domain multiword expressions from general ones.

Acknowledgments

Lucia C. Passaro received support from the Project SEMantic instruments for PubLIc administrators and CitizEns (SEMPLICE), funded by Regione Toscana (POR CReO 2007-2013). We wish to thank Anna Gabbolini for helping us in validating EXTra on the domain of Italian Public Administration.

References

BASILI R., MOSCHITTI A., PAZIENZA M. T. AND ZANZOTTO F. M. (2001). *A contrastive approach to term extraction*. In Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA–2001), Nancy.

BONIN, F., DELL'ORLETTA, F., MONTEMAGNI, S., VENTURI, G. (2010). *A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora*. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta (Malta), May 19-21 2010: European Language ResourcesAssociation (ELRA).

CHUNG T. M. AND NATION P. (2004). *Identifying technical vocabulary*. System, 32, 251–263.

CHURCH, K. W. AND HANKS, P.(1990). *Word association norms, mutual information, and lexicography*. Computational Linguistics, 16(1), 22–29.

DELL'ORLETTA F.,VENTURI G., CIMINO, A., MONTEMAGNI, S., (2014). *T2K2: a System for Automatically Extracting and Organizing Knowledge from Texts*. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14). Reykjavik (Iceland), May 26-31, 2014: European Language Resources Association (ELRA).

DELL'ORLETTA, F. (2009). *Ensemble system for Part-of-Speech tagging*. In Proceedings of EVALITA 2009, Reggio Emilia, Italy.

DUNNING, T. E. (1993). *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics, 19(1), 61–74.

EVERT, S. (2008). *Corpora and collocations*. In A. Lüdeling and M. Kytö (eds.), Corpus Linguistics. An International Handbook, article 58. Mouton de Gruyter, Berlin.

FRANTZI K.T. AND ANANIADOU S., 1999. *The C-Value/NC-Value domain independent method for multi-word term extraction*. Journal of Natural Language Processing, 6(3):145–179.

JACKENDOFF R. (1997). *The Architecture of the Language Faculty*. Cambridge, MA: The MIT Press.

JACQUEMIN, C., (1997). *Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Universitè de Nantes, France (1997).

JUSTESON, J. S. AND KATZ, S. M.. 1995. *Technical terminology: some linguistic properties and an algorithm for identification in text*. Natural Language Engineering, 1:9–27.

KAGEURA, K. AND UMINO, B. (1996). *Methods of automatic term recognition: a review*. Terminology, 3(2), pp. 259–289.

MAYNARD D.G AND ANANIADOU S., 2000. *Identifying terms by their family and friends*. In Proc. of 18th International Conference on Computational Linguistics (COLING). Saarbrücken, Germany, July 31 - August 4, 2000: Association for Computational Linguistics.

MAYNARD D.G., 2000. *Term Recognition Using Combined Knowledge Sources*. PhD thesis, Manchester Metropolitan University, UK, 2000.

PAZIENZA, M.T., (1999). *A domain specific terminology extraction system*. In: International Journal of Terminology. Benjamin Ed., Vol.5.2 (1999) 183-201.

PENAS A., VERDEJO F. AND GONZALO J. (2001). *Corpus-Based Terminology Extraction Applied to Information Access*. In Proceedings of Corpus Linguistics 2001, 458–465.

SALTON,G. AND MCGILL M. J., 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

STATISTICAL AUTOMATIC EXTRACTION OF V-N ITALIAN COLLOCATIONS FROM AN ACADEMIC SPOKEN CORPUS

Diana Peppoloni
University for Foreigners of
Siena - Italy
dianapeppoloni@gmail.com

Abstract

The aim of the present paper is to describe the results of a method of semi-automatic extraction of Italian V-N collocations, taken from the annotated medium size corpus ASIC (Academic Spoken Italian Corpus) (Peppoloni, 2012).

This corpus consists of audio recordings of academic lectures on various subjects, including about 500,000 words.

Collocations, that represent a subclass of multi-word expressions, are crucially involved in many aspects of language research: it emerges from the seminal works of Palmer (1933) on language teaching, from those linked to lexicography (Benson, Benson & Ilson, 1986; Benson, 1990; Cowie, 1981, Granger & Meunier, 2008) and to natural language processing (Smadja, 1993; Calzolari et al., 2002; Sag et al., 2002), going through important corpus linguistics (Sinclair, 1991) and language acquisition (Nesselhauf, 2005) studies.

Being observed from so many different perspectives, we currently relate with a large number of different definitions of collocations; in this study it is adopted that proposed by Evert (2005): "A collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon" (Evert, 2005: 17).

Many researches (Biber, 2006) have analysed academic lexicon starting from collocations, that is to say recurrent combinations of words that tend to occur together to form fixed expressions with a global and conventional meaning. A misuse of the terms that make up collocations, can lead to misunderstandings in communication, altering the message that words bring. To know a word, means to know the other terms with which this usually combines; Firth (1957) says "you shall know a word by the company it keeps". This process promotes an easier way in producing and understanding sentences and concepts. Native speakers do not store individual lexical entries, but rather entire fixed words sequences, not having to rebuild it all the time, but using it already formed.

In order to extract V-N collocations from the corpus ASIC, we tested the suite of statistical tools offered by the CWB platform, brought up by the Institute for Natural Language Processing of the University of Stuttgart. Its central component is the flexible and efficient query processor CQP. Our script provides the possibility to insert optional linguistic constituents, placed between the possibly identified verbs and nouns, in the automatic research; the resulting research pattern is the following: VERB (ADVERB) (ARTICLE) (ADJECTIVE OR NUMBER) NOUN.

The result of this computational operation is a list of 341 V-N collocations, with a frequency of occurrence equal to or greater than 3. Not all the identified word combinations corresponded though in attested Italian collocations. These were then assessed by 50 Italian native speakers non-

linguistic experts, who evaluated respectively groups of 10 collocations, indicating which, according to them, were or were not valid for the Italian language. This crowd sourcing experiment has allowed us to obtain impartial linguistic data, not influenced by the professional background of the speakers as linguists, but only by their linguistic competence as native speakers.

1. DEFINING ACADEMIC LEXICON

The most widely used definition of Academic Lexicon, is that provided by Nation (2001), who postulates that it is made of high-frequency words used in academic texts and discourses, regardless to the subject treated.

Academic lexicon is tightly linked to the learning process and the didactic activities with which foreign students daily come in contact with during lessons, exams, conferences and so on. A right and immediate decoding of some lexical units such as *context*, *theory*, *approach*, or some multi-words highly conventional expressions such as *to introduce a concept* or *to deal with a topic*, is a fundamental step to proceed successfully throughout the academic career and then to get into the working world. Academic Language is not only the written language typically found in textbooks, but also that used daily in classrooms, in conferences or during an oral exam. Mastering Academic Language is a challenge for all students, but it is a difficult task especially for foreigners students. The immediate access to the meaning of a word allows them to focus on the content of what they try to explain, rather than on the way to explain that concept.

The present study focuses on Academic Language (AL) because it is an important tool for gaining, sharing and developing both practical and theoretical knowledge within any field of study. As stressed by Nagy and Townsend (2012:92) "academic language is the specialized language, both written and oral, of academic settings that facilitates communication and thinking about disciplinary content", i.e. it is a tool which enables us to develop and convey abstract and technical ideas and facts about phenomena with precision and subtle nuances. Even if specific fields of study are usually identified by the use of specialised technical vocabularies and certain genre norms, academics are however required to master not only technical vocabulary within their own field of study, but also the more general Academic Vocabulary (AV), which is used across faculty boundaries and which characterizes AL (Coxhead 2000). Academic words (e.g. anticipate, compatible and phenomenon), and phrases (e.g. as a function of, in the case of, to some extent) can be considered the glue of academic language, serving a range of functions in relation to presenting information, building argumentation, scaffolding, signposting, quantifying and stance-setting the content proper, i.e. the technical information to be conveyed in the academic text (Biber 2006; Hyland 2008; Simpson-Vlach & Ellis 2010). Thus, AV in the form of both single and multi-word item, e.g. collocations, must be easily accessible for the language user. Research has, however, shown that even very advanced second-language users have problems in acquiring and using collocations (Arnaud & Savignon 1997, Nesselhauf 2005, Laufer & Waldman 2011, Henriksen 2013, Westbrook & Henriksen 2014). Collocations, recurring two-to-three word syntagmatic units, are a subset of formulaic sentences which primarily serve a referential function (Nattinger and DeCarrico 1992, Howarth 1998). Academic collocations knowledge is important for disambiguating meaning, understanding connotational meaning and presenting factual knowledge with precision. Mastery of academic collocations is often seen as a sign of academic expertise, signaling membership into the academic community (Howarth 1998, Li & Schmitt 2010).

The study of academic vocabulary is quickly growing in the linguistic field, especially considering the possibility of creating some lists of the most frequent words used in the academic communicative situations, that easily students can match with in their career, based on lexical academic spoken corpora (just as the one illustrated in the following sections of the article), having the purpose of facilitating their learning capability (Peppoloni, 2012). The present project is directly related to this area of studies and aims to propose a useful and usable tool for the Italian language, at the disposal both of teachers, who can look it up to elaborate their didactic strategies, and of students, who can recur to this linguistic resource depending on their specific needs.

2. WHAT ARE COLLOCATIONS?

The didactic and linguistic relevance of Formulaic Sequences (FSs) emerges already from the studies conducted by Pawley and Syder (1983), Nattinger and DeCarrico (1992) and Lewis (1993). FSs include different sub-groups of word combinations; among them: idioms (*to bite off more than you can chew*), figurative expressions (*a hot potato*), pragmatic formulas (*have a nice day*), discourse markers (*on the other hand*), and collocations (*key aspect*, *low profile*); the latter being the main focus of the present study. Collocations constitute a key component of the communicative competence (Barfield & Gyllstad 2009, Nation 2001, Schmitt 2004, Wood 2010, Wray 2002), since they are tightly linked both to fluency and naturalness in processing a certain language (Pawley & Syder 1983) and to fulfil basic communicative needs (Wray 2002). Mastery of collocations is extremely important not only for native speakers, but for foreign learners too (Granger & Meunier 2008, Lewis 2000, Schmitt 2004).

Collocations are frequently recurring two-to-three word syntagmatic units which can include both lexical and grammatical words, e.g. verb + noun (address an issue), adjective + noun (high probability), noun+noun (survey data) adj+adv constructions (widely dispersed) which play a central role in the academic text in relation to supporting the more technical, i.e. topic specific content of the text (Henriksen, 2013).

The present study draws on previous studies on the establishment of academic language corpora, extraction of academic word and phrase lists and the construct of AV in relation to domain specific and general language use (e.g. Biber 2006; Granger & Paquot 2009; Simpson-Vlach & Ellis 2010; Gardner & Davies 2013; Ackerman & Chen 2014; Hyland 2008).

2.1 The role of collocations in language learning

According to many studies, collocations play a crucial role in language learning, as they seem to be responsible for fluency and native-like selection both in children and in foreign learners (Peters 1983; Wray 1999). Since many psycholinguistic studies demonstrate that human brain is much confident with memorizing than with processing new linguistic items, and that the availability of large numbers of prefabricated units reduces its efforts, making language usage as fluent as possible (Aitchison 1987; Fillmore 1979; Pawley & Syder 2000; Partington 1996), it emerges the need, from teachers and researchers, for explicitly focusing on multi-word expressions as a viaticum for improving native and non-native speakers linguistic competence reducing processing effort. A deviant usage of collocations can compromise the effectiveness of the produced message, drawing the auditor attention away from its meaning (Hüllen 1971; Hecht & Green 1988; Korosadowicz-Struz˙yn´ska 1980;

Cowie & Howarth 1996). The knowledge of and the ability in using collocations are thus essential for academic foreign learners; unfortunately, due to their conventionality and arbitrariness, they pose considerable difficulties, even for the advanced students.

3. DATA AND METHODOLOGY

Enthusiasm for the pedagogic possibilities of corpora has been steadily growing in the last few years. However, as it is familiar from many other linguistic and language teaching enterprises, the first steps have been taken almost exclusively in the written domain. So far the pedagogical applications of speech corpora have received scant attention, with few exeptions (i.e. McCarthy 1998; Zorzi 2001).

One of the general strengths of corpora is that they can show what is typical or common in a language. So, for instance, if the most frequent use of the verb "pensare" (to think) is not with the meaning referring to some ponderous mental process, but rather with the meaning "to have an opinion", this can be seen in a corpus and shown to students by a teacher, or be discovered by students on their own. On the bases of examples like this, we can replace recommendations of language use which are solely based on tradition or teacher intuition. It has become a common finding that what is thought as functional language use is not necessarily in agreement with what is frequent in a language, or even appears at all. Such findings seem to be particularly typical of speech, so that it is not unreasonable to expect corpus data to be helpful in simply providing more relevant information to base pedagogical practices on (McHardy & Sinclair 2004).

The corpus presented in this article can be seen as an opportunity given to students to exercise and improve their lexical fluidity.

3.1 Collecting and structuring corpus data

At the beginning of this first phase of the present research there is the collection of the data that constitute the corpus ASIC (Academic Spoken Italian Corpus) and from which to extract the list of Italian Academic Collocations. In order to develop the corpus, some academic communicative situations and academic courses taught in Italian at the University for Foreigners of Perugia and at the University of Perugia have been audio-recorded. The corpus has been balanced in two ways:

- *horizontally*, dividing the thematic areas of affiliation of the collected data (humanistic (social sciences, literature courses, linguistics)/economic-juridical (economics, law)/scientific (mathematics, physics, medical subjects));

- *vertically*, sharing the different communicative situations considered relevant for the project (frontal lessons, oral exams, seminars and conferences).

In order to process the data through computational applications, the recordings have been orthographically transcribed, using facilitating software, such as NVivo and Dragon Naturally Speaking, that help in automatizing this procedure, allowing researchers to save time especially when working with big amounts of data. After the phase of transcription, data have been structured through the procedure of *encoding*; which includes marking-up and tagging.. After this phase, we dispose of a spoken corpus, statistically and computationally analyzable, made of about 500.000 words.

3.2 Extracting and filtering collocations from the corpus ASEC

Academic Italian Collocations have been retrieved in the corpus sections using the Corpus Query Processor (CQP), which allows the execution of queries in large text corpora with linguistic annotations. CQP is the program language of a platform developed by the University of Stuttgart, called Corpus WorkBench (CWB), in order to automatically investigate and manipulate the data of a corpus, performing several operations such as collocations extraction, elaboration of frequency word lists and so on. The extraction of the collocations with the POS sequence V-N, has been executed using the following query:

```
[pos="VER.*"][pos="ADV.*"]?[pos="ART"]?[pos="ADJ|DET.*|NUM.*"]?[po
s="NOUN"].
```

Once obtained the list of candidate collocations, it has been filtered by removing all the candidates with frequency ≤ than 3, in order to eliminate the factor of fortuity. After this process we dispose of 341 automatically extracted collocations.

3.3 Validation of the data

Of course not all the 341 combinations included in the initial list could be considered as Italian collocations.

How to assess then which of them were or not Italian collocations, without the a priori intervention of the unique linguistic knowledge as a native speaker of the researcher?

A final filtering operation has been foreseen, consisting in the manual exclusion of non-collocations through an experiment of crowd sourcing on Italian native speakers. The choice of this kind of practice allowed the researcher to limit his intervention in deciding which word-combinations can be considered as Academic Italian Oral Collocations, validating then data in an external and objective way.

Crowd sourcing procedures have been usually used in linguistics with the aim of:

- Resources optimization (Snow et al. 2008, Hsueh et al. 2009);
- Translation (Callison-Burch 2009);
- Corpora development (Post et al. 2012, Wang et al. 2012).

The innovativeness of this study relies in the fact that native speakers are not engaged in the process of data collection or definition, but in the phase of data validation, telling to the researcher which collocations have to be included in the list or not. A sort of decisional power has been assigned to them, due to their linguistic competence. Since we intend language as a social and shared process, native speakers, who intuitively know composition, meaning and usage of collocations proper of their mother tongue, are called to select data and to establish what to include or not in the list.

The subjects involved in the experiment are 50 native Italian speakers, at all levels of education and of different ages and sex. Each participant was given a form with an introduction about the aims of the project, how to participate in the experiment and the list of the ten collocations to be evaluated.

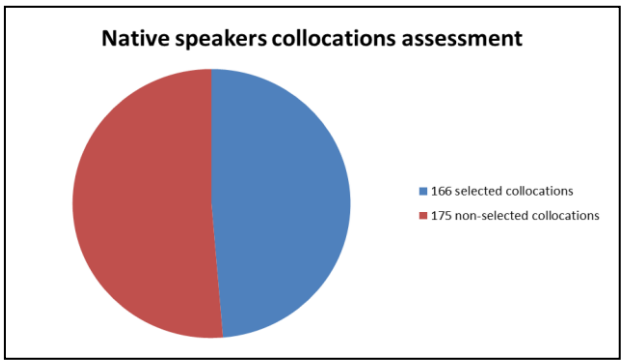
For each of them they had to write down YES or NOT, depending on their positive or negative assessment. If they selected a combination of words as a collocation, then they had to try to define it.

(1) essere parola SÌ (to be word) yes
"essere di parola": essere una persona che mantiene le promesse
(to be someone who keeps promises)

4. RESULTS OF THE CROWD SOURCING EXPERIMENT

Native speakers have evaluated an accuracy of 48,68% of the CQP query, selecting 166 collocations on the 341 given.

Including the query the possibility of inserting many linguistic elements, it probably individuated a very heterogeneous set of data. That is why 175 combinations were discarded.



Some linguistic aspects of the presented word-combinations, may have influenced Native Speakers evaluation:

- it is important to underline that, according to the definitions given by NS, many times each words-combination corresponds to more than one collocation showing different meanings and usages.

(2) **Fare - analisi** (to do - analysis)
Devo fare le analisi (I have to do blood tests)
Fare un'analisi della situazione (To do an analysis of the situation)
Dare – numero (to give - numbers)
dare i numeri (to go crazy)
dare il proprio numero a qualcuno (to give your own telephone number to someone else)

- The higher is the frequency of the words-combination, the higher is the positive assessment of NS. High frequency co-occurrence of two or more words indicates a sort of familiarity in usage because of the strong degree of exposure to this combination;

- Word-combinations containing verbs with a very general meaning (such as *to be*, or *to do*) are difficult to be assessed by NS, that usually are not able to individuate the particular meaning assumed by those verbs in association with other words.

(3) **Essere – sinistra** (to be - left) <u>not selected</u>
Instead it exists the collocation: *essere di sinistra*
(to side with the leftist political party)

- The insertion of other linguistic elements can confuse Native Speakers (NSs)

(4) **Essere – ramo** (to be - branch) <u>not selected</u>
Instead it exists the collocation: *essere del ramo*
(to show expertise in a field)

- Reflexive verbs generate problems for NSs

(5) **Fare – forza** (to do - strength) <u>not selected</u>
Instead it exists the collocation: *farsi forza*
(to be strong)

- If one of the words that constitute the combination, has more meanings, this increases NSs difficulty in individuating it as a collocation

(6) **Essere – piano** (to be – level/flat surface/floor/program) <u>not selected</u>
It exists the collocation: *essere sullo stesso piano*
(to be equal, to have the same weigth)

5. CONCLUSIONS AND FURTHER PERSPECTIVES

The results of the conducted experiments show that even if native speakers do have an intuitive idea of what collocations are and how they have to be used in their mother tongue, they are in difficulty in identifying them, their usage and their meaning, when they have to explicitly reflect on them. Furthermore, the task becomes more and more complex if they only have at their disposal couples of words, without any example of their usage in a linguistic context; in other words, when they have to build by themselves a real-usage example. A possible solution could be that of providing to NSs real attested examples directly extracted from the reference corpus, in order to help them in visualizing how each word combination works in a certain linguistic context, making thus easier to establish if it is or not a collocation.

Another relevant point emerged from this study is that many of the identified collocations could be assigned also to general language and not only to the academic one, since this latter is a sub-group of the first one. A possible solution could be to introduce, in addition to the frequency index, other statistical parameters such as salience and keyness indexes. These two measures allow the researcher to identify keywords in a collection of texts, that is to say those lemma that mostly characterize it, because they show an uncommon frequency for a certain textual genre, such as the academic one.

Other future developments expected to improve the present research are:

- to test different methods of extraction, to compare their results choosing the most convincing and reliable one;

- to increase the size of the corpus ASIC; currently, it is a small-size corpus with its 500.00 entries. This modification could be helpful in eliminating the computational problem of data sparseness;

- to increase the number of NSs involved in the crowd sourcing experiment, in order to guarantee a sufficient representative sample of the population;

- to develop multilingual collocation lists, since these linguistic phenomena constitute a crucial problem in translation, not showing any exact correspondence between one language and another.

References

ACKERMAN, K. AND CHEN, Y., 2014. *The Academic Collocation List*. [online] Available at: http://pearsonpte.com/research/Pages/CollocationList.aspx .

AITCHISON, J., 1987. Reproductive furniture and extinguished professors. In: R. Steele and T. Threadgold, eds. 1987. *Language Topics. Essays in Honour of Michael Halliday*, Amsterdam: Benjamins. Vol. 2 pp.3–14.

ARNAUD, P. J. L. AND SAVIGNON, S. J., 1997. Rare words, complex lexical units and the advanced learner. In: J. Coady and T. Huckin, eds., 1997. *Second language vocabulary acquisition*. Cambridge: Cambridge University Press. pp.157-173

BARFIELD, A. AND GYLLSTAD, H. eds., 2009. *Researching collocations in another language: Multiple interpretations*. Basingstoke: Palgrave Macmillan.

BENSON, M., 1990. Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 31, pp.23-35.

BENSON, M., BENSON, E. AND ILSON, R., 1986. *The BBI Dictionary of English Word Combinations*. Amsterdam: John Benjamins.

BIBER, D., 2006. *University Language: a corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

CALLISON-BURCH, C., 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, August. Association for Computational Linguistics. pp.286–295.

CALZOLARI, N. ET AL., 2002. Towards Best Practice for Multiword Expressions in Computational Lexicons. *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Canary Islands; Spain, 29 May – 31 May 2002. pp.1934-1940.

COWIE, A., 1981. The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, 2, pp.223-235.

COWIE, A. P. AND HOWARTH, P., 1996. Phraseological competence and written proficiency. In: G. M. Blue and R. Mitchell, eds., 1996. *Language and Educaction = British Studies in Applied Linguistics*. Clevedon: Multilingual Matters. pp.80–93.

COXHEAD, A., 2000. A New Academic Word List. *TESOL Quarterly*, 34, pp.213–38.

EVERT, S., 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, IMS, University of Stuttgart.

FILLMORE, C., 1979. On fluency. In: C. Fillmore, D. Kempler and W.S.Y. Wang, eds., 1979. *Individual differences in language ability and language behavior*. New York: Academic, pp.85-101.

GARDNER, D. AND DAVIES, M., 2013 A new academic vocabulary list. *Applied Linguistics*, 35, pp.1-24.

GRANGER, S. AND MEUNIER, F., 2008. *Phraseology. An interdisciplinary perspective*. Amsterdam: John Benjamins.

GRANGER, S. AND PAQUOT M., 2009. In search of a General Academic vocabulary: A corpus-driven study. In: K. Katsampoxaki-Hodgetts, ed., 2009. *Options and Practices of LSP Practitioners*. University of Crete, 7-8 February 2009, University of Crete Publications. pp.94-108.

HECHT, K. AND GREEN, P. S., 1988. KommunkativeWirksamkeit von Schülerbriefen – ein Produkt von Sprachrichtigkeit?. *Englisch*, 1, pp.1–8.

HENRIKSEN, B., 2013. Research on L2 learners' collocational competence and development - a progress report. In: C. Bardel, B. Laufer and C. Lindqvist, eds., 2013. *L2 vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis*. Eurosla Monographs Series 2, EUROSLA. pp.29-56.

HOWARTH, P., 1998. Phraseology and second language proficiency. *Applied Linguistics*, 191, pp.24-44.

HSUEH P., MELVILLE P. AND SINDHWANI V., 2009. Data quality from crowdsourcing: a study of annotation selection criteria. *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. Association for Computational Linguistics. pp 27–35.

HÜLLEN, W., 1995. Englisch als Fremdsprache. In: R. Ahrens, W.-D.Bald and W. Hüllen, eds., 1995. *Handbuch Englisch als Fremdsprache*. Berlin: Schmidt. pp.54–58.

HYLAND, K., 2008. As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, 27, pp.4-21.

KOROSADOWICZ-STRU˙ZY ´NSKA, M., 1980. Word collocations in foreign language vocabulary instruction. *Studia Anglica Posnaniensia*, 12, pp.109–120.

LAUFER, B. AND WALDMAN, T., 2011. Verb-noun collocations in second-language writing: A corpus analysis of learners' English. *Language Learning*, 612, pp.647-672.

LEWIS, M., 1993. *The lexical approach*. Hove: Language Teaching Publications.

LI, J. AND SCHMITT, N., 2010. The development of collocations use in academic texts by advanced L2 learners: a multiple case study approach. In: D. Wood, ed., 2010. *Perspectives on formulaic language: Acquisition and communication*. London/New York: Continuum. pp.23-46.

MCCARTHY, J., 1998. Elaboration tolerance. *Working Papers of the Fourth International Symposium on Logical formalizations of Commonsense Reasoning*, Commonsense-1998.

MCHARDY SINCLAIR, J., 2004. *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

NAGY, W. AND TOWNSEND, D., 2012. Words as Tools: Learning Academic Vocabulary as Language Acquisition. *Reading Research Quarterly*, 471, pp.91-108.

NATION, P., 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

NATTINGER, J. R. AND DECARRICO, J. S., 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

NESSELHAUF, N., 2005. Collocations in a learner corpus. *Studies in Corpus Linguistics*, Vol. 14. Amsterdam: Benjamins.

PALMER, H.E., 1933. *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.

PARTINGTON, A., 1996. *Patterns and Meanings.Using Corpora for English Language Research and Teaching*. Amsterdam and Philadelphia: Benjamins.

PAWLEY, A. AND SYDER, F., 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency. In: J. Richards and R. Schmidt, eds., 1983. *Language and communication*. London: Longman. pp.191-226.

PAWLEY, A. AND SYDER, F., 2000. The One-Clause-at-a-Time Hypothesis. In: H. Riggenbach, ed., 2000. *Perspectives on Fluency*. Ann Arbor: The University of Michigan Press.

PEPPOLONI, D., 2012 Linguistic and computational tools in support of non-native Italian speaking students: the development of the Academic Spoken Italian Corpus. In: A. Llanes, L. Astrid, L. Gallego and R. Mateu, eds., 2012. *La lingüística aplicada en la era de la globalización*. Lleida: Edicions i Publicacions de la Universitat de Lleida.

PETERS, A. M., 1983. *The Units of Language Acquisition*. Cambridge: CUP.

POST M., CALLISON-BURCH C. AND OSBORNE M., 2012. Constructing parallel corpora for six indian languages via crowdsourcing. *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada, June. Association for Computational Linguistics. pp.401 409

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A. AND FLICKINGER, D., 2002. Multiword expressions: A pain in the neck for NLP. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics CICLING 2002*. Mexico City. pp.1–15.

SCHMITT, N., ed., 2004. *Formulaic sequences: Acquisition, processing and use*. Amsterdam: Benjamins.

SIMPSON-VLACH, R. AND NICK C. E., 2010 An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31, 4, pp.463–512.

SINCLAIR, J., 1991. *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

SMADJA, F., 1993. Retrieving collocations form text: Xtract. *Computational Linguistics*, 191, pp.143-177.

SNOW R., O'CONNOR B., JURAFSKY D. AND NG A., 2008 Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*. pp 254–263.

WANG, A., HOANG, C. AND KAN, M. Y., 2012. Perspectives on Crowdsourcing Annotations for Natural Language Processing. *Language Resources and Evaluation*.

WESTBROOK P. and Henriksen, B., 2014: Advanced non-native university lecturers' collocational competence. *Thinking, Doing, Learning: Usage Based Perspectives on Second Language Learning*, 24-26 April, 2013, Odense, Denmark.

WOOD, D., 2010. *Perspectives on formulaic language: Acquisition and communication*. London/New York: Continuum.

WRAY, A., 1999. Formulaic Language. *Learners and native speakers Language Teaching*, 32, pp.213-231.

2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

ZORZI, D., 2001. The Pedagogic Use of Spoken Corpora. In: G. Aston, ed., 2001. *Learning with corpora*. Bologna, CLUEB /Houston, Athelstan. pp.85-107.

PORTUGUESE PROVERBS: TYPES AND VARIANTS

Sónia Reis
University of Algarve
reis.soniamm @gmail.com

Jorge Baptista
University of Algarve
L2F/INESC-ID Lisboa
jbaptis@ualg.pt

Keywords: Proverbs, European Portuguese, Variation, Automatic Identification, Corpus Linguistics

Abstract

Drawing on the methodology and previous results of Rassi *et al.* (2014) on the automatic identification of Brazilian Portuguese proverbs, this paper reports on an extension of that experiment, but now focused on the identification of the European Portuguese proverbs and their variants. Based on a large collection of over 56 thousand Portuguese proverbs and their variants, a database of proverb types was specifically built for natural language processing, along with the finite-state tools that allow for the identification of these strings in texts. Our aim is to make these linguistic resources and language processing tools publicly available, which will undoubtedly be deemed useful assets to other paremiologic studies.

1. INTRODUCTION

Proverbs are an important part of most societies' culture and language. As micro-texts, brought into discourse from the common cultural repository, they are subject to many creative types of variation. On the other hand, functioning as in quotation mode, they integrate discourse in an almost disruptive way, challenging natural language processing (NLP) systems, and requiring their accurate identification and delimitation.

Concerning Portuguese proverbs, and though several, extensive collections of proverbs are available in printed form (Machado, 2011), to the best of our knowledge, no resources have been specifically produced for NLP purposes, even if some digitally available dictionaries (Almeida, 2014) include a few examples, interspersed between other type expressions, like different types of idioms and many forms of slang.

Recently, Rassi *et al.* (2014) have proposed a formal (syntactic) classification of proverbs, based on a collection of over 3,500 proverb variants, organized in 594 proverb types (Rassi, 2014), and taken from several dictionaries from the Brazilian variety of the Portuguese language. The authors presented a finite-state based method for the automatic identification of proverbs in large-sized corpora, and experimented on a 29M tokens corpus of journalistic text (Bruckschein *et al.* 2008), taken from the daily online edition of the Brazilian newspaper *Folha de São Paulo*. The authors report a 60 to 73% precision, depending on the proverb class and the width of the insertion window between the proverbs' keywords. In spite of the corpus size, but not surprisingly, only 137 types and

788 instances were matched, most likely because of the journalistic nature of the texts in this corpus. However, seen from this side of the Atlantic, results from Rassi and colleagues are surprising mostly for the fact that, in spite of some American idiosyncrasies, most proverbs seem to exist also in the European variety, quite unlike the mismatch that has been found for verbal idioms (undisclosed reference).

Drawing on this methodology and these previous results, this paper reports on an extension of that experiment, but now focused on the identification of the European Portuguese proverbs and their variants. This intends to set up the basis for a large collection of Portuguese proverbs and their variants, specifically built for natural language processing, and to make it publicly available, along with the finite-state tools built for retrieving them from texts. These tools and resources will undoubtedly be deemed useful assets to other paremiology studies.

The remainder of this paper is structured as follows: Section 2 presents the methods, starting with the formal classification criteria, and the current state of the collection of Portuguese proverbs' types (§2.1); followed by the criteria to select the proverbs' keywords in order to define those proverbs' types (§2.2); next, the finite-state tools to match proverbs in text are presented (§2.3), followed by the proverbs' collections used to produce a digitized list of Portuguese proverbs and variants (§2.4). Section 3 present some preliminary evaluation of the finite-state tools when applied to this list and discusses some of the issues pertaining to the resulting matches. Finally, Section 4 concludes the paper and hints at future work.

2. METHODS

2.1 Adapting the proverbs classification of Rassi *et al.* (2014) to European Portuguese

Based on the syntactic classification of Brazilian Portuguese proverbs proposed by Rassi *et al.* (2014) we produced a database, in tabular format, with the key elements of each proverb. This formal, taxonomic, approach allows us to determine accurately the concept of variant and base form of a proverb.

In order to better frame this paper, we first present the formal/syntactic classification proposed by Rassi *et al.* (2014) of Brazilian Portuguese proverbs, which may also be adopted by and large to European Portuguese proverbs.

This classification is based on the number of propositions forming the proverb and the part-of-speech (PoS) of their main elements. Other, secondary features are also used. The number of propositions (or clauses) organizes the data in 3 main classes (**P1x**, **P2x** and **P3**). The specific sentence type of **P1x**, or the transformations it may (or may not) undergo are used to further split this type into several classes. Thus, we find the following classes:

P1F1: impersonal constructions; while in European Portuguese this are mostly sentences with the verb *haver* (there be), in Brazilian Portuguese one also finds an impersonal use of *ter* (have); the head noun is often modified by a prepositional phrase (PP) or even a relative subclause; impersonal constructions with indefinite clitic pronoun *se* were also included in this class:

Não há rosa sem espinhos 'There is no rose without thorns'

Não há mal que sempre dure 'There is no evil that lasts forever';

Tem muita estrela pra pouco céu 'There is to many star[s] for little/small sky'

Devagar se vai ao longe 'Slowly one can go far'

P1F2: attributive sentences with copula verb *ser*; the subject is usually a noun (sometimes a verb in the impersonal infinitive); the predicative element can be a noun, an adjective, an verb in the impersonal infinitive, or even a prepositional phrase:

A fome é o melhor tempero 'Hunger is the best seasoning'

O amor é cego 'Love is blind'

Partir é morrer um pouco 'To leave is to die a little'

O silêncio é de ouro 'Silence is of gold'

P1F3: direct-transitive verb constructions:

Muitos cozinheiros estragam a sopa '[Too] many cooks ruin the soup'

Os fins justificam os meios 'The ends justify the means'

Uma mão lava a outra 'One hand washes the other'

P1F4: obligatory negation:

Burro velho não aprende línguas 'Old donkey does not learn languages'

Uma andorinha não faz a primavera 'A single swallow does not make spring'

Gostos não se discutem 'Tastes are not discussed'

P1F5: obligatory fronting of the prepositional complement:

De boas intenções está o inferno cheio 'Hell is full of good intentions'

Em terra de cegos quem tem um olho é rei 'In a land of blind [people], he who has an eye/the one-eyed is king"

Em boca fechada não entra mosca 'In [a] closed mouth no fly enters'

Next, two-clause proverbs are considered (**P2x** classes). These include:

P2F1: proverbs with a main clause and a comparative subordinate clause; as comparison can be expressed in many different ways, different types of comparative structures are considered:

Antes tarde (do) que nunca 'Better late than never'

Mais vale um pássaro na mão do que dois a voar 'Better a bird in the hand that two fling"

Não há pior cego do que aquele que não quer ver 'There is no one more blind that the one who does not want to see'

P2F2: proverbs with two coordinate clauses; in some cases, the coordinative conjunction is not expressed but only implied (asyndeton):

As moscas mudam mas a merda é sempre a mesma 'The flies change but the sheet is always the same'

Vão-se os aneis [*mas, e*] *ficam os dedos* 'The rings go away (but/and) the fingers stay'

Deus não fecha uma porta que não abra logo duas 'God does not close a door without opening two right away'

P2F3: verb-less proverbs with two phrases; in some cases, a coordination or subordination nexus can be inferred but the coordinate/subordinate conjunction is not expressed,

only implied (asyndeton); in other cases, the (implied) main verb of the sentence can be inferred:

Cada roca com seu fuso, cada terra com seu uso 'Each spinning distaff with its spindle, each land with its ways

Muito riso, pouco siso '[Too] much laughter too little judgement'

Tal pai, tal filho 'Like father, like son'

Olho por olho, dente por dente 'Eye for eye, tooth for a tooth'

Cada cabeça [?dá] *sua sentença* 'Each head [gives] its sentence (opinion)'

P2F4: pseudo-interrogative sub-clauses, introduced by interrogative *Qu-* (*Wh-*) pro-forms:

O que não mata engorda 'What doesn't kill [one] makes [one] fat'

Quem avisa amigo é 'He who warns [one] is a friend'

P2F5: proverbs with a main clause and a subordinate clause:

Fazer o bem sem olhar a quem 'To do good without looking to whom'

Devagar que tenho pressa 'Slowly for I am in a hurry'

P2F6: proverbs with a main clause and an obligatorily fronted subordinate clause:

Para morrer basta estar vivo 'In order to die, is enough to be alive'

Enquanto houver vida há esperança 'While there is life, there is hope'

Quando a esmola é muita, o pobre desconfia 'When the charity is too much, the poor gets suspicious'

Finally, class **P3** includes the long proverbs with more than 2 clauses/propositions. Unlike the previous classes, this one has not been subdivided yet, pending on a accumulation of data that would render such sub-classification necessary. Thus, in **P3** we find:

P2F6: proverbs with more than 2 clauses/propositions; the coordinative conjunction can be omitted; often, instead of a clause one finds verb-less phrases:

Mãos frias, coração quente, amor ardente 'Cold hands, hot heart, burning love'

Laranja, de manhã é ouro, à tarde é prata e à noite mata 'Orange, in the morning is gold, in the afternoon is silver and at night it kills'

Um é pouco, dois é bom, três é demais 'One is [too] little, two is good, three is too much'

Following their publication, A. Rassi and her co-authors published the list of 594 proverb types and their classification[44], which we used as the basis for our own classification of European Portuguese proverbs. While, in its general features, their taxonomy seems a useful tool to organize the complex and abundant data already available, in some cases, we disagreed with the authors' classification and decided to assign some of their proverbs to a different class. We also some added Portuguese-specific proverbs and variants, absent from Rassi et al. (2014) list. This careful and close review of the proverbs list, along with the growing number of classified proverbs, will eventually lead to a more granular classification, especially for the larger classes, that is, those with the larger number of types.

[44] https://www.researchgate.net/publication/266852580_Rassi2014?ev=prf_pub.

For instance, the proverbs with the copula verb *ser* (be) constitute such a large set that it could be advantageous to create sub-classes according to the morphological class (PoS) of the element that are associated (adjective, noun, etc.). Table 1 shows the classification in its current state:

Class	Structure	definition	Example/gloss (or wbw translation)	types	%
P1F1	0 V w	impersonal constructions	*Não há rosa sem espinhos.* 'There is no rose without thorns'	**20**	0.03
P1F2	N_0 Vcop (N+Adj)	predicative constr. (copula verb)	*O amor é cego.* 'Love is blind'	**53**	0.09
P1F3	N_0 V N_1	direct transitive (no PP)	*Uma mão lava a outra.* 'One hand washes the other'	**80**	0.13
P1F4	N_0 Neg V w	obligatory negation	*Uma andorinha não faz a primavera.* 'a single sparrow does not makes spring'	**53**	0.09
P1F5	Prep N_1, N_0 V	obligatory fronting of PP	*Pela boca morre o peixe.* 'By the mouth dies the fish'	**45**	0.08
P2F1	F_1 Cs-comp F_2	comparative	*Mais vale um pássaro na mão do que dois a voar.* 'Better a bird in the hand than two flying'	**39**	0.07
P2F2	F_1 (Cc) F_2	coordinate (asyndeton)	Vão-se os aneis, ficam-se os dedos. 'The rings go away, the fingers stay'	**71**	0.12
P2F3	N_1 (Cc) N_2	coordinate (w/o verb)	*Tal pai, tal filho.* 'Like father, like son'	**48**	0.08
P2F4	*Quem V V w*	interrogative subject *Qu-* 'Wh-'	*Quem vê caras não vê corações.* 'Who sees faces does not see hearts'	**90**	0.15
P2F5	F_1 Cs F_2	subordinate	*Fazer o bem sem olhar a quem.* 'To do good without looking to whom'	**20**	0.03
P2F6	Cs F_2, F_1	obligatory fronting of subordinate	*Quando um burro fala, os outros abaixam as orelhas.* 'When a dunkey speaks, the others lower their ears'	**28**	0.05
P3	F_1 C F_2 C F_3	3-clause	*Mãos frias, coração quente, amor ardente.* 'Cold hands, warm heart, burning love'	47	0,08
				594	

Table 1. Classification of Portuguese proverbs (adapted from Rassi *et al.* 2014).

2.2. Compiling the proverbs' keywords

For each class, we defined the key elements that were to be matched in order to unambiguously identify the proverb. As Rassi and co-authors had not published this data (only an example per type and its class are provided), we have re-done most of their work, carefully reviewing the selection criteria that help define the proverb's core elements. These keywords vary depending on the class and, in some cases, even on the proverb itself and its variants. Therefore, only a glimpse of the complex process of selecting the keywords can be provided here.

For keywords, the main content words (nouns, verbs and adjectives) are usually selected, and represented by their lemma (represented inside chevrons, '<' and '>'), rather than the surface (inflected) form, in order to allow for the capture of creative reuse of the proverb. Hence in, for the entry

[**P1F3**] *Os* ***fins*** *(não)* ***justificam*** *os* ***meios*** 'The ends (do not) justify the means'

the lemmas of the two nouns, *fim* and *meio*, and the verb are considered keywords:

<fim> <justificar> <meio>

Usually, copula verbs and auxiliaries are dropped, as in:

[**P3**] ***Um*** *é* ***pouco, dois*** *é* ***bom, três*** *é* ***demais***

'One is too few, two is good, three is too much'

where only the subjects and the adjectives are kept; notice that in this case, the numerals are not determining an noun, as determinants are usually discarded; the specific word order is characteristic of the proverbial nature of the sentence:

<um> <pouco> <dois> <bom> <três> <demais>

In some cases, it is the structure, rather than the specific words, that is key to the proverb. For example, in the proverb:

[**P2F1**] ***Duas cabeças pensam melhor do que uma*** 'two heads think better than one'

we can group the many variants using the following string[45]:

<dois> <N:p> <V:p> (mais+melhor) (do+<E>) que <um:s>

where <N:p> and <V:p> stand for any noun and verb (in the plural), the comparative adverbs *mais* '*more'* and *melhor* 'better' introduce the subordinate comparative conjunction (*do*) *que* 'than', which allows for the zeroing of *do*, and the numeral *um* 'one'.

The number of variants of a proverb can yield quite complex expressions, as in the next case (all variants taken from proverbs' collections; see §2.3, below):

[**P1F4**] ***Não fales*** *de* ***corda*** *em* ***casa*** *de* ***enforcado***

Não *se deve* ***falar*** *em* ***corda*** *em* ***casa*** *de* ***enforcado***

Não *se* ***fala*** *em* ***corda*** *em* ***casa*** *de* ***enforcado****.*

Não *se* ***fale*** *em* ***corda*** *em* ***casa*** *de* ***enforcado***

É ***falar*** *de* ***corda*** *em* ***casa*** *de* ***enforcado***

Em ***casa*** *de* ***enforcado não nomeies*** *o* ***baraço***

Em ***casa*** *de* ***enforcado não*** *se* ***fala*** *de* ***corda***

Em ***casa*** *de* ***ladrão não fales*** *em* ***baraço***

Em ***casa*** *de* ***ladrão não lembrar baraço***

all of them meaning approximately the same: 'Do not speak of rope in a hanged man's house'. Besides the fronting of the prepositional phrase (a locative), in the last for examples, notice the alternation between imperative and the use of the modal auxiliary *dever* 'should' (second sentence), the lexical variation of *corda* 'rope' and *baraço* 'string', the surprising alternation between *enforcado* 'hanged man' and *ladrão* 'thief', and the alternative use of *lembrar* 'remember' instead of *falar* 'speak'. Notice also the (truncated?) form, in the fifth line, introduced by *ser* 'to be', which can be used as a comparison and adapted for commenting on an given event/situation: *Fazer isso é* (*como*) *falar de corda* … 'to do this is like…'. In this more complex case, and because of the changes in word order, the key elements are represented by two strings:

(não+<E>) (<falar>+<lembrar>) (<corda>+<baraço>) <casa>(<enforcado>+<ladrão>)

<casa> (<enforcado>+<ladrão>) (não+<E>) (<falar>+<lembrar>) (<corda>+<baraço>)

Notice that in the case of the negation adverb *não* 'not', this should be treated as a facultative element, since the zeroing of the negation is often one of the strategies to creatively adapt the proverb to new uses. This very phenomenon is illustrated in one of the proverbs' variants.

Thus, a database in tabular format was produced, where the lines contain the proverbs keywords and the number of columns varies according to the class. This matrix, that we call a *lexicon-grammar*, can also be used to include other relevant information. For the moment, this is limited to the proverbs' conventional ID (the code of the class and the proverb number).

[45] We use the notation of the UNITEX system (Paumier 2003, 2014) in representing these regular expressions: elements inside brackets and linked by '+' are interchangeable in that given position; <E> stands for the null string.

2.3. Building finite-state tools for adverb identification

Since the lexicon-grammar of proverbs is in constant update, and can not be directly used to match strings in texts, we used UNITEX[46] linguistic development platform (Paumier 2003, 2014), which is based on finite-state technology, we built the tools that would allows to find, delimit and tag as proverbs candidate strings in running text. In this way, the linguistic information is represented independently of the tools used to apply it to texts. Next we describe this process.

A reference graph was created for each class of proverb, according to its syntactic structure and the number of keywords. Each graph describes the sequences of those key elements of the proverbs. In the graph, variables represent the corresponding cells in the matrix. Then, the reference graph is intersected with the matrix: the system reads each line of the database at a time, replacing the variables for the content of the corresponding cells in the matrix and generates a sub-graph for each proverb. Each sub-graphs is univocally numbered and entire set of is grouped together in a result graph that can be used to match proverbs in texts. To simplify the procedure, for this paper, a single reference graph was produced and a simplified extract is shown in Fig. 1:

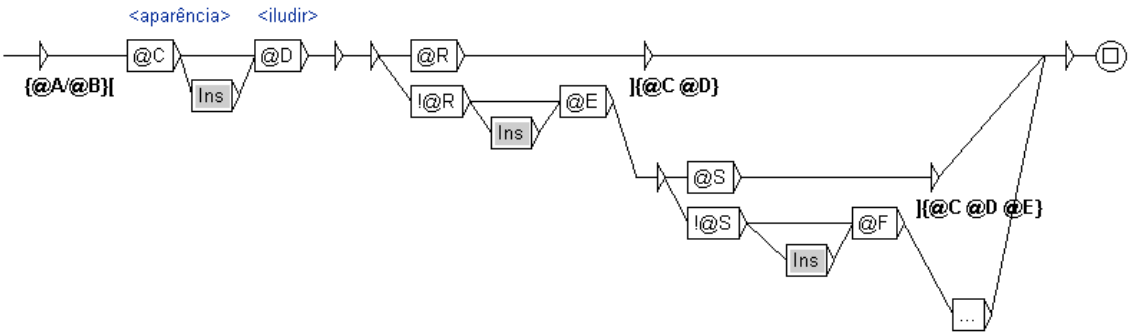


Fig. 1. Reference graph for the proverbs' database.

In this graph, variables @R and @S stand for supplementary or auxiliary columns in the matrix. These columns indicate whether the $n+1^{th}$ column has content (and then this is marked with a plus '-') or not ('+'), the n^{th} column being the last content cell that has been processed; up to 11 columns were used to represent the keywords of the proverbs; in this case, @R stands for the 3rd column, @S for the 4th, and so on; there are always at least two keywords/elements, as illustrated by the proverbial expression *As aparências iludem* 'Appearances deceive'. The system replaces the content variables in the graph (in this case, @C , @D , @E and @F) by their corresponding elements in the matrix. When reaching a variable for the auxiliary columns, the system either continues the path, if this corresponds to a plus sign '+', ending the representation of the proverb; or it moves to the next content variable. The process is repeated until all columns have been explored. Thus, a single reference graph can be used for tables with any given number of columns. The sub-graphs are transducers that can be applied to texts. When matching a given sequence, they insert a right delimiter '**[**' preceded by the proverb ID and the number of keywords (variables @A and @B); and a left delimiter '**]**' followed by the proverb's keywords (variables @C, @D, etc.). An insertion window form 0 up to 3 words is represented by an auxiliary sub-graph **Ins** (shown in a grey box). Fig. 2 shows the sub-graph automatically generated for proverb *Roma e Pavia não se fizeram num dia* 'Rome and Pavia were not made in a single day' [**P1F4**]:

[46] http://www-igm.univ-mlv.fr/~unitex/

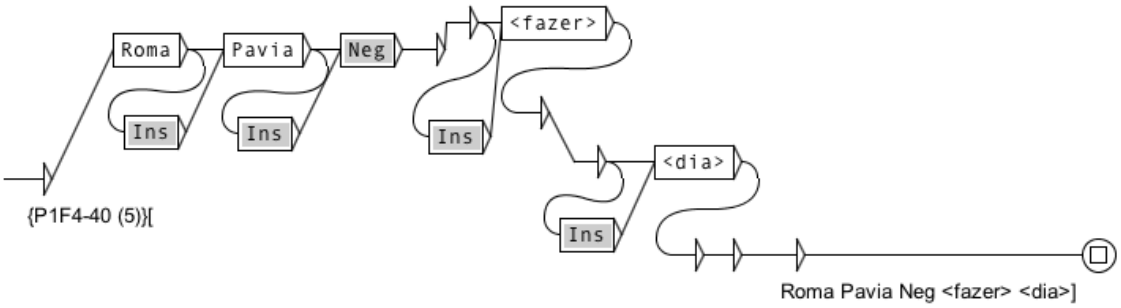


Fig. 2. Sub-graph describing an individual proverb.

2.4. Corpus building

For the automatic identification of proverbs in texts, we digitized a list of approximately 56,150 proverbs (and variants), taken from two large proverb collections:

- Costa (1999): 32,709 entries and 278,217 words;
- Machado (2011): 23,437 entries and 211,257 words.

This list will function not only as a source for the completion of the lexicon-grammar, but also as a testing ground for the finite-state tools built to identify proverbs in texts.

3. PRELIMINARY EVALUATION

The 594 sub-graphs (one for each proverb type), resulting from the process described in the previous section, were applied to the list of 56,150 proverbs (and variants) taken from the two large proverb collections mentioned above. Table 2 shows some preliminary results.

Number of proverbs	**LPP (Costa, 1999)**	**GLP (Machado, 2011)**
Proverbs in each *corpus* (matches)	# 446	# 391
Different proverbs in each *corpus* (types)	# 199	# 194
Proverbs in both *corpora* (matches)	# 347	# 310
Different proverbs in both *corpora* (types)	142	

Table 2. Number of proverbs in corpora

For the 594 proverb types represented in the graphs, only 446 and 391 were captured in each sub-corpus. Notice that this difference may be related to the fact that the list from Machado (2011) is substantially shorter than the one from Costa (1999). Notice also that in spite of the number of matches, the number of different types is much smaller, corresponding to a ratio variants/type of 2.24 in Costa (1999) and 2.02 in Machado (2011). On the other hand, the matched types correspond to about a third of all the types in the

lexicon-grammar. This can be due to several reasons: an incorrect formalization of the data in the matrix; or some problem in the graphs; or the fact that several proverbs are specific of the Brazilian variant. However, it is interesting to notice that in such large collections, so many types were of the lexicon-grammar were not found. Finally, we remark that only 142 types were common to both collections, corresponding to 347 and 310 matches, respectively. This hints at a large non-overlap of the two lists, not due to variation alone, but mostly to lexical coverage.

The most common type of variation in matched proverbs is the left and right extensions of the core elements: while the keywords of the proverb are found, the graphs also capture variants that present some extra material. This is the case, for example with proverb *Não há regra sem exceção* 'The is no rule without exception', from the class P1F1, that produced three different matches (extra material in bold):

Não há regra sem excepção

Em toda a afirmação *não há regra sem excepção*

Não há regra sem excepção, ***nem mulher sem senão***

As one can see, the first line corresponds to the basic form of the proverb, while in the following lines there is a fronted preposition phrase and coordinated clause, respectively.

Variation in vocabulary is also observed. The most affected grammatical classes are nouns and verbs, showing morphological and lexical variation but no differences regarding the proverbs meaning or function. The next two proverbs illustrate these types of variation, respectively (keywords in bold):

O ***fim justifica*** *os* ***meios***

Nem sempre os ***fins justificam*** *os* ***meios***

Os ***fins*** *não* ***justificam*** *os* ***meios***

Burro velho *não* ***aprende letra***

Burro velho *não* ***aprende línguas***

Burro velho *não* ***aprende o caminho***

Representing in the lexicon-grammar most keywords by their lemma already captures morphologic variation. However, in order to capture the lexical variation, since not all variants are yet represented in the lexicon-grammar, a good strategy would be to semi-automatically create new graphs replacing one keyword at a time by its PoS code (e.g. <N> for the nouns, <V> for the verbs, and so on), and then automatically retrieve the variants that were not yet registered in the lexicon-grammar.

4. CONCLUSION AND FUTURE WORK

This paper presented a general framework for the formal (syntactic) classification of proverbs, based on their structure and syntactic properties. The first steps towards the construction of a lexicon-grammar of Portuguese proverbs by building a database of proverb types, consisting of the keywords that univocally identify each proverb, and allowing for some lexical and morphologic variation. Preliminary experiments on the conversion of these lexical matrices into finite-state tools were carried out, which enable to identify, delimit and tag proverbs and their variants in texts. An extensive list of 56,150 proverbs (and variants), taken from two large collections was digitized and will function

not only as a source for the completion of the lexicon-grammar, but also as a testing ground for the finite-state tools built to identify proverbs in texts.

From the obvious contrast between the number of proverbs from each collection and the number of types already represented in the lexicon-grammar, the next first step is to extend the lexical coverage of the matrix. One of the methods to be explored is the semi-automatic construction of new graphs replacing each content keyword (mainly nouns, verbs and adjectives) by the corresponding PosS. As a consequence, the initial classification may need to be refined, particularly in the case of the most productive structures, which may require sub-classification or even the creation of new classes. Attention must be given to devise a reference procedure that will allow relating, in a single lexical-grammatical unit, those variants belonging to different formal classes, often involving changes in word order. A more precise understanding of the variation phenomena, and the construction of reference graphs for each formal class will reflect of a more precise definition of the insertion window and the sentence alternations that proverbs often allow. It is also a future objective, once a satisfactory lexical coverage is achieved, to measure the frequency of the proverbs types (and their variants) in real corpora, in order to associate frequency information to the database.

Acknowledgements

Research for this paper was partially funded by Portuguese national funds through Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2014.

References

ALMEIDA, J.J. 2014. *Dicionário aberto de calão e expressões idiomáticas*. [online] Available at: <http://natura.di.uminho.pt/~jj/pln/calao/dicionario.pdf> [Accessed 14 March 2015].

BRUCKSCHEIN, M., MUNIZ, F., SOUZA, J., FUCHS, J.T., INFANTE, K., GONÇALEZ, P.N., VIEIRA, R. and ALUISIO, S.M. 2008. *Anotação linguística em XML do corpus PLN-BR. Série de Relatórios do NILC*, São Carlos (SP): NILC-ICMC-USP.

COSTA, J. R. M. 1999. *O Livro dos Provérbios Portugueses*, Lisboa: Editorial Presença.

MACHADO, J.P., 2011. *O Grande Livro dos Provérbios*, 4ª ed., Lisboa: Casa das Letras.

RASSI, A.P., 2014. *List of Proverbs in Brazilian Portuguese*. <https://www.researchgate.net/publication/269165152_Rassi2014> [Accessed 14 March 2015]. DOI: 10.13140/2.1.4907.7280

RASSI, A.P., BAPTISTA, J. AND VALE, O. 2014. Automatic Detection of Proverbs and their Variants. In: M. Pereira, J. Leal, J. and A. Simões, eds. *Proceedings of the Symposium on Languages, Applications and Technologies* (SLATE`14). Leibniz (Germany): Schloss Dagstuhl - Leibniz-Zentrum fur Informatik, Dagstuhl Publishing. pp. 235-249.

ROCHA, P. AND SANTOS, D. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Nunes, M.G. *et al.*, eds., *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, São Paulo: ICMC/USP. pp. 131–140.

IDENTIFICATION D'UNITES PHRASEOLOGIQUES ET EQUIVALENCE SEMANTIQUE DANS LA TRADUCTION

Dorota Sikora
Université du Littoral – Côte d'Opale
& HLLI (EA 4030)
Dorota.Sikora@univ-littoral.fr

Abstract

Although lexical resources tend to offer more and more extended coverage of multi-word expressions, the available machine translation programs still encounter difficulties in dealing with phraseological units. In some contexts, sentences they generate are infelicitous, or even grammatically incorrect. Such translation problems rise questions of the ways online dictionaries describe multi-word expressions. Considering the case of four French and English strong idioms (fully non-compositionnal phrasemes, cf. Mel'čuk, 2013), the present paper discusses the encoding of phraseological units in two lexical databases : the *French Lexical Network* and the *English Lexical Network* built in the same theoretical and methodological frame, as implementations of a graph model named lexical system (Polguère, 2009, 2014). Since each of these resources is basically monolingual, multilingual connexions between two or more lexical systems can be implemented. Beyond the particular case of the idioms the present study focuses, we suggest that a homogeneous and coherent treatment of multi-word expressions in lexical resources underlying machine translation programs can significantly improve rendering of idiomaticity in the target language.

1. INTRODUCTION

Le présent article propose une réflexion sur ce qui – selon notre hypothèse – est l'une des principales sources d'erreurs dans la traduction automatique d'unités phraséologiques (désormais UP) : un diagnostic imprécis de leur statut linguistique. Nous présenterons le traitement des locutions fortes (ou phrasèmes complets, cf. Mel'čuk, 2013 : 134) dans deux bases de données lexicales développées actuellement à l'Atilf (CNRS UMR 7118) : *Réseau Lexical du Français* (*RL-fr*, cf. Gader et al., 2012, Polguère, 2014), et *English Lexical Network* (*en-LN*, Gader et al., 2014). Ces ressources sont construites en parallèle, selon les principes théoriques et méthodologiques homogènes, définis initialement dans le cadre de la Lexicologie Explicative et Combinatoire et développés récemment par Polguère (2009, 2014) dans un modèle de *système lexical* (cf. section 2).

Si l'intérêt des lexicographes pour l'intégration des UP dans les ressources dictionnairiques est aujourd'hui manifeste (cf. Osherson et Fellbaum, 2010, Fellbaum,

2014), les utilisateurs n'en constatent pas moins de problèmes dans les résultats obtenus. Trois types de difficultés sont régulièrement repérés.

Premièrement, dans un contexte monolingue, la couverture que proposent les dictionnaires reste faible (cf. Vaguer, 2010). Elle est de plus hétérogène : même les locutions les plus fréquentes ne s'y retrouvent pas systématiquement. Dans une perspective bilingue, cela signifie qu'il est difficile de trouver l'équivalent d'une UP de la langue L_1 dans une ressource de la langue L_2.

Deuxièmement, il arrive souvent que tout en préservant l'équivalence sémantique entre la phrase source en L_1 et la traduction proposée en L_2, l'idiomaticité de l'orginal se perd, comme danse l'exemple (1b) :

(1) a. Cynthia et Ted étaient aux anges. (Frantext)

b. Cynthia and Ted were ecstatic. (https://translate.google.fr)

D'un point de vue communicatif, il s'agit certes d'un moindre mal, même si comme l'observe Svensén (2009 : 156), « [i]dioms in the source language must as far as possible be paralleled in the target language by idioms with the same content ».

Un troisième cas de figure est celui où une même ressource produit des résultats corrects dans certains cotextes, alors que dans d'autres, elle restitue des phrases inacceptables. L'exemple (1c) montre que Reverso établit correctement l'équivalence entre le syntagme *être aux anges* dans (1a) et le phrasème anglais *to be in seventh heaven*. Cependant, la phrase (2a) pose un problème d'analyse aux deux utilitaires.

(2) a. Tu as une voix qui met aux anges... (FrWac)

b. You have a voice(vote) which puts *?*in the angels... (traduction Reverso, 12 juin 2015)

c. You have a voice that *?*makes angels… (traduction Google, 12 juin 2015)

Notre étude se concentre sur ce dernier type de problèmes. Pour identifier leur source, observons la figure (1) qui reproduit l'information fournie par Reverso à ses utilisateurs : *être aux anges* est encodé comme une locution verbale, équivalente de la collocation *be extremely happy*[47], considérée à son tour comme quasi-synonyme de plusieurs UP en L_2. Selon notre hypothèse, l'origine des erreurs constatées en (2b) et en (2c) se trouve dans cet encodage non conforme à la réalité linguistique.

Dans la section 2, nous présenterons le modèle de *système lexical* (cf. Polguère, 2009, 2014), ainsi que le mode d'encodage et la structuration d'informations dans le *RL-fr* et dans le *en-LN*, ressources lexicales qui sont des implémentations de ce modèle. La section 3, consacrée à l'analyse d'un corpus d'exemples français et anglais qui attestent des caractéristiques de quatre phrasèmes : *être aux anges*, *être au septième ciel* en français, *to be over the moon*, *to be in seventh heaven*. Notre objectif sera de bien distinguer la locution elle-même et les collocations qu'elle peut former. Les conclusions proposées en 4 porteront sur les possibilités d'interconnexion de systèmes lexicaux.

[47] L'amalgame évident entre la définition proposée (la collocation *be extremely happy about smth*) et une série de quasi-synonymes en L_2 (*be walking on air*, etc.) présentés comme des locutions verbales n'est certainement pas étranger aux problèmes de traduction constatés en (2b).

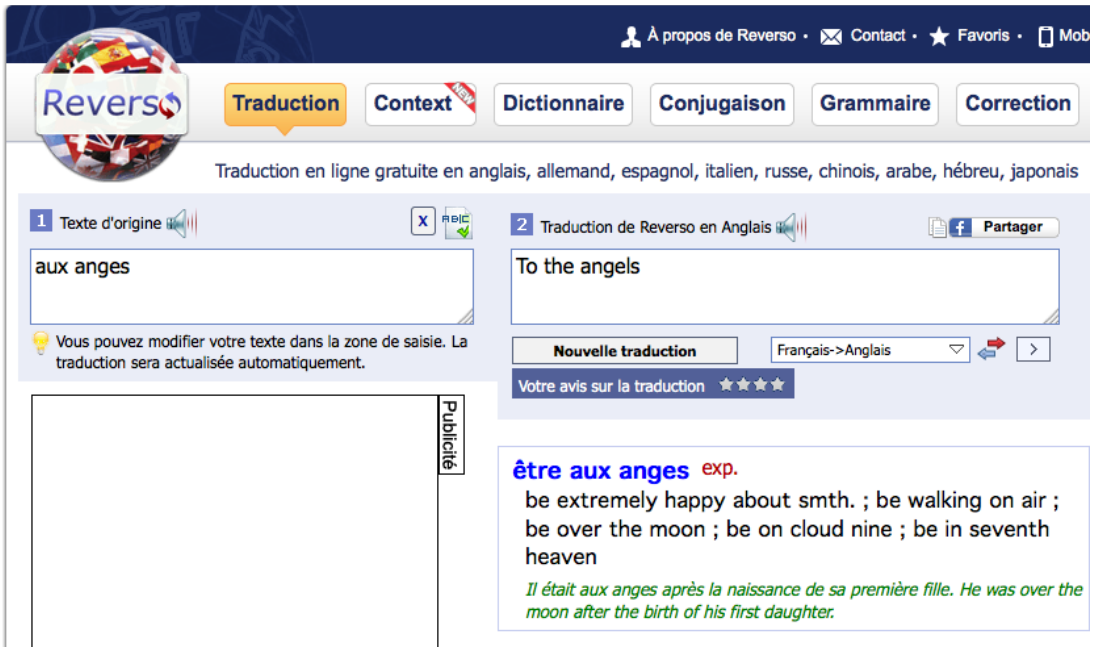


Figure 4. Traduction du syntagme *aux anges* et description lexicographique de *être aux anges* proposées par Reverso (18/03/2015)

2. TROIS RESSOURCES, DEUX SYSTEMES LEXICAUX

Malgré l'homogénéité méthodologique qui les caractérise, les deux ressources ont des histoires différentes. Le *RL-fr* a démarré en juin 2011, avec une nomenclature dite d'amorçage de 3734 vocables lexémiques, compilée à partir de ressources pédagogiques (cf. Polguère et Sikora, 2013), pour couvrir le vocabulaire de base (McCarthy, 1999) du français contemporain. Depuis, sa croissance se poursuit selon les règles décrites dans plusieurs publications (cf. Lux-Pogodalla et Polguère, 2011, Polguère et Sikora, 2013). Les UPs, notamment les locutions fortes, qui ne figuraient pas dans la nomenclature d'amorçage, sont systématiquement décrites par les lexicographes.

La base *en-LN* a été créée en 2012 par conversion des données de WordNet. Pour une présentation détaillée de cette opération, on pourra se reporter à Gader et *al.* (2014), mais il est important de souligner qu'elle a consisté dans une transformation d'un modèle ontologique du lexique reposant sur une hiérarchisation conceptuelle des lexies en un système non ontologique (i. e. en un système lexical, cf. *infra*), organisé par une multitude de relations de différente nature (paradigmatiques, syntagmatiques, de copolysémie entre les lexies d'un vocable polysémique et d'inclusion formelle entre les UPs et les lexies qui les composent). Techniquement, cela signifie qu'un graphe lexical dont les nœuds sont formés par des synsets, ensembles de lexies avec une composante conceptuelle commune (WordNet), devait être converti en un graphe connectant des lexies, avec des éléments descriptifs de nature purement linguistique associés à chacune d'elles. Les figures 2 et 3 représentent respectivement la locution *seventh heaven* dans WordNet et son entrée dans *en-LN* issue de l'injection.

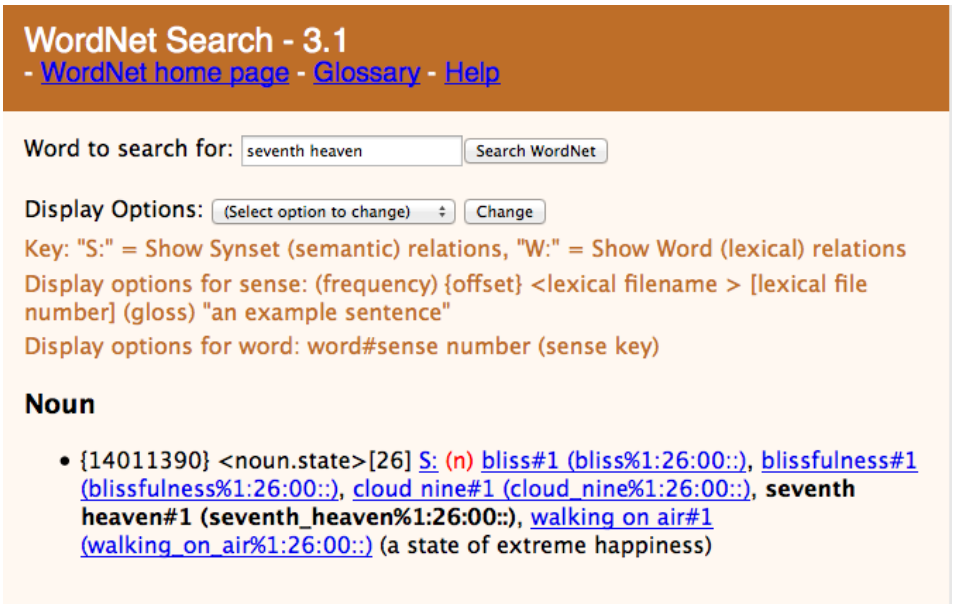


Figure 5. Locution *seventh heaven* dans WordNet.

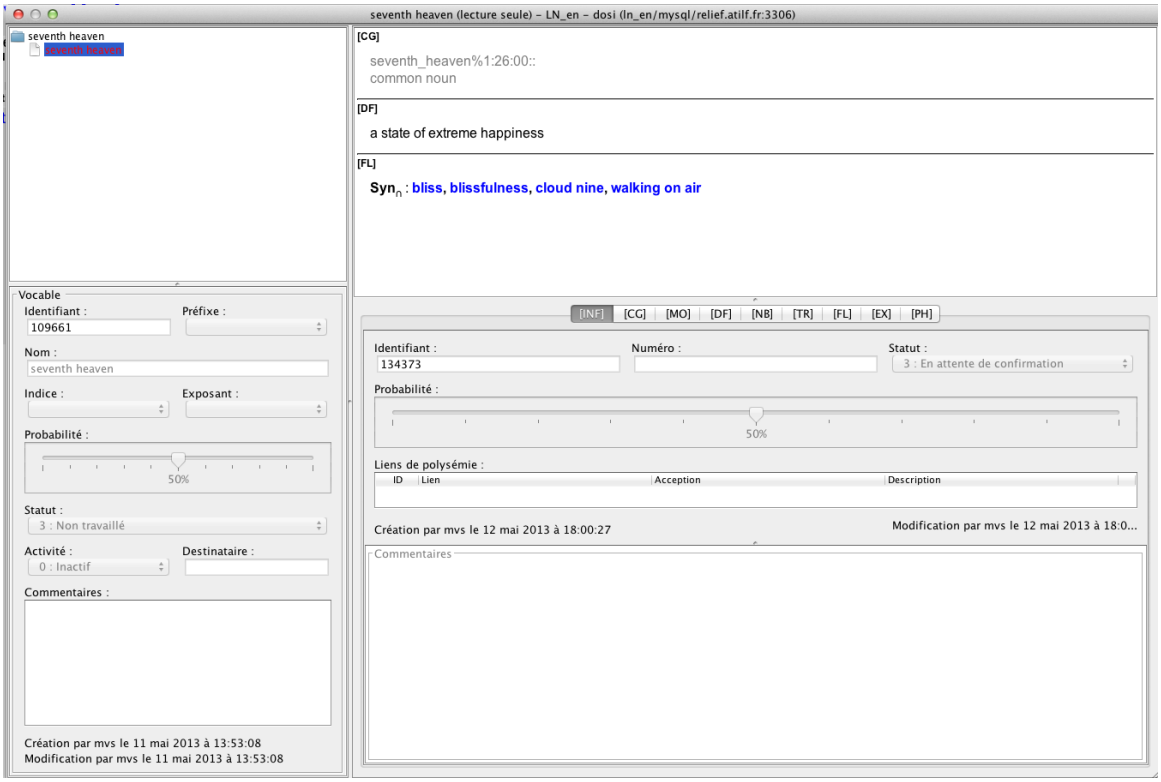


Figure 6 : Locution *seventh heaven* injectée automatiquement dans *en-LN* par conversion des données de WordNet.

Un système lexical tel qu'il est proposé par Polguère (2014) est une modélisation du lexique sous forme de réseau. C'est un graphe au sens mathématique, fait de sommets que constituent les unités lexicales (lexies) et des arcs représentant des relations lexicales. Cette structure est une première des quatre caractéristiques des systèmes lexicaux. Deuxièmement, il s'agit – rappelons-le – d'un modèle non ontologique, qui ne repose sur aucune classification hiérarchique des concepts, ni de connaissances. Le réseau est tissé par encodage de multiples relations lexicales sans que celles-ci soient hiérarchisées.

La structure interne d'un nœud lexical est non-atomique : elle encapsule un ensemble d'informations sur la lexie vedette, allant des indications générales [l'onglet INF] (identifiant de la lexie, état d'avancement de la description, ainsi que – le cas échéant – la nature du lien de copolysémie qui la relie à un ou plusieurs copolysèmes du même vocable), par les caractéristiques grammaticales [CG], flexionnelles [MO] jusqu'aux exemples [EX]. La description sémantique dans la rubrique [DF] comprend une étiquette sémantique (cf. Polguère, 2011), une spécification d'actants sémantiques dans la forme propositionnelle et la définition structurée en composante centrale et en composantes périphériques. Enfin, les relations sémantico-syntaxiques, c'est-à-dire les fonctions lexicales, qui structurent le système lexical et qui en forme l'ossature, sont visibles dans la rubrique [FL]. Les exemples illustrent le sens et l'emploi de la lexie vedette, alors que sous [PH] sont listés – pour chaque unité lexémique – les différents types de phrasèmes dans lesquels celle-ci se trouve incluse.

Un système lexical est de nature relativiste, ce qui permet de rendre compte d'un certain flou dans l'information dont dispose le locuteur et – souvent – le lexicographe. C'est la raison pour laquelle chaque élément de la description sémantico-syntaxique est identifié avec un taux de probabilité qui va de 60 % à 100 %, lorsqu'il s'agit d'un travail effectué « manuellement »[48].

Un système lexical est une modélisation du lexique d'une langue donnée : il doit rendre compte de sa structure et de son fonctionnement, sans s'orienter d'office vers des correspondances interlinguistiques. Ainsi, le *RL-fr* et *en-LN* sont implémentés de manière autonome, en tant que modèles monolingues, leur interconnexion étant cependant envisageable.

3. STATUT LINGUISTIQUE DE LA LOCUTION ET CONSTRUCTION DU NŒUD LEXICAL

Les exemples (1) et (2) présentés dans l'Introduction suggèrent que la traduction se déroule sans encombre, lorsque le phrasème est un syntagme gouverné par le verbe *être*, alors que d'autres cotextes – les verbes attributifs (*sembler*, *paraître*, *avoir l'air*[49]) et le causatif (*mettre*) – posent des difficultés[50]. *Septième ciel* apparaît dans des configurations diverses : l'individu X peut, certes, *être au septième ciel*, mais il n'est pas rare de trouver également *vivre au septième ciel* ou *atteindre le septième ciel*.

Faut-il dès lors admettre plusieurs locutions verbales, plus ou moins synonymes, contenant la séquence *aux anges* et *septième ciel*, en implémentant autant de nœuds lexicaux du graphe ? Pour le vérifier, nous avons observé, dans la section 3.1, le degré d'autonomie qui caractérise les syntagmes prépositionnels *aux anges* et *au septième ciel*, en nous intéressant aux emplois sans le verbe *être* et sans la préposition *à*. Dans 3.2, nous procédons de façon analogue pour examiner deux de leurs équivalents anglais *to be over the moon* et *to be in seventh heaven*.

[48] Toute information injectée automatiquement se voit associer un taux de probabilité de 50 %.

[49] Notons que *avoir l'air* est une collocation dont *air* est la base et *avoir* le verbe support collocatif (cf. *Elle a l'air aux anges -> Son air aux anges*).

[50] Globalement, les résultats obtenus par Google translator sont meilleurs que ceux retournés par Reverso, dans la mesure où le premier de ces outils propose des traductions adéquates, mais qui ont perdu l'idiomaticité de l'orginal.

3.1 (Être) aux anges et (être) (au) septième ciel dans le RL-fr

3.1.1 Locution prépositionnelle aux anges

Le syntagme prépositionnel *aux anges* apparaît fréquemment avec des verbes attributifs (exemple 3 ci-dessous) ou bien rattaché directement à un gouverneur nominal, sans support verbal (exemple 4). Son fonctionnement syntaxique est ainsi caractéristique des adjectifs.

(3) Le couple semblait aux anges mais une consigne très précise avait été donnée à tous les journalistes présents : aucune question sur les rumeurs de grossesse. (FrWac)

(4) Assez fière du repas d'anniversaire 2005, avec des petits plats équilibrés et une famille aux anges, et surprise. (FrWac)

De même, il s'emploie régulièrement en position détachée à gauche (exemple 5), comme à droite (exemple 6).

(5) Aux anges, Géraldine berçait ce poupon mou de paroles encourageantes, expliquait comme il est plus avantageux, question vanité, d'être un peu plaqué que très cocu. (Frantext)

(6) Lambert écoutait, aux anges. (Frantext)

D'autres propriétés transformationnelles mettent en évidence la nature adjectivale de ce syntagme. Il peut en effet être modifié par des adverbes (7) et s'inscrire dans des coordinations avec d'autres adjectifs (8).

(7) « Ma petite fille Azura Sienna est née la semaine dernière ! Je suis euphorique et complètement aux anges », a-t-elle posté sur le réseau social. (Internet)

(8) Amélie N. (Hollywood Girls 4), amoureuse et aux anges, elle dévoile sa nouvelle conquête. (Internet)

Plutôt que de traiter *être aux anges* comme une locution verbale, il convient d'y voir une collocation construite sur la base *aux anges*, syntagme prépositionnel qui forme une locution à caractère adjectival. L'identification de la partie du discours permet de situer la locution dans un ensemble de relations paradigmatiques[51]. *Aux anges* est lié par des liens de quasi-synonymie avec, entre autres, les lexèmes *ravi*, *enchanté*, *heureux*, ainsi qu'à *sur un nuage*. *Aux anges* peut en outre commuter avec ses quasi-synonymes, en partageant leurs propriétés distributionnelles : tout comme ces adjectifs, il forme des collocations avec *être* et avec des verbes attributifs. Si cette locution s'emploie avec le verbe *se sentir*, son causatif (cf. exemple 2a dans l'Introduction) n'est pas celui que l'on utilise avec les autres adjectifs de sentiments. Les collocations formées avec *mettre* conduisent à penser que le sens de *aux anges* est mieux caractérisé par l'étiquette sémantique 'qui est dans un certain état psychique' que par 'qui éprouve un sentiment'. La figure 4 visualise l'article lexicographique de notre locution, *i.e.* le nœud lexical qu'il forme.

[51] On notera à ce propos une différence notable qui oppose le traitement des locutions dans un système lexical à celui que propose WordNet (cf. Osherson et Fellbaum, 2010).

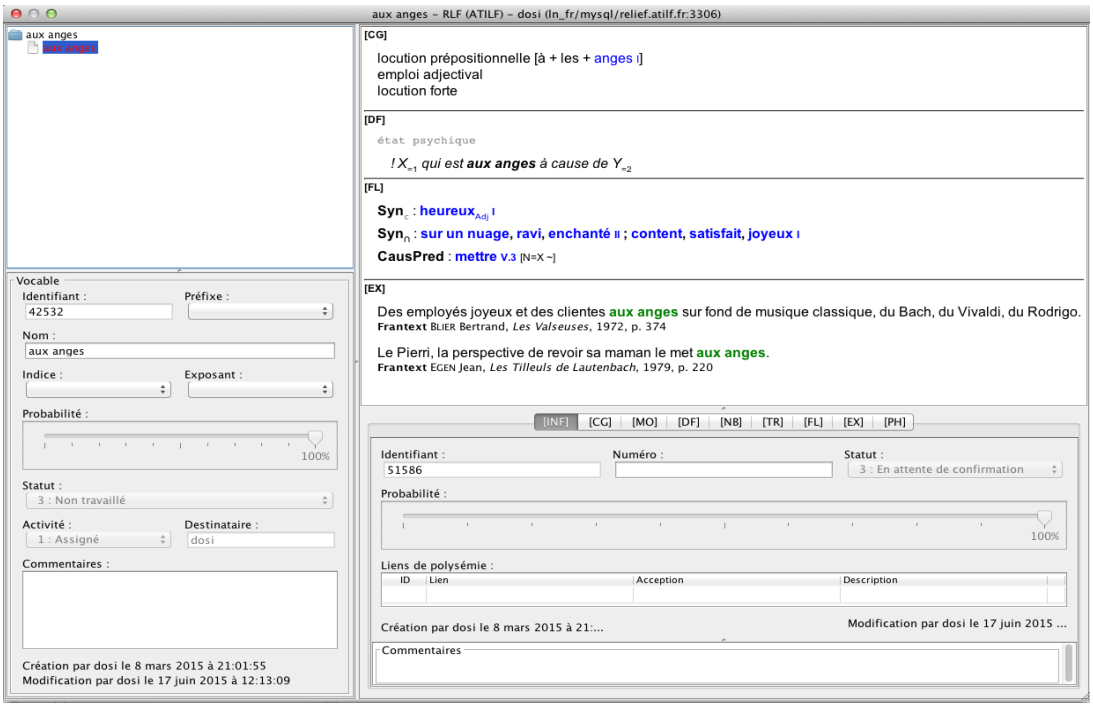


Figure 7. Locution *aux anges* dans la base *RL-fr* (vue d'article).

3.1.2 (Être) (au) septième ciel : locution nominale enchâssée

Le jeu de parenthèses dans l'intitulé de cette sous-section renvoie aux trois types d'emplois attestés dans les corpus. *Septième ciel* s'utilise en effet en syntagme verbal avec le verbe *être* ou bien avec des verbes attributifs tels que *paraître* dans (9).

(9) Alors que M^lle^ Lavoix souriait, paraissait au septième ciel, voulant ajouter à son ravissement, j'employai quelques épithètes flatteuses. (Sabatier, R. 1997. *Le Lit de la merveille*)

Le syntagme prépositionnel *au septième ciel* fonctionne de façon adjectivale. On le retrouve en position détachée à gauche (10) et à droite (11), ainsi qu'en coordination avec des adjectifs : dans (12), nous l'avons substitué à *aux anges* de l'exemple (8). En revanche, il se montre réfractaire à la modification adverbiale : nous n'en avons trouvé aucune attestation et l'insertion d'un modificateur dans l'exemple (9) renuméroté en (13) aboutit à une phrase difficilement acceptable.

(10) Au septième ciel, ma fille, ravie, elle veut se précipiter. (Frantext)

(11) Ils regardaient ses gestes fiévreux, ses yeux luisants, sa pose... trop penché en avant... perdant la tête, oubliant où il est, se croyant seul avec elle dans les limbes, au septième ciel... (Frantext)

(12) Amélie N., amoureuse et au septième ciel, elle dévoile sa nouvelle conquête.

(13) *?* Alors que M^lle^ Lavoix souriait, paraissait complètement au septième ciel, voulant ajouter à son ravissement, j'employai quelques épithètes flatteuses.

Contrairement à *aux anges* cependant, les corpus attestent de nombreux emplois purement nominaux de *septième ciel*. Les exemples (14) et (15) en offrent une illustration :

(14) Pour rejoindre votre septième ciel, *Épices & Vous* vous propose de faire entrer l'Asie dans votre cuisine... (FrWac)

(15) Les partiels, c'est le septième ciel. (Internet)

La structure interne du nœud que forme la locution *septième ciel* sera donc celle d'une lexie nominale, insérée dans un réseau de relations de quasi-synonymie avec *bonheur*, *plaisir*, *bien-être*, *contentement*, *ravissement*, *enchantement*, *félicité*, etc. Notons cependant que, contrairement à la plupart de ses quasi-synonymes, *septième ciel* n'est pas un sentiment à proprement parler : s'il est possible de parler de *sentiment de bonheur*, de *sentiment de plaisir* ou de *sentiment de contentement*, *sentiment de septième ciel* semble difficile à accepter. Le sens de cette locution est à classer plutôt parmi les états psychiques.

Le syntagme verbal *être au septième ciel* est ainsi un exemple de double enchâssement. Pour former une collocation à valeur adjectivale, *septième ciel* sélectionne la préposition *à*. Or, à l'image des adjectifs simples, les collocations à valeur adjectivale telles que *au septième ciel*, *au désespoir de*, *en colère* se combinent avec *être* et avec les verbes attributifs. Dans la mesure où il s'agit d'une caractéristique commune aux adjectif, aucune mention n'en est fait dans l'entrée lexicographique du nœud lexical.

La figure (6) représente le nœud *septième ciel* dans le système lexical *RL-Fr*.

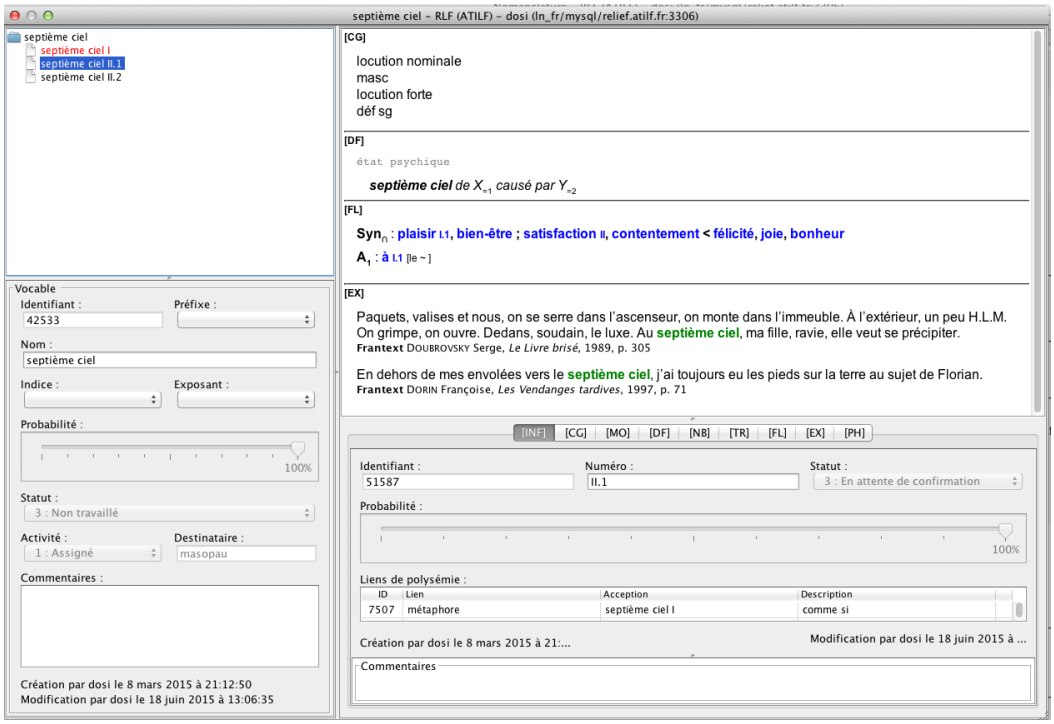


Figure 8. Locution *septième ciel* dans la base *RL-Fr* (vue d'article).

3.1.3 Aux anges et *septième ciel* dans le système lexical du français

Le sous-graphe du système lexical français organisé autour des nœuds locutionnels *aux anges* et *septième ciel* (figure 6) montre que leur interconnexion est indirecte : elle se fait par un lien de quasi-synonymie avec le nom *contentement*, lié sur le plan paradigmatique à la fois à *septième ciel* et à l'adjectif *content*, quasi-synonyme de *aux anges*.

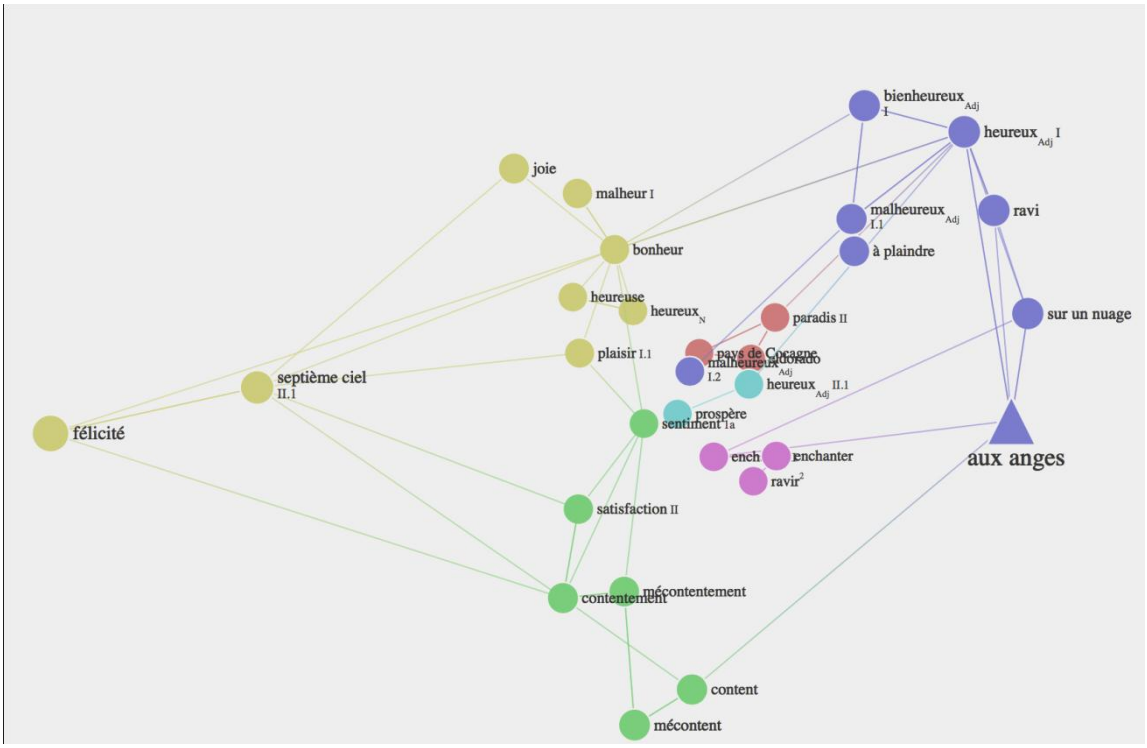


Figure 9. Interconnexion des nœuds lexicaux *aux anges* et *septième ciel* dans *RL-Fr*.

3.2 (To be) over the moon et (to be) (in) seventh heaven dans en-LN

3.2.1 (To be) over the moon dans le système lexical *en-LN*

Le traitement de *(to be) over the moon* dans les ressources lexicales de l'anglais n'est pas homogène. Certaines d'entre elles[52] lui accordent un statut de locution verbale, syntagme gouverné par *to be*, alors que d'autres[53] décrivent le phrasème *over the moon*, c'est-à-dire un syntagme prépositionnel de forme *Prep Det N*.

Comme *(to be) over the moon* ne figure pas dans WordNet, cette locution n'a pas été injectée automatiquement dans *en-LN*. Le nœud lexical a été créé et situé dans le réseau par les lexicographes au terme d'une analyse linguistique dont l'objectif premier était de déterminer le statut lexical – locutionnel ou collocationnel – du phrasème.

En effet, si *over the moon* apparaît fréquemment en position de complément du verbe *to be*, il n'en manifeste pas moins de propriétés adjectivales. Dans la phrase (16), il se rattache directement à son support nominal, alors que (17) et (18) illustrent des emplois détachés, respectivement à gauche et à droite. Malgré une ponctuation qui en fait un énoncé à part dans (18), *over the moon* y qualifie anaphoriquement l'antécédent *mum* de la phrase précédente.

(16) Gloria Taylor's mom over the moon (Internet).

[52] Voir, à titre d'exemple, Cambridge Dictionaries Online (dictionary.cambridge.org/dictionary/british/be-over-the-moon), Reverso, etc. Cette locution n'est pas intégrée dans WordNet, ni en tant qu'une forme verbale, ni en tant qu'une forme adjectivale).

[53] C'est le cas du dictionnaire Collins (www.collinsdictionary.com/dictionar/english).

(17) Although over the moon with it generally, I am disappointed with the lack of power in the naturally aspirated diesel engine. (COCA)

(18) He's fast asleep, so is mum! Over the moon. (Internet)

La possibilité de modification adverbiale dont témoigne l'exemple (19) et de coordination avec d'autres adjectifs (*humbled*, *happy* dans 20) pointe également vers un statut adjectival de la locution prépositionnelle *over the moon*.

(19) Absolutely over the moon with this job! (BNC)

(20) As the first from Great Britain to win a world title, I'm just humbled and happy, and over the moon. (Internet)

Le nœud *over the moon* du réseau *en-LN* (figure 7 ci-dessous) contient donc un ensemble d'informations propres aux adjectifs. La possibilité de sélectionner le collocatif *to be* fait partie des propriétés des lexies adjectivales, on peut donc se demander s'il est indispensable de l'indiquer dans la zone de fonctions lexicales. Une telle démarche semble en effet superflue dans le cas des nœuds monolexémiques quasi-synonymes *ecstatic*, *delighted*, *happy*, *thrilled*, *overjoyed* : en tant qu'adjectifs, ils s'emploient régulièrement avec le collocatif *to be*, tout comme les locutions *on the top of the world*, *on cloud nine*.

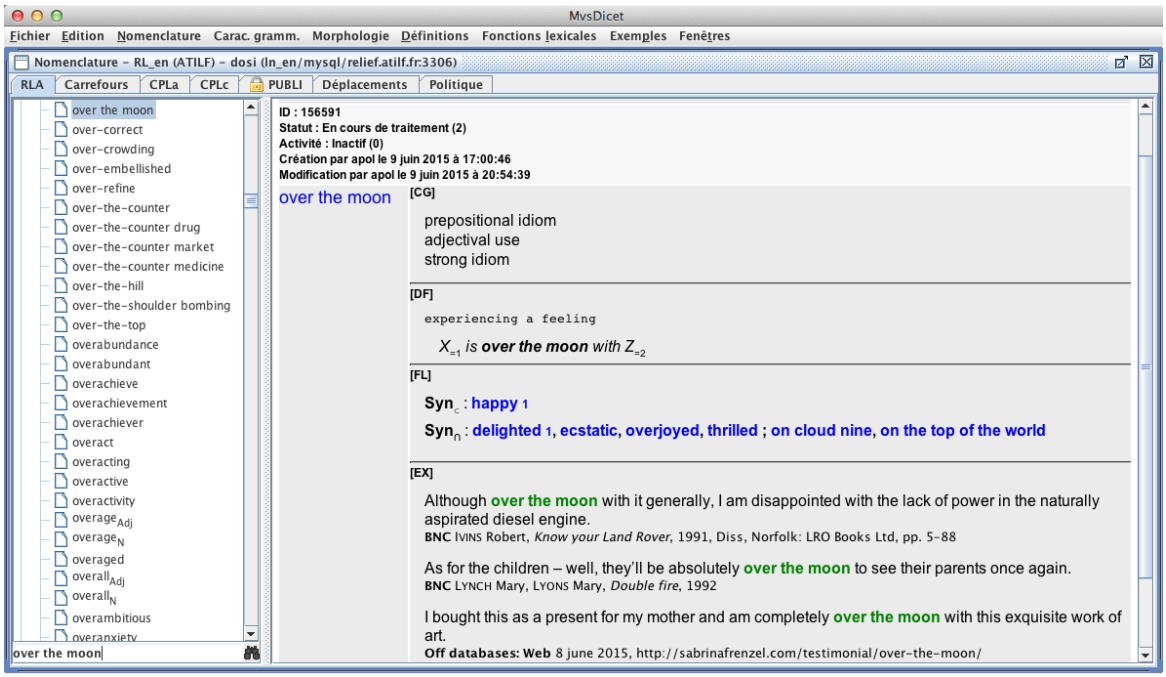


Figure 10. Locution *over the moon* dans la base *en-LN* (vue d'article).

3.2.2 (To be) (in) seventh heaven : une locution nominale enchâssée

Conformément à l'analyse effectuée par les lexicographes de WordNet, *seventh heaven* est une locution nominale et c'est en tant que telle qu'elle a été injectée dans le *en-LN*. À la conversion, elle s'est trouvée extraite de son synset (cf. figure 2 dans la section 2) pour former un nœud dans le système lexical de l'anglais. Le résultat de cette opération a

néanmoins nécessité une vérification manuelle permettant de s'assurer qu'il s'agit bien d'un nœud nominal.

Les corpus attestent des emplois sans le verbe *to be*, tels que l'exemple (21).

(21) It's funny to see a guy in seventh heaven because of a girl. (Internet)

Les phrases (22) et (23) montrent cependant que la préposition *in*, qui autorise un emploi adjectival, n'est nullement obligatoire. Les collocatifs *some form of* ~ et *my state of* ~ indiquent qu'en tant que locution nominale, *seventh heaven* dénote une entité : un état psychique qui caractérise l'individu X (instancié dans l'exemple 23 par le possessif *my*) en réaction à une situation Y.

(22) The more he wished time away, the more slowly it passed although when, at last, the ship returned to Saint-Iean-de-Luz, life assumed the aura of some form of seventh heaven. (Pike, Richard, 2006, *Seven Seas, Nine Lives. A Biography of Captain A. W. F. Sutton*, p. 144)

(23) Sarah shook her head ruefully. 'In my state of seventh heaven it's all going over the top of my head.' She sobered suddenly. 'I'm afraid I was a bit of fool.' (Armstrong, Lindsay. 2011. *Accidental Nanny*, ch. 6)

Notre analyse s'accorde donc avec celle de WordNet. La figure 8 montre le nœud lexical *seventh heaven* dans le réseau *en-LN* après un traitement lexicographique. Le lien syntagmatique vers la préposition *in* permettant de former une collocation à valeur adjectivale a été encodé dans la zone FL. Remarquons que tout comme son équivalent français, cette collocation se trouve régulièrement enchâssée dans celle qu'elle forme avec le verbe *to be*.

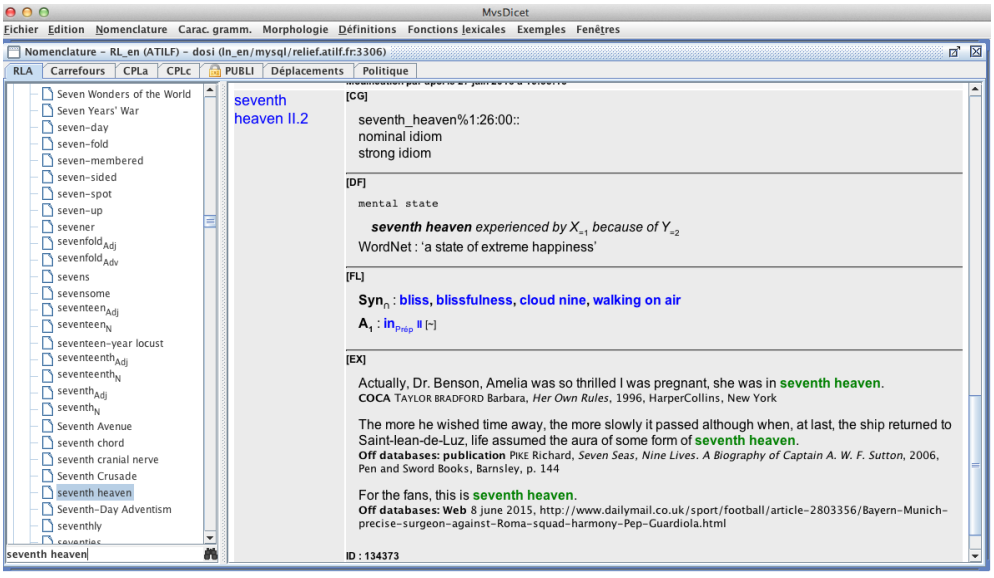


Figure 11. Locution *seventh heaven* dans la base *en-LN* (vue d'article).

4. CONCLUSIONS

Les quatre locutions dont cet article détaille l'analyse nous ont permis d'illustrer le traitement des UPs dans deux ressources lexicales de type dictionnaires virtuels. En identifiant le statut grammatical de chaque phrasème dans le cadre strictement défini d'un même modèle lexical, on vise à départager ce que celui-ci contient de locutionnel et ce qui relève de la collocation. Les liens syntagmatiques et pourront ensuite être tissés pour d'une part, situer le nœud dans le lexique et d'autre part, fournir aux utilisateurs humains et aux logiciels les informations nécessaires pour en maîtriser la combinatoire syntaxique.

Une fois qu'un nœud est inséré dans un système lexical construit de manière autonome, deux modes d'interconnexions interlinguistiques sont proposés par Polguère (2009). Premièrement, des nœuds peuvent être connectés directement entre eux : on établira ainsi un lien d'équivalence entre *pluie* I, nœud du système lexical *RL-fr* et *rain* 1 dans *en-LN*, dénotant tous les deux le même phénomène atmosphérique. Un deuxième mode d'interconnexion s'applique aux collocatifs des lexies équivalentes. La connexion s'effectue par l'intermédiaire de relations sémantico-lexicales (FL), en aboutissant à des équivalences que l'on pourrait qualifier de locales. Ainsi, l'application de la FL Magn (intensificateur) aux arguments *pluie* I en français et *rain* 1 en anglais retourne respectivement les valeurs telles que *violent(e)* II, *diluvien(ne)* en français, *heavy* 2 en anglais. Or, il est clair que des lexies telles que *diluvien(ne)* et *heavy* 2 ne peuvent être reliées qu'à travers la relation sémantico-lexicale qui se construit sur le plan collocationnel avec leurs bases respectives.

Le mode d'encodage et l'organisation des informations proposés dans le modèle de systèmes lexicaux permettra, selon notre hypothèse, de limiter, voire d'éviter certaines erreurs de traduction, notamment celles liées aux collocatifs non identifiés comme tels. Si les deux ressources sont encore trop jeunes pour des tests effectués sur l'ensemble de leurs réseaux, leurs structures autorisent des vérifications locales à partir des sous-graphes que forment des lexies bien décrites des champs lexicaux sur lesquels le travail de description a déjà été effectué, par exemple celui des sentiments.

Bibliographie :

GADER, N., LUX-POGODALLA, V., AND POLGUERE, A., 2012. Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor. *Proceedings of the Third Workshop on Cognitive Aspects of the Lexicon (CogALex III)*, The COLING 2012 Organizing Committee, Mumbai, pp. 109–125.

GADER, N., OLLINGER, S., AND POLGUERE, A. 2014. One Lexicon, Two Structures: So What Gives? In: H. ORAV, Ch. FELLBAUM and P. VOSSEN, eds 2014. *Proceedings of the Seventh Global Wordnet Conference (GWC2014)*. Tartu (Estonie), Global WordNet Association, pp. 163-171.

FELLBAUM, CH. 2014. Large-Scale Lexicography in Digital Age. *International Journal of Lexicography*, 27/4, pp. 378-395.

LUX-POGODALLA, V. AND POLGUERE, A. 2011. Construction of a French Lexical Network. Methodological Issues. *Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011. An ESSLLI 2011 Workshop*, Ljubljana, pp. 54–61.

MCCARTHY, M. 1999. What constitutes a basic vocabulary in spoken communication? *Studies in English Language and Literature*, 1, 233-249.

MEL'CUK, I., 2013. Tous ce que nous voulions savoir sur les phrasèmes, mais... *Cahiers de Lexicologie*, 102, 1, pp. 129-149.

MEL'CUK, I., CLAS, A. AND POLGUERE, A. 1995. *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

OSHERSON, A. AND FELLBAUM, Ch. 2010. The Representation of Idioms in WordNet. *Proceedings of Global WordNet Conference 2002, CFILT, IIT*, Bombay, Mumbai, 2010. < http://www.cfilt.iitb.ac.in/gwc2010/pdfs/16 Representation of Idioms in WordNet Osherson.pdf> [Accessed 4 May 2015]

POLGUERE, A., 2009. Lexical systems: graph models of natural language lexicons. *Language Ressources and Evaluation*, 43, pp. 41-55.

POLGUERE, A. 2011. Classification sémantique des lexies fondée sur le paraphrasage. *Cahiers de Lexicologie*, 98-1, pp. 197-211.

POLGUERE, A. 2014. From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 27/4, pp. 396-418.

POLGUERE, A. AND SIKORA, D. 2013. Modèle lexicographique de croissance du vocabulaire fondé sur un processus aléatoire, mais systématique. In: C. Masseron, C. Garcia-Debanc, Ch. Ronveaux eds, 2013. *Enseigner le lexique. Pratiques sociales, objets à enseigner et pratiques d'enseignement*. Laval (Québec): Presses universitaires de Laval, pp. 35-63.

SIKORA, D. 2014. Brève histoire d'une erreur lexicale. Polysémie et liens lexicaux dans l'enseignement du vocabulaire. In: F. NEVEU, P. BLUMENTHAL, L. HRIBA, A. GERSTENBERG, J. MEINSCHAEFER and S. PREVOST, eds 2014. *4e Congrès Mondial de Linguistique Française – CMLF 2014*, Paris: Institut de Linguistique Française, pp. 1157-1172.

SVENSEN, B., 2009. *A Handbook of Lexicography: The Theory and Practice of Dictionary-making*. Cambridge, New York: Cambridge University Press.

VAGUER, C., 2010. *Être aux anges*, *sortir de ses gonds*... Comment les langues traduisent-elles des états émotionnels ? *Cahiers Sens public*, 1, n° 13-14, pp. 253-269.

Websites

BNC, BRITISH NATIONAL CORPUS. [online] Available at http://www.natcorp.ox.ac.uk/

CAMBRIDGE DICTIONARIES ONLINE. [online] Available at http://dictionary.cambridge.org

COCA, 1990-2012, *Corpus of Contemporary American English*. [online] Available at http://corpus.byu.edu/coca/

Collins. [online] Available at < www.collinsdictionary.com/dictionar/english>

FRANTEXT, 2014, *Base textuelle Frantext*. [online] Available at www.frantext.fr

FRWAC, 2008, *Corpus of French Web*. [online] Available at http://nl.ijs.si/noske/wacs.cgi/first_form?corpname=frwac;lemma=;lpos=

GOOGLE TRADUCTION, 2015. [online] Available at https://translate.google.fr/?hl=fr

REVERSO, 2015. [online] Available at www.reverso.net

WORDNET, 2015, *A lexical database for English*. [online] Available at https://wordnet.princeton.edu/

CROSS-LINGUAL EXTRACTION OF MULTIWORD EXPRESSIONS

Shiva Taslimipoor
University of Wolverhampton
shiva.taslimi@wlv.ac.uk

Abstract

A multiword expression (MWE) can be defined as a combination of two or more words that, together, result in a special meaning. Statistical association measures are widely used in MWE identification/acquisition due to the collocational behavior of these expressions. MWEs like *take place* in English and *tener lugar* in Spanish, which are translation equivalents of each other, have high degrees of association measures (in their own languages). In this paper, we propose to combine the association measures of translation-equivalent expressions in order to rank expressions. In this way, expressions are scored according to a new cross-lingual association measure, which more reliably identifies them as being MWEs.

1. INTRODUCTION

Multiword expressions (MWEs) are defined as idiosyncratic interpretations that cross word boundaries or spaces (Sag et al. 2002). Examples are *frying pan*, *take a look* and *take part* in English and *base de datos*, *tener lugar* and *dar la bienvenida* in Spanish. One of the standard approaches to dealing with MWEs in Natural Language Processing is to rank them according to the degrees of association between their components (Ramisch et al. 2010). Existing statistical measures of collocation and words association are computed for any sequence of words according to their occurrences in monolingual corpora.

In this paper, we consider the association measure of each expression in one language together with the association measure of its translation equivalent in the other language in order to improve MWE identification. Combining the association measures, which are computed for translation equivalent expressions separately, will boost the performance of extracting MWEs using association measures. In this way, truly MWEs like *take place*, which have translation equivalents with high association degrees, would achieve higher ranks according to the new cross-lingual association measure (since e.g. the translation equivalent, *tener lugar*, has a high association degree). On the other hand, some expressions with high association degrees like *have nothing* in English, which do not have translation equivalents in Spanish with high association degrees, would achieve a lower rank according to the new cross-lingual association measure. To extract translation equivalent expressions, we follow our previous work (Taslimipoor et al. 2015).

2. RELATED WORK

There is a large body of recent studies describing different properties of various MWEs (Fazly 2007, Baldwin and Kim 2010). However, the cross-lingual analysis of these expressions and automatic extraction of their translation equivalents is still an under-researched topic (Bouamor et al. 2012). Contrastive and cross-lingual studies based on parallel and comparable bilingual corpora can benefit from statistical analysis of the various categories of such expressions (Colson 2008, Corpas Pastor 2013). The works on identifying MWEs have proved to benefit from the use of standard statistical association measures in determining the acceptability of different expressions, i.e. the degree to which they can be considered MWEs (Ramisch et al. 2010).

Inspired by previous research where the performance in one language is used to improve the performance in other languages as in the case of bilingual pronoun resolution (Mitkov and Barbu 2002) and bilingual term extraction (Ha et al. 2008), we seek to benefit from the extraction of MWEs in one language in order to boost the extraction performance in the other.

3. METHODOLOGY

The main purpose of our investigation is to combine the association measure of an expression with the association measure of its translation equivalent to achieve a bilingual association measure which ranks the expressions better as being MWE or not. We focus on two statistical association measures: Frequency and Likelihood (Ramisch et al. 2010). They capture in different ways the qualification of how often given words co-occur with one another.

In this work, we consider Verb + Noun constructions and calculate how likely they are co-occurring and so can be considered as MWEs. For every Verb + Noun expression, Frequency, is simply, how many times the verb and the noun co-occur consecutively and Likelihood is the logarithmic probability of the co-occurrences of the verb and the noun tokens. For each expression, the summation of its association measure with the association measures of its translation equivalents are computed. To find the translation equivalents of expressions we use our previous distributional approach in automatically translating MWEs in parallel corpora. Applying that distributional approach, every expression is vectorised according to its surrounding context and the expressions with most similar vectors are considered as translation equivalents (Taslimipoor et al. 2015).

4. EXPERIMENTS AND EVALUATION

4.1. Resources and Experimental Setup

In the evaluation of the approach, we focus on the 20 most frequent verbs in English and every Verb + Noun combination they construct. We extract the 20 most frequent verbs from the BNC corpus. These verbs are commonly and productively used in Verb + Noun combinations. To illustrate, in the English side of the Europarl corpus, they co-occurred (with frequency >15) with 4930 different nouns from which 156 items are MWEs, according to the annotators.

The English-Spanish Europarl parallel corpus is used to extract Verb + Noun combinations and their frequencies. To our knowledge, Europarl is the most reliable parallel corpus that exists for English and Spanish (Koehn 2005). We automatically POS tag the corpus using TreeTagger. We first extract all Verb + Noun sequences (bigrams) from the corpus considering the 20 selected verbs. We then augment these Verb + Noun candidates with their frequencies of occurrences through the corpus. Ignoring the expressions with frequencies less than 5, we randomly select 1000 expressions among which MWEs should be identified. The Verb + Noun English expressions were annotated by two native speakers with 1 (as being a MWE) or 0 (as not being a MWE). The inter-annotator agreement between annotators according to the Kappa measure is 0.69.

The first experiment is to evaluate MWE identification when ranking expressions according to a monolingual statistical measure compared with MWE identification while ranking them according to the bilingually combined statistical measure which considers translation equivalents. To evaluate the performance of involving translation equivalent in the first experiment we used the Glosbe online dictionary[54] to automatically retrieve the translations of expression. Firstly English V+N expressions were ranked according to their frequencies in the English side of Europarl. Then those English expressions were ranked according to the new frequency measure which is the summation of the frequency of that English expression and the frequency of its Spanish equivalents. Furthermore, to compute the Likelihood based measure, we use the collocation modules of the NLTK package (Loper and Bird, 2002) to calculate likelihood measures for such combinations.

In the second experiment, we focus on a smaller sample of 100 English expressions. We find the Spanish translation equivalents of the expression using our distributional similarity approach and repeat the calculation of Frequency and Likelihood measures like the first experiment. The results of the experiments are reported in section 4.3.

4.2. Evaluation Measures

The performance of MWE identification regarding different automatic ranking schemes are reported using F-score which is the combination of the Precision and Recall scores. For each ranking schemes, i) ranking according to monolingual Frequency or Likelihood, and ii) ranking according to combined bilingual Frequency or Likelihood, we set the threshold at various points, and count the number of true MWEs above the threshold (true positive) and the number of non-MWEs above the threshold (false positive), as well as the number of MWEs and non-MWEs below the threshold (false positive and true negative, respectively). We compute Precision, Recall and F-measure regarding each threshold and plot the corresponding curves.

4.3. Results and Discussion

The results of improving MWE identification for English expressions are presented in this section. As can be seen in figure 1, the F-measure performance of ranking expressions according to the new measure is better than the F-measure performance of ranking them according to their monolingual statistical measures (old measures). This improvement applies for both ranking schemes: Frequency (in the left sub-plot) and Likelihood (in the right sub-plot). To illustrate for the left sub-plot, while the old measure is the standard Frequency of expressions in a monolingual corpus, the new measure is the combination of the Frequency of expressions with the Frequency of their translation equivalents. The illustration is the same for the Likelihood in the right sub-plot.

[54] https://glosbe.com/

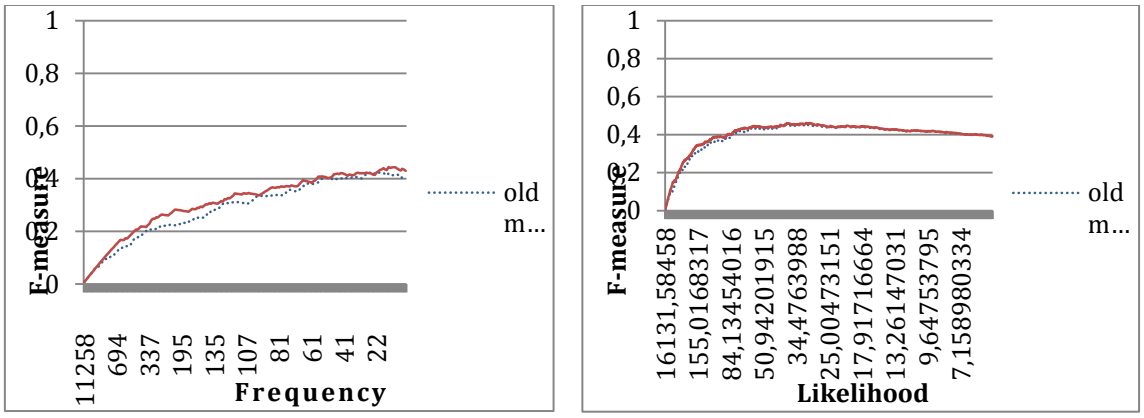


Figure 12. F-measure comparison of MWE identification, ranking the expressions using Monolingual (old) and Bilingual(new) Frequency-based and Likelihood-based measures.

Moreover, comparing the left frequency based subfigure with the right Likelihood based one in Figure 1, clearly, the Likelihood score works better than Frequency in ranking expressions as being MWEs.

In our second experiment, Ranking expressions using the same measures but applying the automatic distributional approach for finding translation equivalents, the trend in the performance is slightly similar. According to the F-measure curve regarding all thresholds, we set the best threshold of 157 for Frequency and 112.66 for Likelihood. Resulted F-measure performance is reported in Table 1.

	Monolingual	Bilingual
Frequency	0.5	0.59
Likelihood	0.74	0.76

Table 8. F-measure scores comparison of MWE identification, ranking 100 selected English expressions using Monolingual (old) and Bilingual(new) Frequency-based and Likelihood-based measures.

The results in Table 1 for the second experiment support the conclusion of the first experiment in that Likelihood, in general, overcome Frequency in ranking expressions as MWE. Furthermore, if we combine the statistical measures of expressions with the statistical measure of their equivalents, we receive better results for the new bilingual measure than the old monolingual association measure.

5. CONCLUSION AND FUTURE WORK

In this paper an improvement in the identification of MWEs using statistical association measures is presented. The improvement is achieved by combining the statistical association measures of expressions with the statistical association measures of their translation equivalents. Specifically, we conclude that ranking expressions according to a new association measure which is the summation of the association measure of the

expressions and the association measures of their translation equivalents would result in a better F-measure performance in MWE identification.

This method can not only be beneficial for (and enhance the performance of) machine translation systems but can also offer new opportunities for cross-lingual studies on MWEs based on their occurrences in bilingual corpora. This methodology could also assist lexicographers when deciding which MWEs should be listed in bilingual dictionaries and could speed up the semi-automatic compilation of such dictionaries.

References

BALDWIN, T. and KIM, S. N., 2010. "Multiword expressions". *Handbook of Natural Language Processing, Second Edition*, N. Indurkhya and F. J. Damerau, eds., CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

BOUAMOR, D., SEMMAR, N., and ZWEIGENBAUM, P., 2012. "Identifying bilingual multi-word expressions for statistical machine translation". *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* Istanbul, Turkey.

COLSON, J. P., 2008. "Cross-linguistic phraseological studies: An overview". *Phraseology: An interdisciplinary perspective*, S. Granger and F. Meunier, eds., John Benjamins Publishing Company, Amsterdam/ Philadelphia: Meunier, John Benjamins Publishing Company.

CORPAS PASTOR, G., 2013 "Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas". *In: Inés Olza / Elvira Manero (eds.): Fraseopragmática*, Berlin: Frank & Timme, 335-373.

FAZLY, A., 2007. "Automatic acquisition of lexical knowledge about multiword predicates". Ph.D. thesis, Department of Computer Science, University of Toronto, Department of Computer Science, University of Toronto.

HA, L. A., FERNANDEZ, G., MITKOV, R., and PASTOR, G. C., 2008. "Mutual bilingual terminology extraction". *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008*, Marrakech, Morocco.

KOEHN, P., 2005., Europarl: A Parallel Corpus for Statistical Machine Translation. *Conference Proceedings: the tenth Machine Translation Summit*, AAMT, Phuket, Thailand, AAMT, 79-86.

LOPER, E. and BIRD, S., 2002., Nltk: The natural language toolkit. *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, Association for Computational Linguistics, 63-70.

MITKOV R. and BARBU C. 2002., "Using corpora to improve pronoun resolution". *Languages in contrast*, 4(2), 201-211.

RAMISCH, C., VILLAVICENCIO, A., and BOITET, C. 2010., "MWEToolkit: a framework for multiword expression identification", Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias, eds., Valletta, Malta, European Language Resources Association (ELRA)

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A. A., and FLICKINGER, D., 2002. "Multiword expressions: A pain in the neck for nlp". *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing,* CICLing '02, London, UK, Springer-Verlag, 1-15.

TASLIMIPOOR, S, MITKOV, R, and CORPAS PASTOR, G., 2015. "Using cross-lingual contexts to extract translation equivalents for Multiword Expressions from parallel corpora". *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies* (AIETI'15). Malaga, Spain, January, 2015, 103-105.

INTEGRATING VERB+NOUN COLLOCATIONS INTO A FRENCH - ROMANIAN LEXICAL ALIGNMENT SYSTEM FOR LAW DOMAIN

Amalia Todirascu
LiLPa
(Linguistique, Langues, Parole),
Université de Strasbourg
todiras@uni stra.fr

Mirabela Navlea
LiLPa
(Linguistique, Langues, Parole),
Université de Strasbourg
mirabela_abe@yahoo.com

Abstract

In this article, we compare two methods to integrate a specific class of multiword expressions, Verb+Noun collocations, into a French - Romanian lexical alignment tool. In our experiments, we use a French - Romanian parallel corpus for law domain. This corpus is tokenized, tagged, lemmatized and chunked. The first method uses a dictionary-based approach to complete Verb+Noun collocations alignment. The second method proposes an alignment algorithm which uses a set of MWEs candidates previously extracted from the monolingual part of the training corpus. These candidates were detected by a hybrid extraction method combining statistical measures and linguistic filters. The best results were obtained with the hybrid method.

1. MULTIWORD EXPRESSIONS AND LEXICAL ALIGNMENT

Multiword expressions (MWEs) are characterized by several degrees of lexical, syntactic, morphosyntactic and statistical idiomaticity (Baldwin and Kim, 2010). MWEs include idiomatic expressions, collocations, verb or noun compounds, domain specific terms, named entities. Each subclass of MWEs has its own set of properties and it requires specific identification, extraction or translation methods. Some classes, such as idiomatic expressions are characterized by their fixedness and their highly non-compositional sense. Other classes such as verbal collocations (verb particle constructions such as *give up, take off,* light verb constructions such as *make a mistake, take a break, faire appel* 'to appeal', *donner force* 'bring to force') have more variable syntactic properties and more compositional sense.

Lexical alignment is a necessary step for statistical machine translation (SMT) systems. Generally speaking, lexical alignment methods aim to identify translation equivalents. Some lexical alignment methods apply statistical methods to compute possible word-to-word matches (Och and Ney, 2003). Frequent bilingual pairs of words occurring in parallel sentences are identified as possible translations. Other methods take into account linguistic information to complete lexical alignment (Tufis ***et al,*** 2006).

The existing alignment methods generally fail to handle MWEs alignments, due to their specific syntactic, lexical or semantic behavior. Indeed, if highly fixed expressions might be identified by statistical methods, more variable constructions (such as verb+particle constructions, verbo-nominal collocations) are not properly aligned. Moreover, MWEs have strong lexical preferences: the translation equivalent is valid only in specific contexts (the French expression *donner lieu* 'to give place' is aligned with the Romanian expression *a da nastere* 'to give birth', but *nastere* is not equivalent of ***lieu*** in other contexts, except if the verb ***a da / donner*** 'to give' is occurring in the context). Non-compositional sense is also a difficult task for alignment methods due to the lack of external knowledge (for example, we should align the Romanian expression *a pune paie pe foc* 'to put fuel on the fire' with the French equivalent expression *mettre de l'huile sur le feu***,** having similar meaning). Many-to-one alignments are possible: a MWE is equivalent with a single noun or a single verb (the Romanian Verb+Noun collocation *a repara pagubele* 'to make good any damage' is equivalent to the French verb *dédommager*). Free word order (such as in Romanian language) may generate alignment errors. Several possible translations might be available, so statistical lexical alignment systems generally fail, due to data sparseness.

To solve these problems, several alternative methods were proposed. Melamed (1997) develops a specific alignment algorithm to build many-to-many alignments, by identifying longest MWEs candidates. Tiedemann (2004) uses linguistic knowledge to complete alignment. Other methods use single word alignment and extra knowledge (lemma, POS tag, etc.) to identify many-to-many alignments (Villadamoiron and Tiedemann, 2006; Pal *et al,* 2013). Several systems proceed with MWEs extraction from monolingual corpora before alignment (Ren *et al.,* 2009; Ramisch *et al.,* 2013). The extraction combines linguistic filters and statistical methods to identify possible MWEs candidates in monolingual corpora. Then, the extracted monolingual candidates are aligned via specific algorithms combining linguistic information and statistical measures. Some methods identify noisy word links and propose new algorithms to repair these problems (Okita, 2012). Lambert and Banchs (2006) extract n-grams of words to find possible bilingual phrases and apply some linguistic information to filter out invalid candidates. Other specific alignment methods focus on specific classes of MWEs (NE, terms) (Bouamor et al., 2013).

In this paper, we focus on the presentation of the collocation definition adopted for our project, the architecture of the French - Romanian lexical alignment system, the methods used to align the Verb+Noun collocations, and their evaluation. We compare two collocation alignment methods. The first method uses external resources, such as a multilingual dictionary (Todirascu *et al.,* 2008), combined with a specific alignment algorithm. The second method exploits a hybrid strategy to extract Verb+Noun collocations from monolingual corpus (Todirascu *et al.,* 2009) and then an alignment algorithm to map these monolingual candidates in the parallel corpus. Both methods use rich morphosyntactic and syntactic properties to detect many-to-many alignments.

2. COLLOCATION DEFINITIONS

In this project, we are particularly interested in aligning Verb+Noun collocations. This is a specific class of MWEs, which is relatively independent across domains. Collocations have no unique, generally recognized definition by all the researchers from several domains. Computational linguistic researchers propose functional definitions, to be able to automatically identify them, while linguistic researchers focus on detailed analysis of these classes of MWEs. In this article, we consider collocations as being MWEs, sometimes discontinuous, with specific morphosyntactic, semantic or pragmatic behavior (Gledhill, 2007). In our project, we apply three criteria to identify collocations: the frequency, the syntactic idiomaticity and the semantic idiomaticity (Gledhill, 2007; Gledhill and Todirascu, 2008). Indeed, the collocations are frequent word associations (Firth, 1956; Williams, 2003), which are related by syntactic links (Hausmann, 2004) and having a non-compositional sense. In this project, we focus on two classes of Verb+Noun collocations (constructions) according to Gledhill (2007):

- **complex predicators**, presenting high morphosyntactic and syntactic fixedness
and a non-compositional sense *(jeter l'éponge* 'throw the towel', *avoir peur* 'to be scared'). Indeed, the complex predicators have strong preferences for some specific morphosyntactic properties of the verb or of the noun. The noun is always in singular or in plural, a specific determiner (zero, definite or indefinite) is preferred, for most of the occurrences. The verb may only by used in active voice. No adverbial modifiers are accepted between the verb and the noun. In this category of Verb+Noun collocations, the noun specifies the range of the process expressed by the verb, according to Halliday's analysis (Halliday, 1985) (Gledhill, 2007). These constructions usually express relational processes. According to Baldwin and Kim (2010) classification, this class includes idiomatic verbal constructions;
- **complex predicates**. This class of constructions is more variable in terms of
syntactic contexts *(a lua masuri* 'to take measures'; ***a tine o conferinta*** 'to give a talk'). The noun accepts several classes of determiners and it is more variable in number, gender or case. The verb is found in both active and passive voice and modifiers are accepted between the verb and the noun. From the syntactic point of view, the noun is the direct object of the verb. In this category of Verb+Noun collocations, the noun also specifies the range of the process expressed by the verb (Gledhill, 2007). The sense of the collocation usually is an abstract, mental process. According to Baldwin and Kim (2010) classification, this class is equivalent to the light verb constructions classes.

The properties characterizing each class are useful to semi-automatically identify

3. THE ARCHITECTURE OF THE LEXICAL ALIGNMENT SYSTEM

The architecture of the French - Romanian lexical alignment system (Navlea, 2014) integrates the GIZA++ aligner (Och and Ney, 2003) as a first step. This aligner is applied in both directions (French -> Romanian and Romanian -> French). For the next step, only sure links are kept. These links are identified by intersecting the alignments obtained in both directions (Koehn *et al.,* 2003). The resulting alignment is considered as the baseline of the system. To improve this baseline several strategies are applied:

cognate detection and alignment, linguistic heuristic rules based alignment, collocation integration. We focus here on this last step of the collocation alignment.

For our experiments, in order to train the alignment system, we use a law parallel corpus of about 1,500,000 tokens per language, while in order to test the system, we use a small parallel corpus of about 30,000 tokens per language (1,000 parallel sentences) (Navlea, 2014). These corpora are extracted from the DGT-TM law corpus (Steinberger *et al,* 2012), being also sentence-aligned. Additionally, they are tokenized, tagged, lemmatized and chunked with TTL, a POS tagger available for Romanian (Ion, 2007) and for French (Todirascu *et al.,* 2011). We give in the Figure 1 an example of output of this tagger for the French language.

```
<seg lang="fr"><s id="ttlfr.2">¶
<w lemma="Les" ana="Da-fp.Da-mp">Les</w>¶
<w lemma="droit" ana="Ncmp" chunk="Np#1">droits</w>¶
<w lemma="et" ana="Cc">et</w>¶
<w lemma="obligation" ana="Ncfp" chunk="Np#2">obligations</w>¶
<w lemma="découler" ana="Vmpp">découlant</w>¶
<w lemma="de" ana="Spd" chunk="Pp#1">de</w>¶
<w lemma="le" ana="Da-fs" chunk="Pp#1.Np#3">l'</w>¶
<w lemma="adjudication" ana="Ncfs" chunk="Pp#1.Np#3">adjudication</w>¶
</s></seg>¶
```

Figure 1 Example of output of TTL for the French language; each word <w> is associated with lemma (the *lemma* attribute), morphosyntactic properties (the *ana* attribute) and chunks (the *chunk* attribute)

The POS tagsets used by TTL are the tagset from MULTEXT project (Ide and Veronis, 1994), for French, and the tagset from MULTEXT-EAST (Erjavec, 2004), for Romanian language.

As we mentioned previously, the lexical alignment system first uses a specific method to detect and to align cognates (Navlea and Todirascu, 2012) in order to improve the baseline system. This module combines statistical techniques, linguistic information and orthographic adjustments. A second module implements a set of specific linguistic heuristic rules to complete the previous alignments (Navlea and Todir^fcu, 2011; Navlea, 2014). These linguistic rules are based on the morphosyntactic differences between both languages.

Concerning the collocation alignment, as other approaches proposed in the literature, we experiment two methods. As a first strategy, we use an external French - Romanian bilingual dictionary (Todirascu *et al.,* 2008) to identify many-to-many and many-to-one mappings. However, bilingual collocation dictionaries are not always complete or available for different pairs of languages. So, we propose to use an alternative method: we extract MWEs from monolingual parts of the training corpus and we try to map bilingual candidates using complex POS tags or lemmas.

Navlea (2014) presents experiments with Verb+Noun collocations applied as a last step in the lexical alignment system, after cognate and heuristic based strategies, by using collocation dictionary alignment method. In this paper, we apply the dictionary, but also the MWEs extractor, directly to the baseline system, in order to compare the results and then to find the best way to integrate collocations in the system. We give in the Figure 2 the architecture of the French - Romanian lexical alignment system.

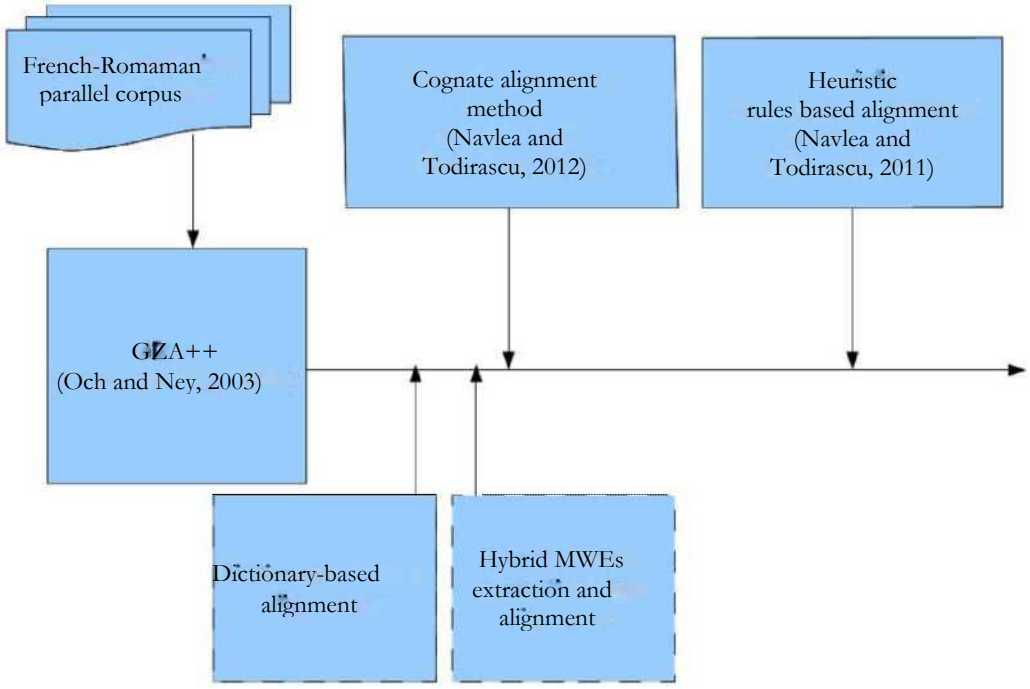


Figure 2 The architecture of the French - Romanian lexical alignment system

4. THE VERB+NOUN COLLOCATIONS ALIGNMENT METHODS

In our experiments, we apply two various strategies which consist of completing the baseline alignment with multiple links. The collocation identification methods are the following:

- applying an external resource (a bilingual collocation dictionary (Todirascu *et al,* 2008)) (Navlea, 2014);
- using a hybrid collocation extraction method (Todirascu *et al,* 2009).

For the two methods, specific algorithms taking into account the morphosyntactic properties of the collocations are applied.

4.1. The Dictionary-based Method

For these experiments, we use a small bilingual dictionary, manually developed from law corpus (Todirascu *et al,* 2008) to identify many-to-many or many-to-one alignments.

4.1.1. The Dictionary

The dictionary is composed of 250 bilingual Verb+Noun collocations and their translation equivalents. Each entry contains rich information about the morphosyntactic, syntactic and semantic environment of the construction. The bilingual entry represents the translation equivalents, single units or collocations. For each collocation (<complexitem>), the information represented in the dictionary is the lemmas of the Verb+Noun constructions, their type (complex predicate or complex predicator), and the properties of the collocations (preference for one preposition <prep>, their class

<construction>, some examples of usage and the corpus where the sample was identified). An example of bilingual entry *jeter l'ancre* (FR) vs. *a ancora* (RO) 'to anchor' is given in the Figure 3.

```
<entry id="56">
<te lang="fr"><complexitem> <construction>jeter l'ancre</construction>
<v_spec><lemma>jeter</lemma><voice freq="75">
passive</voice><voice freq="25">active</voice></v_spec>
<prep>null</prep>
<n_spec><lemma>ancre</lemma>
<det freq="100">def</det>
<nb freq="100">sg</nb>
</n_spec>
<colloc_spec><arg type="indirect object"><prep>de</prep></arg></colloc_spec>
</complexitem>
<colloc_documentation>
<example corpus="JRCAcquis"></example>
</colloc_documentation>
</te>
<te lang="ro"><simpleitem>a ancora
<v_spec><lemma>a ancora</lemma></v_spec>
<prep/>
<n_spec/>
<colloc_spec/>
</simpleitem>
</te>
</entry>
```

Figure 3. An example of bilingual entry, with a French complex item corresponding to a single verb in Romanian

The entry also specifies the properties of the verb <v_spec> with additional information about specific lexical preferences or voice. For the noun (<n_spec>), the information represented in the entry concerns the number, the type of possible determiners or the modifiers which occurs between the verb and the noun. Each property is represented as the percent of occurrences found in all the contexts. For each collocation, the dictionary identifies one or several translation equivalents in the other language. Some of these translation equivalents are collocations from the same classes or single words (as in the example <simpleitem>). This information is used to correctly build many-to-many or many-to-one lexical alignments.

4.1.2. *Aligning Verb+Noun Collocations using the Dictionary*

A first method used to align Verb+Noun collocations exploits the dictionary to search for all the collocations (the pairs of verbs and nouns) and their translation equivalents into the parallel French - Romanian corpus.

We start searching for the pairs of verb and noun in the French sentences. For each French sentence containing a candidate, we select the Romanian sentence and we look for the corresponding translation equivalents (the corresponding verb and noun, or a single word if this is the case). The algorithm adds some new links between the verb or the noun and their correspondent translation equivalents. Then, we use the

information about the morphosyntactic and syntactic environment of the collocation to complete the alignment. We include in the alignment mandatory properties such as determiners or auxiliaries found between the pairs of verb and nouns. In the example of the parallel sentences given in the Figure 4, we generate many-to-many new links (marked with dashed lines).

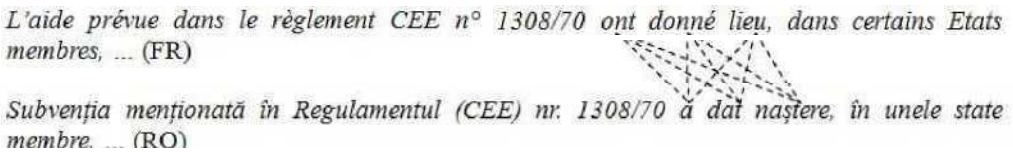


Figure 4 Example of many-to-many alignments using the dictionary

4.2. The Hybrid Extraction and Alignment Method

Our second method aims to identify collocations in monolingual parts of the training corpus and then to map the possible bilingual candidates (Ramisch *et al,* 2013). Thus, to detect monolingual candidates we apply a hybrid extraction method (Todirascu *et al,* 2009). Then, we align these collocation candidates using a specific algorithm.

4.2.1. The Hybrid Method

The method extracting the collocation candidates combines linguistic and statistical information (Todirascu *et al,* 2009). Indeed, MWEs are frequent word associations, characterized by specific morphosyntactic, syntactic and semantic properties. Thus, our method searches first for the frequent Verb+Noun associations. The statistical measure used to extract collocations is Log-likelihood (LL) (Dunning, 1990). We select all the pairs of verb and nouns and their contexts, in a window of 11 words, with LL > 9. Then, linguistic filters are applied to identify possible Verb+Noun collocations and to delete several irrelevant candidates.

Linguistic filters identify expressions with strong preferences for a specific class of determiners, a specific number or case. For example, a filter built to identify Romanian complex predicators will check for verbs *(pos=V.*)* followed by singular noun with a null determiner *(pos=Nc.s-n***)** and followed by a dative determiner (T-determiner, i - indefinite, f - feminine, s - singular, o- dative):

[pos=V.*][pos=Nc. s-n][ana=Tifso]

This pattern checks if this condition is fulfilled by all the contexts of the verb and of the noun. It identifies collocations such as *face fata* 'to face', *a da nastere* 'to give birth'.

Several filters are used to delete invalid candidates: pairs of verbs and nouns separated by several prepositional groups (at least two PP), pairs of nouns modified by the past participle verb.

4.2.2. Aligning MWEs candidates using the hybrid method

We extracted from the corpus a set of MWEs candidates for each language, by applying the method described in the previous section. For a French monolingual candidate we select the sentences containing the verb and the noun within a distance of

5 words. Then, from the aligned target sentences, we identify several target pairs of verbs and of nouns. For each target pair found in the aligned sentences, we compute a weight by taking into account:

- the existing alignments between verb - verb, noun - noun ;
- if some determiners are found in a window of 3 words around the nouns ;
- if the distance between the source and the target pairs is less than 5 words ;
- if no other noun or verb are used between the elements of the target pair.

From several candidates we order the pairs according to their score. We do the same process starting with the Romanian candidates. We search the possible corresponding French collocation equivalents for each Romanian candidate. If the initial French pair is in this list, then the Romanian candidate is selected as a possible translation equivalent. For our experiments, we select 250 monolingual candidates for each language, with a high log-likelihood score (more than 100).

5. COMPARING THE METHODS

For comparison purposes, we use a lexically aligned French - Romanian reference corpus (Navlea, 2014). This corpus is composed of 1,000 pairs of sentences, which was manually aligned. All Verb+Noun collocations are also aligned.

The evaluation measure is the AER (Alignment Error Rate) score (Och and Ney, 2003) which is usually computed by taking into account the sure, but also the probable alignments. However, during the manual alignment process, all alignments were considered as sure. Then the AER score is computed as:

$$AER = 1 - F\text{-measure}$$

The two collocation alignment methods (dictionary and MWEs extractor based alignment) were applied independently to the baseline alignment, which does not contain collocation alignments.

The results of the methods comparison are given in the Table 1 :

Systems	Recall	Precision	F-measure	AER
Baseline	49.71 %	95.51 %	65.39 %	34.61 %
Dictionary-based method	50.71 %	95.15 %	66.16 %	33.84 %
Hybrid method	53.13 %	93.99 %	67.89 %	32.11 %

Table 1. A comparison of the results obtained by the various systems

The baseline system (Navlea, 2014) obtained an AER score of 34.61%. The dictionary improves the AER score with 0.77%, which is not very significant. In a previous experiment (Navlea, 2014) the dictionary, applied after cognate and heuristic rules based alignments, improved the AER score with 0.32%. The dictionary method shows low improvements in both experiments because of the few collocations common to the dictionary and to the reference corpus. In this case, it seems that the size of this resource should be increased for further experiments.

The second method is more effective improving the baseline system with 2.50%, at least for the list of 250 candidates selected for the experiments. At the moment, this

second method does not find many-to-one links, while the dictionary proposes such alignments. This method also should be evaluated in the system architecture, after cognate and heuristic rules application (see section 3), to study its influence to the overall alignment results and then to find the best way of collocation alignment.

6. CONCLUSION AND FURTHER WORK

We present a comparison of two methods for building multiple links in order to improve a French - Romanian lexical alignment system. The dictionary-based method uses a small bilingual dictionary with rich morphosyntactic properties. The dictionary has low influence on the overall system performance, but this is due to its incompleteness and small coverage between the reference corpus and the bilingual dictionary. Further experiments will be performed with a larger dictionary including a set of about 230 Verb+Noun additional collocations and their nominalizations (Navlea, 2014).The second alignment method uses a hybrid strategy to extract Verb+Noun collocations candidates from monolingual parts of the training corpus and then an algorithm is applied to build multiple alignments. This method is more effective, even if it finds several irrelevant candidates and so the resulting alignments contain some false links. Further experiments should be continued with a larger list of monolingual candidates. Moreover, the MWEs extractor method, applied after the cognates and the heuristic rules step in the lexical alignment system, will be evaluated in order to study its influence to the overall alignment results.

References

BALDWIN, T. & KIM, S. N. (2010). Multiword expressions. In N. Indurkhya, & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (pp. 267-292), Second edition. Boca Raton (USA, FL): CRC Press, Taylor and Francis Group.

BOUAMOR, D., SEMMAR, N., & ZWEIGENBAUM, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of Eigth International Conference on Language Resources and Evaluation* (pp. 674-679). Istanbul, Turkey: ELRA.

de GISPERT, A., GUPTA, D., POPOVIC, M., LAMBERT, P., MARINO, J., FEDERICO, M., NEY, H., & BANCHS, R. (2006). Improving Statistical Word Alignments with Morpho-syntactic Transformations. In *Proceedings of 5th International Conference on Natural Language Processing, FinTAL '06* (pp. 368-379).

DUNNING, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics,* 19(1), 61-74.

ERJAVEC, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (pp. 1535-1538). Paris: ELRA.

GLEDHILL, C. (2007). La portée : seul dénominateur commun dans les constructions verbo-nominales. In Frath, P., Pauchard, J., & Gledhill, C. (Eds.), *Actes du 1er colloque, Res per nomen, pour une linguistique de la dénomination, de la référence et de l'usage* (pp. 113-125)*,* Université de Reims-Champagne-Ardenne.

GLEDHILL, C., & TODIRAS CU, A. (2008). Collocations en contexte : extraction et analyse contrastive. *Texte et corpus,* 3, *Actes des Journées de la linguistique de Corpus 2007* **(pp.** 137-148).

HAUSMANN, F. J. (2004). Was sind eigentlich Kollokationen?. In K. Steyer (Ed.), *Wortverbindungen -mehr oder weniger fest* (pp. 309-334). Institut fur Deutsche Sprache Jahrbuch.

IDE, N., & VERONIS, J. (1994). Multext (multilingual tools and corpora). In *Proceedings of the 15th CoLing* (pp. 90-96). Kyoto (Japon).

ION, R. (2007). *Metode de dezambiguizare semantics automaM. AplicMi pentru limbile engleM 0 româÆ [Méthodes de désambigiiisation sémantique automatique. Application pour les langues anglaise et roumaine]*. Ph.D.Thesis. Bucharest (Romania): Romanian Academy

LAMBERT P. & BANCHS R. (2006). Grouping multi-word expressions according to Part-Of-Speech in statistical machine translation. In *Proceedings of the EACL Workshop on Multi-word expressions in a multilingual context* (pp; pp. 9-16). Trento, Italy.

MELAMED D. I. (1997). Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing* (pp. 97-108). RI, USA: Providence.

NAVLEA, M., 2014. *La traduction automatique statistique factorisée : une application à la paire de langues français - roumain.* Ph.D. Thesis, Université de Strasbourg, Strasbourg.

NAVLEA, M., TodiralSm, A. (2012). Using Cognates to Improve Lexical Alignment Systems. *in* Petr Sojka, Ales Horak, Ivan Kope^ek, and Karel Pala (eds). *Text, Speech and Dialogue (15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings), Lecture Notes in Computer Science,* Volume 7499, Springer Verlag Berlin Heidelberg, 370-377. ISBN: 978-3-642-32789-6 (Print)

PAL, S., NASKAR, S. K., & BANDYOPADHYAY, S. (2013). MWE Alignment in Phrase Based Statistical Machine Translation. In K. Sima'an, M. L. Forcada, D. Grasmick, H. Depraetere, & A. Way (Eds.), *Proceedings of the XIVMachine Translation Summit* (pp.61-68).

RAMISCH, C., BESACIER, L., & KOBZAR, A. (2013). How hard is it to automatically translate phrasal verbs from English to French?. In J. Monti, R. Mitkov, G.. Corpas Pastor, V. Seretan (Eds.), *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology* (pp. 53-61), Nice (France).

REN, Z, LU, CAO, J., LIU, Q, & HUANG, Y. (2009). Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009* (pp. 47-54).

OCH, F. AND NEY, H., 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics,* 29(1), pp. 19-51.

SMADJA, F. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics,* 19, pp. 143-175.

STEINBERGER, R., POULIQUEN, B., WIDIGER, A., IGNAT, C., ERJAVEC, T., AND TUFIS, D., 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th LREC Conference*, Genoa, Italy, 22-28 May, 2006, pp.2142-2147, ISBN 2-9517408-2-4, EAN 9782951740822

TODIRASCU, A., HEID, U., |ETEFANESCU, D., TUFE, D., GLEDHILL, C., WELLER, M., AND ROUSSELOT, F, 2008. Vers un dictionnaire de collocations multilingue. In Xavier Blanco Escoda, Marie-Claude L'Homme et Marc Van Campenhoudt (éd) Special Issue on "Lexique, dictionnaire et connaissance dans une société multilingue", *CAHIERS DE LINGUISTIQUE, Revue de sociolinguistique et de sociologie de la langue française,* Éditions Modulaires Européennes, vol. 33/1, pp 161-186, 2008, ISSN : 0771-6524, ISBN : 978-2-930481-52-4

TIEDEMANN J. (1999). Word alignment -step by step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics* (pp. 216-227). University of Trondheim, Norway.

TODIRASCU A., GLEDHILL C., & STEFANESCU D. (2009). Extracting Collocations in Contexts. In Z. Vetulani, & H. Uszkoreit (Eds.), *Responding to Information Society Challenges: New Advances in Human Language Technologies, LNAI 5603* (pp. 336-349). Berlin Heidelberg: Springer-Verlag.

TODIRASCU, A., ION, R., NAVLEA, M., & LONGO, L. (2011). French text preprocessing with TTL. In *Proceedings of the Romanian Academy,* Series A, 12(2), 151-158.

TUFIS, D., ION, R., CEAUSU, A. AND STEFANESCU, D., 2006. Improved Lexical Alignment by Combining Multiple Reified Alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006),* Trento, Italy, 3-7 April, 2006, pp. 153-160, ISBN 1-9324-32-61-2

COMPUTER-AIDED AND/OR CORPUS-BASED ANALYSIS OF PHRASEOLOGICAL UNITS

ANÁLISIS DE UNIDADES FRASEOLÓGICAS BASADO EN CORPUS O ASISTIDO POR ORDENADOR

IDIOMS IN SPOKEN CORPUS. A SAMPLE OF CZECH DATA

František Čermák
Institute of the Czech National Corpus,
Faculty of Arts,
Charles University in Prague
Frantisek.Cermak@ff.cuni.cz

Marie Kopřivová
Institute of the Czech National Corpus,
Faculty of Arts,
Charles University in Prague
Marie. Koprivova@ff.cuni.cz

Abstract

It is a generally accepted view that at least some idioms (phrasemes) occur in the spoken communication primarily. Though this old idea needs a corroboration on extensive data, it is spoken corpora that are being built up in many languages that seem to be an ideal resource for this. Moreover, the situation of the Czech language appears to be favourable in that there exists a large multi-volume dictionary of idioms which can be used as a reference in background while a series of spoken corpora is now being built, too, some of them being already available.

Even a cursory look into a volume of the Czech idiom dictionary (Čermák, 2009) containing sentence-type idioms offers an interesting part in that many of its very first records start in an "iambic" manner, namely with an unstressed component or word (usually a particle) *a*. There are some 27 such idioms starting with *A* and 25 of those containing *No* standing at the beggining.

All of these (and other) have been used for searching the Prague Spoken Corpus (and other spoken corpora); the corpus is based on spoken unscripted and unprepared conversation between equal partners. It has been found out that these idioms are, as a rule, very short (see the examples in the part 2), and are often made up of grammar (synsemantic) words. Functionally and pragmatically, these are heavily loaded in the sense that they express various types of often short emotional reactions and evaluative attitudes. These idioms belonging mostly to propositional idioms, i.e. they form a sentence in each case, have a feature that has not been explored in phraseology much, namely a specific intonation, which is recorded in the dictionary, too.

The corpus exploration of these idioms, against the background of the idiom dictionary is concerned, next to finding their frequency (however low) and verification of form, in which they occur, with inspection of their use. That will focus on types of reactions of people using them (positive or negative) and a specification of this use, including a discrimination of polysemy which may occur here, too. (Obviously, each meaning is strictly distinguished by a particular intonation type. It seems, in so far, as the corpora used allow this, that there is a tendency for some of these idiomatic reactions to be used in specified types of spoken text, such as, for example, formalized texts, brief dialogues, etc.)

1. CORPORA

Several spoken corpora have been used, which have been, if need be, compared with a large written corpus. In survey:

Corpus	*Size*	*Time*	*Type of recording*
PMK	675.000	1988-1996	formal + informal
ORAL-F	1,691,000	2002-2007	informal
ORAL-S	4,785,000	2002-2011	informal
SYN2010	100,000,000	20th century	written

Prague Spoken Corpus (PMK) is the first of its kind, recorded in Prague between 1988-1996, being, basically, just like all the other spoken corpora, based on dialogues. It has been balanced by four sociolinguistic variables, taking in each case one of two values. Three of them are related to the speaker: gender (male - female), age (20 up to 35 - older than 35), education (higher - lower), and the fourth to the type of speech (formal and informal). By formal such situations are understood where the recording person asked general topical questions and the speaker replied to them in longer answers. The informal situations relate to topics that have not been influenced by anything, being parts of a free conversation between the speakers. The transcription aimed to capture the spoken language as accurately and clearly as possible, is close to the folkloristic transcription. The transcribed text was also manually tagged and lemmatized. As a part of this lemmatization there is also a marking of idioms added whereby idioms assigned a multi-word lemma just like other collocation types.

Later ORAL corpora (ORAL2006, ORAL2008 and ORAL2013, collectively referred to as ORAL-S, Lukeš et al. 2015) cover the whole Czech-speaking territory. They have been linked into ORAL-F corpus where an experimental automatic idiom identification has been tried. The tool use is based on a large Idiom Dictionary of Czech covering all types of Czech idioms (Kopřivová, Hnátková 2014). All recordings following the practice of PMK have been made only in informal situations. Thus conversations of only those speakers who knew each other having a friendly relationship were recorded in their natural environment. Due to the extension to all Czech regions a new category was added here, though not recorded, namely the dialectal area in which the speaker spent his or her childhood, because idiolects of the speakers had been formed there.

SYN2010 corpus is a synchronous written representative and balanced corpus, having 100,000,000 words (40% fiction, 27% professional literature, 33% newspapers). Some examples are from Intercorp, version 8.

2. A SAMPLE OF IDIOMS IN SPOKEN LANGUAGE

Two problems related to data had to be solved, first. It is virtually impossible to obtain a sufficient number of idiom items from spoken corpora, i.e. except for very few. Second, idioms based mostly on synsemantic (grammar) words, though very much in use, are not so numerous if compared to those based on nouns, verbs, etc. Hence, the decision went to two formally most prominent synsemantic words standing in the initial position of idioms, namely *A* and *No*. In this sense, the choice has been partly arbitrary.

The final selection has been made on the basis of some 75 machine-selected combinations of A-idioms and No-idioms and other collocations arriving at some 20 items with the maximum frequency (with the cut-off point around 5) from the corpus ORAL-F. This has been subsequently modified and suplemented from other corpora.

a *tak dále 195, a podobně 92, a když už 16, a co 15, a tak podobně 14, a hele 10, a sice 8, a co ty 7, a je to 5, a jo/jó 4, a hotovo 3, a tim pádem 3, a tak 2, a nazdar 1, a máš to 1, a sakra 1, a co potom 1, a jéje 1, a dost 1, a co teď 1, a tím to končí1.* (= 21 A-idioms altogether)

no *tak 1242, no ale 363, no jo 228, no né 70, no jó 60, no jistě 50, no prostě 46, no myslim 44, no dobře 42, no asi 41, no nevim 37, no právě 32, no dyť 29, no jasně 20, no samozřejmě 18, no rozhodně 14, no počkej 12, no třeba 12, no hele 11, no vidíš 11, no ano 10, no nic 8, no a co? 7, no co 7, no teda 6, no snad 6, no vopravdu 6, no fakt 5, no prosim tě 4, no proto 4, no vlastně 4, no přesně tak 3, no a 2, no hrozný 2, no prej 2, no podívej 2, no uvidíme 1, no dobrý 1, no copak 1* (39 No-idioms alogether)

The two types are somewhat different in their spoken use. While *A-idioms* often serve as (A) signal that the current topic is to be changed, to draw the listener´s attention to a forthcoming important question to be mentioned, or immediate reaction to what has just been said, cf.

a když už *mluvíme o tvých narozeninách, co takhle malý večírek? (And speaking of your birthday, what about a small party?)*

(B) other are used as an utterance closing formula at the end intimating that there is nothing important left to be said.

zajímalo mě všechno od sportu přes turistiku, divadlo, kino ***a tak dále*** *(I´d be interested in everything including hiking, theatre, cinema and so on).*

The *No-Idioms* always stand at the beginning of one´s speech turn, expressing various shades of consent, either direct, modified or moderate, though sometimes also an objection or surprise, etc. Being heavily laden with evaluative pragmatic aspects, these idioms are very common with a high frequency of use, cf.

To budeš mít co dělat – ***No právě****. a požadavky takový že se z toho zbláznim*

(That´ ll be a lot to do - That´s just it! requirements driving me crazy.).

3. AN ANALYSIS OF IDIOM BEHAVIOUR IN CORPUS

Czech data have been selected from three spoken corpora (described in 2 and 3), being based on frequency, primarily that of the Prague Spoken Corpus, supplemented by those from Oral corpora. The nature of spoken corpus texts is highly varied and their records reflect many situational shortcuts, allowing the student, in their diversity, to follow only some aspects.

Let us look briefly at a single No-idiom (*No a co?*) in a context. The speaker A being rather fond of or appreciative of fruit orange compote raises a reproach towards speaker B

because she has eaten it, probably without any advance notice. However, B does not consider what she did being such a big deal and dismisses it as unimportant (*So what*?), cf

A: sou tady i takoví, který normálně snědí kompot, jo? Mandarinkovej. No protože byl vod babičky a já sem prostě ...B: ***No a co****? A: Tak já kdybych věděla, že prostě ... tys sežrala kompot, jo?*

B´s reaction is both a refusal and disparagement of the reproach of A. The No-idiom used here standing in the immediate first position as a direct answer is negatively loaded and rather unfriendly and impolite.

Such being the nature of the idioms inspected where the context is rather complicated to be analysed, only few of the prominent features were followed and picked up.

Inspected formal features include:

- all the idioms are provided with English equivalents and,
- illustrated by a select corpus example (which is translated too, usually)
- type of text (where *Mon* stands for monologue, *Dia* stands for dialogue),
- place of the occurrence of each idiom in the sentence (*Beg.* standing for beginning position in the sentence and *End* standing for its end). Where necessary (in Dia) speakers have been distinguished by figures (*A:B*)

1. ... a tak dále (Eng. *likewise, in a similar way, and so on***)**

... vysvětloval, jak se fotí za umělého světla, jak se fotí za denního světla, práci s bleskem a tak dále

He explained how photos are taken in artificial light, in dayligt, how to use the flash and so on.

Function/Meaning: Mon, statement, indication of repeated continuation at the end of a list, Position: End

2. ... a podobně (Eng. *and in a similar way*)

Nejvíc času strávim s tim, sakra, kde co má bejt, jak má bejt a struktura tý databáze a podobně. Damn, I do spend most of time explaining where things are supposed to be, the database structure and so likewise.

Function/Meaning: Mon, statement, indication of continuation at the end of a list, Position: End

3. **a když už** (Eng. *and while, and just as*)

A když už o tom mluvíme – And while we are talking about it.

Function/Meaning: Dia, as a reaction mitigating the refusal that has preceded it or as a rather wavering reaction to the preceding statement, sometimes as an additional information, Position: Beg.

4. **a je to.** (Eng. *And the job´s done., That´s that*).

... je to tady utřený a nikdo nic nepozná. No, a je to. It has been cleaned up, nobody will see any diffrence. Ant that´s that!

Function/Meaning: Mon/Dia, statement of agreement and satisfaction. Position: End

5. ... a tak podobně (Eng. *likewise, in a similar way*)

Vy prý víte hodně o autech a tak podobně. You 're supposed to be some kinda expert in automobiles .

Filmy, fotky a tak podobně. Like movies and just pictures of different stuff.

Function/Meaning: Mon/Dia, statement, indication of continuation at the end of a list, Position: End

6. A hele (Eng. *Look!, Oh look!*)

Ale, už jsme skoro tam. Look, we´re nearly there.

Function/Meaning: Dia, reaction to to a changed situation, eg. arrival of someone, expression of astonishment or alerting someone, Position: Beg.

7. a co This combination, offered by an automatic analysis, corresponds to more than one idiom, in fact, some of them not being supported by corpus data, but known from elsewhere. These "extensions" include

7.A *a co/což potom/teprve* (Eng. *let alone, what about*)

A co potom ty oznámení. What about the notices?

7.B *A co teď?* (Eng. *What (to do) now?)*

Nemůžeme ho lokalizovat. A co teď? We can´t locate him. Now what?

7.C *A co tohle/tamhle/takhle?* (Eng. *How about this/there/this way?*)

7.D *A co ty/já...?* (Eng. *What about you/me?*)

A: ...no to je psychická nemoc, že jo. -B: No a co ty ? A:...it´s a kind of psychic illness, isn´t it? - B: And what about you?

Function/Meaning: Dia, reaction and question about the partner´s intention, Position: Beg.

7.E *No a co?* (Eng. *So what?*)

A: To je jako barák, chápeš? B: No a co ? Tobě by se to líbilo. A: It´s a kind of house, you see? B: So what? You´d like it.

Function/Meaning: Dia, reaction and disagreement with the objection, Position: Beg.

8. ... a tim pádem (Eng. *thus, therefore, this way*)

Jestli tě nevypletu, tak mě nikdy nechytíš ***a tím pádem*** *ani neuškrtíš, ne ? If I don't untangle you, you 'll never catch me to murder me, will you ?*

Function/Meaning: Dia/Mon, statement and reaction, introduction of a consequence, Pos centre.

9. A jó (Eng. *O yes*)

A:... je to na Freemusic cé zet. -B: A jó, to mě nenapadlo. A: it´s on Freemusic CZ. B: O yes, it did not occur to me.

Function/Meaning: Dia, reaction and realization of a forgotten fact, Position: Beg.

10. ... a hotovo. (Eng. *That´s that*!)

Dál už nepojedem a hotovo. We ain't going no further. This is it.

Function/Meaning: Dia, disapproving reaction to the wish of other people and an announcement of a final decision, Position: End.

11. ... a tak (Eng. *an so on*)

Pak proběhne velká ceremonie a tak. Then there's a big ceremony and so on.

Function/Meaning: Mon, statement, indication of repeated continuation at the end of a list, Position: End

12. No tak (Ang. *come on now, come on then; come, come, come on*!*, well*) having more than one function:

12.A *A: Co když se někdo přihlásí ? -B: No tak když se přihlásí jeden, tak*

A: What if someone registers himself? B: Well, if only a single person registers, then...

Function/Meaning: Dia, reaction and hesitant introduction of a possibility, Position: Beg.

12.B *A: Ty vypadaj skvěle, babi. -B: No tak. Ták. Mám ještě v ledničce. These look wonderful, gran. Come on, take some. I´ve got more in the fridge.*

Function/Meaning: Dia, statement, encouragement to do something, Position: Beg.

12.C *...no tak proč neni umytá skleníčka? Why then hasn´t this glass been rinsed?*

Function/Meaning: Dia, statement/question categorically demanding something, Position: Beg.

13. No jo (Eng, *I guess so, you see*)

A: Tak na to nechoď, já bych na to hned přestala chodit.-B: No jo, ale když na to nejsou skripta

A:Well, do not go there then, I wouldn´t in your place. B: Well you see, there´s no textbook.

Function/Meaning: Dia, reaction, agreement and explanation of an obstacle, Position: Beg.

14. No ale

No ale co ty? Anyway, how about you?

No ale to je přece pravda. And of course that´s true.

Function/Meaning: Dia, a mildly surprised reaction, disagreement, Position: Beg.

15. No jó (Eng. *I see, Well, but*)

No jó, tak on to byl jeden z těch obchodů! One of those shops. That explains it, then.

Function/Meaning: Dia, reaction suggesting one´s grasp of connection, Position: Beg

16. No né! (Eng. *Well, I never!; Oh no*) standing for two different idioms:

16.A *No né, musí se to natřít, že jo. Oh no, it has to be repainted, doesn´t it?*

Function/Meaning: Dia, reaction introducing a possibility or necessity, Position: Beg.

16.B *No né, tak by sme si šli třeba sednout do mekáče. Well we might go to a Macdonald and sit there.*Function/Meaning: surprised reaction, exclamatory comment, Position: Beg.

17. No právě! (Eng. *That´s just it, That´s the point*!)

A: *To je dobrý . -B: No právě. A já sem ti přinesla vochutnat, Terezko***.**

A: What a surprise! - B: That´s the point, I´ve come to let you taste it, Theresa.

Function/Meaning: Dia, statement and intensified consent with another statement, Position: Beg.

18. No ne! (Eng. *Well, I never!,)*

A: Váš NJ vstává brzo, viď, tak .. -B: No ne, myslim u NJ, že to mám dnes a denně.

Function/Meaning: Dia, statement and disagreement correcting or adding sth. to a previous one, Position: Beg.

19. No a co? (Eng. *What about it?*)

A: No protože byl vod babičky a já sem prostě ...-B: No a co?

Function/Meaning: Dia, reaction disagreeing with an objection and disparaging it, Position: Beg.

20. No co ?(Eng. *So what!*)

A: Já sem tě varovala. No. B: No co, to je na vodtažení lavice

Function/Meaning: Dia, reaction and disparagement of objection, Position: Beg.

21. No ano (Eng. *But of course! Right!*)

No ano, ty o tom nevíš. That´s right, you don´t know.

Function/Meaning: Dia, astonished reaction to realization that the other does not know the information, also an emphatic agreement, Position: Beg.

22. no tak co (Eng. So what!, cf 8E)

No tak co, co bych nešel. So what, what I would not go?

Function/Meaning: Acknowledgement of a reproof expressing at the same time "I don´t care" attitude, Dia, Position: Beg.

23. No nic. (Eng. *Too bad!*)

A: Ráno vstáváš, holčičko, ve čtyři! -B: No a co ? - A: No nic.

A: You´re getting up early in the morning, my girl! At four: - B: So what? - A: Nothing.

Function/Meaning: Dia, reaction to an unsuccessful negotiation, demand etc., Position: End

24. No jasně. (Eng. *Of course, All right*)

A: Už musím jít. B: No jasně, běž. A: I have to go. B: All right. You do that!

Function/Meaning: Dia, astonished reaction to realization that the other does not know the information, also an emphatic agreement, Position: Beg.

To sum this up briefly, it is evident that most idioms belong functionally to dialogues that are being used at the sentence beginning (though not entirely), are declarations and statements and only rarely they take the form of question. Semantically, they seem to be statements and/or reactions (which is difficult to distinguish sometimes), expressing, most specifically, disagreement and expressing sometimes, a kind of textual function, such as in Nos 1, 2 and 4. Due to the low number of spoken occurrences more exact or generalized analysis of the meaning is impossible here.

It must be added, however, that similar and functionally equal idioms found in the Czech Idiom Dictionary (Čermák et al. 2009) are much more numerous, due to the limited size of the corpora. Thus, to give you an insight, none of the following idioms is included

***A**: a co by ne?, a co potom?, a co teď?, a co takhle/tohle/..., a co víc, a hele, a just ne, a kdyby?, a ne snad?, a taky že jo, a to jo, a to zas ne, a to zas prr, a vono přece, a vůbec!*

No: no tak (included, but under multiple meanings, see 13), *no to jo, no teda!, no tohle!, no tak co?, no vida!*

4. CONCLUSIONS

It is hardly possible to offer any conclusions, given the double sample character of data, i.e. its choice being limited to the initial position, and the choice of A and No in that position followed, mostly by grammar words. It is obvious, that reactions expressed by those idioms included and inspected offer a number of various types of reactions that are, in the bulk of such idioms, a prevalent type, being, generally, limited to answers of what has preceded them. Most of these A- and No-idioms, however, have, at the same time, a linking function in the (dialogue) text, pointing to their precedents. As for their sentence function, most of these idioms may be viewed as either multiword Adverbials or Particles.

Functionally, the idioms in question are often heavily loaded pragmatically expressing an evaluative, i.e. a negative or positive attitude of the speaker. What that attitude is pointing to, is semiotic "IT" (the situation behind the preceding statement of somebody else, mostly), and, less obviously, to the speaker itself "I/Myself" and the speakers preferences along the positive-negative scale in general.

Naturally, idioms registered in the great dictionary of Czech Idioms (Čermák a kol. 2009) are different from those in corpora. It seems that in spoken language No-idioms are more frequent (some of which not being registered in the dictionary yet) this being a question to be resolved in future research in more data.

Acknowledgements

This study was written within the Programme for the Development of Fields of Study at Charles University, No. P11 Czech National Corpus, sub-programme Czech National Corpus.

References

ČERMÁK, F. ET AL, 2009. *Slovník české frazeologie a idiomatiky. Výrazy větné*. Praha: Leda.

ČERMÁK, F., 2006. Mluvené korpusy. In: F. Čermák and R. Blatná: *Studie z korpusové lingvistiky*. Praha: NLN, pp. 53-67.

ČERMÁK, F., 2008. Partikule, jejich syntagmatika a kumulace v mluvené češtině. In: *Čeština v mluveném korpusu*, eds. M. Kopřivová, M, Waclawičowá, NLN Praha, 2008, pp 63-74.

HNÁTKOVÁ, M., 2002.: Značkování frazémů a idiomů v ČNK s pomocí SČFI. *SaS* 63 (2), pp. 117-126.

HNÁTKOVÁ, M. AND KOPŘIVOVÁ, M., 2013: Identification of idioms in Spoken Corpora. In: K. Gajdošová and A. Žáková: *Proceedings of the Seventh International Conference Slovko*. Lüdenscheid, Germany, pp. 92 -99.

KOPŘIVOVÁ,M., 2008. Frazeologie v mluvených korpusech na základě PMK. In: *Čeština v mluveném korpusu*, eds. M. Kopřivová, M, Waclawičowá, NLN Praha, 2008, pp. 149-160.

LUKEŠ, D., KLIMEŠOVÁ, P., KOMRSKOVÁ, Z., KOPŘIVOVÁ,M. 2008. Experimental Tagging of the ORAL Series Corpora: Insights on Using a Stochastic Tagger. *TSD Proccedings*. (in print).

Corpora

ČERMÁK, F., ADAMOVIČOVÁ, A., PEŠIČKA, J. , ŠIMANDL, J., ŠONKOVÁ, J., SAVICKÝ, P., SMETANOVÁ, Z. *: Pražský mluvený korpus. Ústav Českého národního korpusu FF UK, Praha 2005*. Accessible at *http://www.korpus.cz*

KOPŘIVOVÁ, M., WACLAWIČOVÁ, M.: *ORAL2006: korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha 2006. Accessible at WWW: http://www.korpus.cz

WACLAWIČOVÁ, M. – KOPŘIVOVÁ, M. – KŘEN, M. – VÁLKOVÁ, L.: *ORAL2008: sociolingvisticky vyvážený korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha 2008. Accessible at WWW: http://www.korpus.cz

BENEŠOVÁ, L., KŘEN, M., WACLAWIČOVÁ, M.: *ORAL2013: reprezentativní korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha 2013. Accessible at http://www.korpus.cz

KŘEN, M. , BARTOŇ, T. , CVRČEK, V. , HNÁTKOVÁ, M. , JELÍNEK, T., KOCEK, J., NOVOTNÁ, R., PETKEVIČ, V., PROCHÁZKA, P., SCHMIEDTOVÁ, V., SKOUMALOVÁ, H.: *SYN2010: žánrově vyvážený korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha 2010. Accessible at: http://www.korpus.cz

ROSEN, A. AND VAVŘÍN, M.: *Korpus InterCorp – čeština, verze 8 z 4. 6. 2015*. Ústav Českého národního korpusu FF UK, Praha 2015. Accessible at: http://www.korpus.cz

Appendix

The Appendix gives tables with frequencies of the two idiom types researched in the corpora. The first part gives absolute frequency, the second shows frequencies normalized to millions of occurrences, so that a comparison is made possible. This comparison shows clearly domination of these idioms in the spoken language.

					IPM			
Corpus	ORAL-F	PMK	ORAL-S	SYN2010	ORAL-F	PMK	ORAL-S	SYN 2010
a tak	628	2	792	1348	237.30	2.36	133.17	11.08
a tak dále	175	195	269	1196	66.13	230.34	45.23	9.83
a co	173	15	2561	619	65.37	17.72	430.61	5.09
a jó/jo	157	4	13	75	59.33	4.72	2.19	0.62
a je to	73	5	1081	284	27.58	5.91	181.76	2.33
a podobně	55	92	139	2752	20.78	108.67	23.37	22.62
a co ty	39	7	90	88	14.74	8.27	15.13	0.72
a to je všechno	34	0	43	127	12.85	0.00	7.23	1.04
a hele	31	10	51	88	11.71	11.81	8.58	0.72
a hotovo	31	3	117	150	11.71	3.54	19.67	1.23
a tim pádem	29	3	83	445	10.96	3.54	13.96	3.66
a když už	28	16	32	821	10.58	18.90	5.38	6.75
a nazdar	12	1	23	9	4.53	1.18	3.87	0.07
a máš to	10	1	64	36	3.78	1.18	10.76	0.30
a sakra	10	1	16	101	3.78	1.18	2.69	0.83
a co potom	8	1	12	59	3.02	1.18	2.02	0.48
a sice	7	8	14	791	2.65	9.45	2.35	6.50
a jeje/ a jéje	7	1	6	36	2.65	1.18	1.01	0.30
a dost	7	1	12	25	2.65	1.18	2.02	0.21
a co teď	6	1	6	81	2.27	1.18	1.01	0.67
a je to tady	6	0	9	87	2.27	0.00	1.51	0.72
a tím to končí	5	1	10	150	1.89	1.18	1.68	1.23

					IPM			
Corpus	ORAL-F	PMK	ORAL-S	SYN2010	ORAL-F	PMK	ORAL-S	SYN2010
no tak	2516	1242	8367	2248	950.71	1467.11	1406.85	18.48
no ale	891	363	3737	229	336.68	428.79	628.35	1.88
no jo (no)	1948	228	4835	1455	736.09	269.32	812.97	11.96
no né/ne	267	70	901	364	100.89	82.69	151.50	2.99
no jó	128	60	4324	1455	48.37	70.87	727.05	11.96
no jistě	128	50	269	303	48.37	59.06	45.23	2.49
no prostě	120	46	368	131	45.34	54.34	61.88	1.08
no myslím	8	44	21	0	3.02	51.97	3.53	0.00
no dobře	132	42	302	684	49.88	49.61	50.78	5.62
no asi	78	41	288	16	29.47	48.43	48.43	0.13
no nevim (nevim)	39	37	233	58	14.74	43.71	39.18	0.48
no určitě	37	37	138	10	13.98	43.71	23.20	0.08
no právě	321	32	752	176	121.30	37.80	126.44	1.45
no dyť/vždyť	190	29	547	35	71.79	34.26	91.97	0.29
no jasně	577	20	1286	400	218.03	23.62	216.23	3.29
no samozřejmě	29	18	82	63	10.96	21.26	13.79	0.52
no rozhodně	6	14	12	3	2.27	16.54	2.02	0.02
no počkej	78	12	218	31	29.47	14.17	36.66	0.25
no třeba	21	12	82	23	7.94	14.17	13.79	0.19
no hele	43	11	121	18	16.25	12.99	20.35	0.15
no vidíš	116	11	196	2	43.83	12.99	32.96	0.02
no ano	21	10	121	227	7.94	11.81	20.35	1.87

no nic	55	8	104	339	20.78	9.45	17.49	2.79
no a co?	35	7	111	214	13.23	8.27	18.66	1.76
no co	26	7	58	186	9.82	8.27		1.53
no teda	26	6	63	69	9.82	7.09	10.59	0.57
no snad	8	6	35	25	3.02	7.09	5.88	0.21
no vopravdu	5	6	16	0	1.89	7.09	2.69	0.00
no fakt	17	5	47	24	6.42	5.91	7.90	0.20

PHRASÉOLOGIE ET TRADUCTION : PERSPECTIVE CONTRASTIVE À BASE D'UN CORPUS BILINGUE FRANÇAIS-ARABE TUNISIEN

Abdellatif Chekir
Institut Supérieur des Langues de Nabeul
Université de Carthage (Tunisie)

Résumé

Plusieurs expressions figées en français ont été transférées vers l'arabe tunisien et ont été adoptées par les locuteurs tunisiens. Elles appartiennent aux différentes catégories grammaticales. Certaines sont des traductions littérales alors que d'autres sont des calques hybrides puisqu'ils empruntent des lexèmes de la langue source. Cependant certaines expressions traduites sont fidèles aux patrons de départ et respectent les mêmes contraintes alors que d'autres s'en écartent pour se conformer aux normes de la langue cible.

1. INTRODUCTION

L'arabe dialectal parlé en Tunisie a subi plusieurs influences tout au long de son histoire. C'est ainsi qu'il a emprunté plusieurs lexèmes à l'italien, l'espagnol, le maltais, le turc. Mais la langue avec laquelle il a eu le plus de contact est le français. En effet, après plus de soixante-dix ans de présence française sur le sol tunisien, cette langue a laissé des traces indélébiles avec l'emprunt de beaucoup de mots mais également d'expressions figées qui ont enrichi l'arabe tunisien. Ces expressions ont été transférées par le procédé du calque qui consiste en une traduction littérale de leur sens particulier dans la langue de départ avec les ressources linguistiques de la langue emprunteuse. Cependant, en arabe tunisien le procédé de transfert est un peu différent de celui auquel on recourt en arabe standard puisqu'il permet fréquemment d'effectuer une double opération : on emprunte et le sens de l'expression dans L1 et certains mots français. Notre travail s'articule autour de plusieurs axes. Dans un premier temps nous présenterons l'arabe tunisien et le corpus sur lequel nous allons réfléchir puis nous esquisserons une typologie des expressions calquées pour dégager leurs spécificités. Nous vérifierons enfin la fidélité et l'écart de ces expressions par rapport au modèle d'origine.

2. PRÉSENTATION DE L'ARABE TUNISIEN, DU CORPUS ET DU PROCEDE DU CALQUE

2.1. Présentation de l'arabe tunisien

La langue officielle en Tunisie comme d'ailleurs dans tous les pays arabes est l'arabe standard qui se distingue par ses spécificités tant au niveau morphosyntaxique que sémantique. Mais chaque pays arabe dispose d'un dialecte particulier avec des caractéristiques qui lui sont propres.

L'arabe tunisien est le dialecte parlé en Tunisie. Il est légèrement différent des dialectes parlés dans les autres pays arabes essentiellement au niveau phonétique et lexical. Cette différence n'empêche pourtant pas l'intercompréhension entre ceux qui utilisent ces différents dialectes. Mais comme tout système linguistique, l'arabe tunisien dispose de plusieurs expressions figées qui lui sont propres comme :

Bass fil lu:za littéralement il a pété dans l'amande (il s'est enfoui)

Sab ezzit littéralement il a versé l'huile (il est arrivé)

Jaðrab ɤma:su fi sda:su littéralement il multiplie les cinq par les six (il réfléchit)

jtallaʕ fil ma: lis saʕda littéralement faire monter l'eau dans la pente (compliquer les choses)

jtallaʕ iddam lirra:s faire monter le sang à la tête (énerver)

Mais cela n'empêche que l'arabe tunisien peut être enrichi par des expressions qui viennent d'ailleurs et qui sont très fréquentes.

2.2. Présentation du corpus

Nous avons déjà élaboré un dictionnaire des calques français-arabe standard qui est en cours de publication et qui contient plus de deux milles expressions. Ce dictionnaire a été élaboré à partir de documents écrits constitués essentiellement de journaux tunisiens et de documents audiovisuels tels que certains débats dans les télévisions tunisiennes et la chaîne française France 24 en arabe. Par contre, pour notre corpus sur l'arabe tunisien, nous ne disposons d'aucun document écrit puisque ce dialecte n'a pas encore été grammatisé et nous ne disposons pas encore de dictionnaires et de grammaire pour le décrire ; de même, les conventions de son écriture n'ont pas encore été fixées. Par conséquent, nous avons collecté les expressions de notre corpus à partir des enquêtes de l'Atlas linguistique de Tunisie enregistrées dans les différentes régions de la Tunisie, des échanges verbaux dans les chaînes tunisiennes et de notre connaissance de ce dialecte. Mais une des conditions qui favorise la détection de ces calques est la parfaite connaissance et de L1 qui est la langue prêteuse et de L2 qui est la langue emprunteuse. Bref, une première enquête nous a permis de relever plus de mille exemples mais ce travail doit être poursuivi pour repérer tous les calques afin d'en élaborer un dictionnaire en arabe dialectal.

Mais peut-on parler de calque chaque fois que nous sommes en présence d'une expression figée traduite littéralement du français?

2.3. Calque et traduction littérale

Le calque désigne et le processus suivi par une expression et le résultat de ce processus. Mais toutes les expressions traduites n'ont par le même destin. En effet, beaucoup d'expressions sont traduites du français vers l'arabe mais cela n'empêche qu'elles n'accèdent pas toutes au statut de calque. En effet, certaines expressions restent au stade de la traduction littérale parce qu'elles n'ont pas atteint une telle fréquence d'emploi qui leur permette d'être assimilées dans le système de L2. Celles-ci sont vouées à l'oubli et à la disparition. D'autres, au contraire, sont d'un emploi très fréquent ; ce qui leur permet de ne plus être ressenties comme étrangères et d'être adoptées puisqu'elles ont suivi tout un processus qui leur a permis d'être baptisées comme calque. D'ailleurs, leur détection est très difficile et exige une attention très particulière qui permette de les dépister à travers un regard croisé sur les deux langues. Toutefois l'observation de ces expressions nous montre qu'elles obéissent à des moules différents.

3. TYPOLOGIE DES EXPRESSIONS CALQUÉES

3.1. Les patrons

L'analyse de notre corpus montre que des expressions figées de différentes catégories sont décalquées en arabe :

-des phrases figées :

C'est plus fort que moi	ʃaj ʔaqwa minni:
C'est toujours la même histoire	Di:ma nafs laħka:ja
C'est à prendre ou à laisser	Hiz willa ɤalli

-des adverbiaux

De gré ou de force	Bisja:sa willa bil qɔwwa
A la loupe	Bil mukabra
A chaud	ʕasɤa:na
En un clin d'œil	Fi ramʃet ʕi:n
A l'ombre de quelqu'un	ʕa:jiʃ fi ðillu

Ces expressions se présentent ici sous forme d'un groupe prépositionnel composé d'une préposition : à, de, en suivie d'un syntagme nominal.

-des noms figés

Rire jaune	Ðaħka safra:wijja
Chaude journée	Nha:r sɤu:n
Cervelle de moineau/ d'oiseau	mɔɤ ʕasfu:r
Dernière cartouche	Lkartu:ʃa liɤra
Coup de maître	Ðarbit mʕallim
Dos d'âne	Dos d'âne

Ces noms obéissent en général au moule nom-adjectif et nom complément du nom.

Cependant le verbe figé est la partie la plus productive. En effet, d'après notre corpus, 90 % des calques portent sur les verbes figés. Ce sont en général des expressions préfabrtiquées qui admettent les transformations au niveau de l'actualisation de la personne et du temps et qui peuvent s'adapter au système morphologique de la langue arabe où la dérivation se conforme aux exigences des langues chamito-sémitiques. En voici quelques exemples :

Se noyer dans un verre d'eau	jɔɤrik fi: ka:s ma:
Mettre le doigt sur la plaie	jħɔt sɔbʕu: ʕa ʒʒɔrħ
Attendre quelqu'un au tournant	Jistanna fi:h fiddu:ra
Tomber dans l'eau	Ṯa:ħ fil ma:

3.2. Les caractéristiques

Contrairement au calque en arabe standard qui se présente sous forme de traduction littérale du français vers l'arabe avec le transfert du sens opaque de L1 vers L2 qui est véhiculé par les ressources linguistiques propres à l'arabe, certains calques, en arabe tunisien, contiennent des emprunts c'est-à-dire que l'expression est traduite mais avec la reprise de mots français puisés dans l'expression de départ. C'est ce que Michel Chansou appelle l'emprunt hybride[55]. Cet emprunt partiel rapproche le calque du code switching pratiqué très couramment par les locuteurs tunisiens. C'est l'exemple de :

Tourner la veste	Qlib lfista
Renvoyer l'ascenseur	raʒaʕlu l'ascenseur
Perdre la boussole	ðajaʕ lbu:sla
Dernière cartouche	Lkartu:ʃa liɤra
Donner carte blanche à quelqu'un	ʕta:h carte blanche
Ne pas être du même calibre (qa:lib en arabe)	Mu:ʃ min nafs lkalibre
Le courant ne passe pas entre eux	L'courant ma: jitʕadda:ʃ mabina:thɔm

[55] Chansou M. « Calques et créations linguistiques », *META* vol. 29, n° 3, 1984, p. 281.

Partager le gâteau	ʃaqsam lgateau
Forcer la dose	Za:d fiddoza
En forme Calque de l'anglais *in form*	Fi: fu:rma
Placer, mettre la barre très haut	ʒallaʕ lba:ra ya:ser lfu:q
Jouer cartes sur table	ʔilʕab cartes sur table
Donner un chèque en blanc à quelqu'un	ʕta:h chèque en blanc

Toutefois ces mots sont souvent adaptés au système phonétique et morphologique de l'arabe tunisien : le phonème [v] de *veste* devient [f] en arabe tunisien alors que le phonème p dans *paquet* devient b. Par ailleurs, le [ə] dans les noms *veste*, *barre*, *ceinture* devient [a] en arabe tunisien qui est la marque distinctive du féminin dans cet idiome. Cette adaptation est une forme d'adoption de l'expression dans la langue d'arrivée, adaptation qui se conforme au mode d'assimilation de l'emprunt dans la langue d'accueil.

Nous remarquons également que, dans certains cas, l'arabe tunisien emprunte un mot simple alors que dans d'autres on emprunte le nom suivi de son expansion comme dans les exemples suivants :

-ʕta:h chèque en blanc

-ʔilʕab cartes sur table

- ʕta:h carte blanche

Cependant ces expressions sont l'apanage d'une classe de gens cultivés qui ont une connaissance poussée de la langue française.

Certaines expressions sont, par ailleurs, traduites et en arabe standard et en arabe tunisien mais de manières différentes et montrent la distinction qui existe entre le dialecte et l'arabe standard. C'est le cas de :

Le courant ne passe pas entre eux	Arabe dialectal : L'courant ma: jitʕadda:ʃ mabina:thɔm Arabe littéral : ʔattajja:ru la: jamɔrru bajnahuma :
Donner un chèque en blanc à quelqu'un	Arabe dialectal : ʕta:h chèque en blanc Arabe littéral : ʔaʕta:hu sak ʕala: baja:ð

Ces exemples montrent la différence entre le littéral et le dialectal. Celui-ci se caractérise par le recours fréquent aux expressions hybrides alors qu'en arabe littéral le recours à l'équivalent arabe est de rigueur. Mais ces exemples constituent une exception car, en général, les expressions sont traduites ou bien en arabe tunisien ou bien en arabe standard.

Nous remarquons également que certaines expressions traduites sont déjà des calques de l'anglais. C'est l'exemple de *donner le feu vert*, *en forme*, *gratte-ciel, guerre froide* qui sont devenues des calques en arabe tunisien à partir du français.

Mais est-ce que tous ces calques restituent la structure sémantico-syntaxique des expressions de départ ?

4. FIDÉLITÉ ET ÉCART

4.1. Fidélité

Comme nous l'avons remarqué à partir de notre corpus, les calques reproduisent les mêmes moules que les expressions de départ, ils restituent généralement la structure syntaxique d'origine et par conséquent ils obéissent aux mêmes contraintes spécifiques à l'expression initiale. C'est l'exemple des phrases figées qui sont réfractaires à toutes formes de transformation ou de substitution, des verbes figés qui reprennent la structure syntaxique verbe-complément :

Tourner la veste qlib lfi:sta

Peser ses mots ju:zin kla:mu:

C'est également le cas des adverbiaux qui rétablissent la structure préposition-nom :

A la loupe	Bil mukabra
A chaud	ʕasχa:na

Il en est de même pour les noms figés qui se conforment au patron nom-adjectif ou nom complément du nom

Dernière cartouche	Lkartu:ʃa liχra
Coup de maître	ðarbit mʕallim

Toutefois beaucoup d'expressions s'écartent par rapport au modèle de départ.

4.2. Écart

Ces écarts peuvent se situer aussi bien au niveau syntaxique que sémantique

4.2.1. Écart au niveau syntaxique

Nous constatons à travers certains exemples que les expressions ne gardent pas le même ordre des mots que les expressions de départ comme nous pouvons le vérifier à partir des exemples suivants :

Dernière cartouche	Lkartu:ʃa liχra
Chaude journée	Nha:r sχu:n

En effet, les expressions françaises se caractérisent par l'ordre adjectif-nom alors qu'en arabe, elles adoptent l'ordre nom-adjectif qui est l'ordre canonique en arabe tunisien. C'est ce que Lerat appelle le composé décroisé.[56]

Il en est de même pour les expressions construites selon le moule nom-complément du nom car en français le complément du nom est introduit par une préposition alors qu'en arabe il s'emploie directement sans joncteur comme le montre les exemples suivants :

Cervelle de moineau/ d'oiseau	mɔχ ʕasfu:r
Coup de maître	Ɗarbit mʕallim

Ces écarts s'expliquent par les spécificités syntaxiques spécifiques à l'arabe tunisien auxquelles doivent se conformer les calques. Nous pouvons parler d'un calque adapté à l'arabe tunisien.

Nous constatons également un écart au niveau du nombre. En effet, dans beaucoup de cas, le pluriel devient singulier en arabe. Cet écart est perceptible à travers les exemples suivants:

Peser ses mots	ʃu:zin kla:mu
Lécher les bottes à qqun	ʃilħislu fi sabba:tu/ jilħas
S'arracher les cheveux	qattaʕʃaʕru :
Mettre des bâtons dans les roues	ħɔt laʕsa fil ʕiʒla

Nous constatons parfois une troncation de l'expression lors du transfert en arabe.

Lécher les bottes à qqun	ʃilħislu fi sabba:tu/ jilħas
Monter au créneau	saʕʕad

Dans ces deux cas l'expression est réduite au verbe en arabe.

[56] LERAT, P. (1988), « Les internationalismes dans les langues romanes », *Cahiers d'études hispaniques médiévales : Hommage à Bernard Pottier*, Lyon : ENS Éditions, annexe, 7(2), p. 487.

4.2.2. Écart sémantique

Ce transfert ne restitue pas toujours l'expression de départ comme le montrent ces expressions :

Ne pas remuer un cil	Ma: jħarrakʃ ʃaʕra
Raccrocher les crampons	ʕallaq is sabbat
Crever l'abcès	ifqaʕ iddimma:la

cil devient ʃaʕra (cheveu) dans le premier exemple, *crampons* devient sabbat (soulier) dans le deuxième exemple et *abcès* devient dimma:la (bouton) dans la troisième expression. A travers ces expressions décalquées nous constatons un passage de l'hyponyme à l'hyperonyme en arabe tunisien.

Mais il arrive que la traduction trahisse l'expression de départ comme l'attestent ces deux exemples :

Je ne suis pas dans mon assiette	Mani:ʃ fi saħni
Prendre une veste	Kla: kabbu:t

En effet, *assiette* est pris dans le sens de « pièce de la vaisselle » et non comme « état d'esprit, disposition habituelle » sens véhiculé par l'expression. De même, *veste* devient *manteau* dans l'exemple arabe.

5. CONCLUSION

Même si le calque réinvestit certains mécanismes de l'emprunt il en est une forme atténuée, facilement assimilable dans la langue d'arrivée. En effet ces expressions calquées ont été intégrées dans le système linguistique de l'arabe tunisien à tel point que l'on oublie leur origine étrangère surtout si le calque n'est pas doublé d'un emprunt comme dans le cas des calques hybrides. Ces expressions calquées gardent également les mêmes contraintes qu'en français et peuvent permettre de reconnaitre une expression préfabriquée en français pour ceux qui connaissent les deux langues. Il faut enfin noter que malgré leur caractère intraduisible qui constitue un critère définitoire des expressions figées, elles sont traduites en arabe tunisien et constituent ainsi une source de renouvellement et d'enrichissement du dialecte tunisien à l'instar de tous les procédés de néologie.

Bibliographie

CHANSOU M., 1984, « Calques et créations linguistiques », Méta vol. 29, n° 3, p. 281-284. CHEKIR A., 2014, « Traduction et calque en arabe : des collocations aux expressions figées », *Traces*, p. 75-88.

CHEKIR Abdellatif, « Phraséologismes et traduction en arabe : pour un dictionnaire bilingue des calques » (à paraître dans les actes du colloque Europhras 2014)

GROSS, G., 1996, *Les expressions figées en français : noms composés et locutions*, Ophrys. HOBEIKA-CHAKROUN, F., 2010, « Les collocations arabes intensives N+Adj dans deux romans Les Filles de Ryad et l'Immeuble Yacoubian », *Revue Interdisciplinaire "Textes & contextes*, n° 5 .

LERAT, P. (1988), « Les internationalismes dans les langues romanes », *Cahiers d'études hispaniques médiévales : Hommage à Bernard Pottier*, Lyon : ENS Éditions, annexe, 7(2), p. 483-491.

MEJRI S.,1997, *Le figement lexical. Descriptions linguistiques et structuration sémantique*, Publications de la Faculté des Lettres de la Manouba.

MEJRI, Salah, « L'idiomaticité, problématique théorique » in MEJRI S., (dir), *L'espace euro-méditerranéen : Une idiomaticité partagée,* Tunis, Ceres, p.231-243, 2004.

A NEW SCORE TO CHARACTERISE COLLOCATIONS: LOG-R IN COMPARISON TO MUTUAL INFORMATION

Itsuko Fujimura
Nagoya University (Japan)
fujimura@nagoya-u.jp

Shigenobu Aoki
Gunma University (Japan)
aoki@si.gunma-u.ac.jp

Abstract

This paper proposes a new score named Log-r as a simple measure for calculating the strength of association between the constituent words of bigrams and argues that Log.r is more appropriate than Mutual Information for characterising collocation types. Arguments are based on the visualisation of one million English bigrams taken from a corpus of 1.1 billion words and of 0.4 million French bigrams taken from a corpus of 0.1 billion words. A three-dimensional analysis of each bigram will be made with its Log-r, its logarithmized frequency, and vocabulary level of its constituent words. Transparent typological study of collocations can be conducted using this procedure, which is only based on the frequency of words and bigrams, Pearson's *r* and Zipf's law.

1. INTRODUCTION

Linguistic research on collocations based on large-scale corpora is flourishing. A collocation is a string of two or more words that frequently co-occur.

There are various types of collocations, and there is a need to categorise and describe the characteristics of each. Currently, however, there is no generally accepted typological framework. Terms such as 'compound word', 'phrase', 'fixed expression', 'collocation', 'idiom', 'lexical bundle', and 'multi-word unit' are used without being given distinctive definitions.

The frequency of co-occurrence and the degree of association between the constitutive words have been recognised as basic properties characterising collocation types (Ellis, 2012; Evert, 2009; Wray, 2012). While frequency is easy to understand and measure, strength of association is more complex; it is referred to using different names ('degree of association', 'degree of compositionality', 'degree of fixedness', 'degree of coherence', etc.) and is not measured in a unified fashion. Furthermore, while Mutual Information (MI) is often used as a method to measure strength of association, various other methods have been proposed

(Pecina, 2010), and research on this topic is still developing (Bybee, 2010; Evert, 2009; Gries, 2013).

This paper proposes a new score named Log-r as a simple measure for calculating the strength of association between the constituent elements of two-word collocations and argues that Log-r is more appropriate than MI for describing collocation types.

For the sake of simplicity, this paper addresses only bigrams –sequences of two words. A collocation is a bigram in which the words are more or less habitually associated; it includes all of the aforementioned terms ('compound word', 'lexical bundle', 'idiom', etc.).

1.1. A Frequency- and Strength-based Typological Model: Wray (2012)

There are various types of word strings that fall under the above-mentioned definition of 'collocation'. Wray (2012: 241) proposes the diagram in Figure 1 as a model for comprehensively expressing some of these.

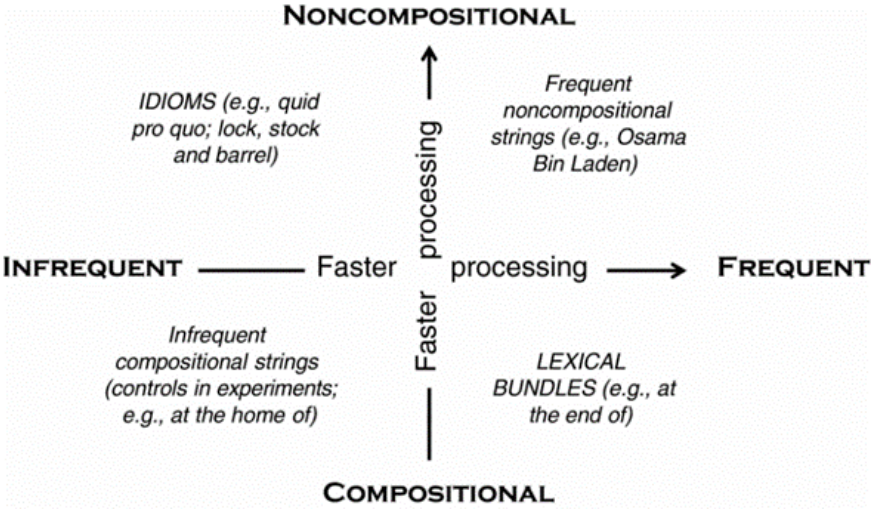


Figure 1. Wray's (2012) Typological Model of Collocations

The vertical axis in Figure 1 represents the level of compositionality, and the horizontal axis, the frequency. The compositionality is the same concept as the strength of association between constitutive elements. An exemplary type of word sequences is shown in each quadrant. Word sequences that often function as single words and appear frequently (e.g. *Osama Bin Laden*) are positioned in the top-right, first quadrant; word sequences that often function as single words but appear infrequently are in the top-left, second quadrant (idioms, e.g. *quid pro quo*); word sequences that generally do not function as single words, appear infrequently, and cannot be called collocations are in the bottom-left, third quadrant (e.g. *at the home of*); and word sequences that appear frequently but generally do not function as single words are in the bottom-right, fourth quadrant (lexical bundles, e.g. *at the end of*).

Examples of bigrams of each exemplary type are as follows: Quadrant 1 – *White House*, *Hong Kong*; Quadrant 2 – *lingua franca, bovine spongiform*; Quadrant 3 – *pink roses*, *familiar enough*; Quadrant 4 – *I am*, *of the*.

1.2. Problems with the MI score

MI is one of the most frequently mentioned methods for measuring the strength of association between constituent elements. Church & Hanks (1990:23) propose MI as an 'association ratio', Ellis (2012:28) and Hunston (2002:71) introduces it as measure of the strength of association, and A Glossary of Corpus Linguistics describes it as follows:

> Mutual information: (…) In corpus linguistics it is often used as a measure of the strength of a collocation between two words. (Baker, Hardie, & McEnery, 2006: 120)

However, it is hard to say that MI can be relied upon as a measure of collocation strength. Firstly, based on English-language newspaper corpora introduced in Table 1 below, the MI value of *Hong Kong* (12.8) was lower than that of *Jacqueline Onassis* (13.9); also the MI value of *human rights* (11.0) was lower than that of *human societies* (12.3). Intuitively, one would think that the association between the two words in *Hong Kong* is stronger than that of those in *Jacqueline Onassis*, and that of those in *human rights* is stronger than that of those in *human societies*. Secondly, as the developers of MI have pointed out from the beginning, a problem with this measure is that low-frequency bigrams are overvalued (Church & Hanks, 1990:24) and it has become customary to exclude them from measurements. However, this feature has not yet been adequately explained. Thirdly, the fact that various measures are being proposed to calculate collocations (Pecina, 2010, Gries, 2012) and that there are considerable discussions around them (Evert, 2009; François & Manguin, 2006; Gries, 2013), indicates itself that MI does not satisfy the conditions for it to be admitted as an appropriate indicator of collocation strength.

In this study, we will propose the use of Log-r instead of MI to measure the strength of association between the constituent elements of bigrams. In terms of Wray's model (Figure 1 above), Log-r would be the vertical axis scale.

2. LOG-R PROPOSAL

2.1. Definition

As a measure expressing the strength of association between two words, we propose Log-r which is a common logarithm of the correlation coefficient *r* that expresses the attribute correlation of two variables (word x and word y). The Pearson's product-moment correlation coefficient is defined as:

$$r = \frac{cov_{xy}}{\sigma_x \sigma_y} \quad (1)$$

This study assumes a Poisson distribution and uses the approximation formula (2), where f_{xy} is the frequency of the successive words xy, and f_x and f_y are respectively the frequency of word x and word y. A Poisson distribution can be assumed when large-scale data are used and the frequency of words xy is low.

$$r \doteq \frac{f_{xy}}{\sqrt{f_x f_y}} \quad (2)$$

Log-r is therefore defined as:

$$\text{Log-r} = log_{10} \frac{f_{xy}}{\sqrt{f_x f_y}} \quad (3)$$

2.2. Examples

Log-r's values are less than or equal to 0. Typical values and English and French examples of each are shown below.

- Log-r = 0, *r* = 1, e.g. lingua franca (en), statu quo (fr)

lingua franca: 100% of the occurrences of word x (*lingua*) are co-occurrent with 100% of the occurrences of word y (*franca*).

- Log-r = -1, *r* = 0.1, e.g. *apple pie* (en), *sud ouest* (fr)

apple pie: 10% of the occurrences of word x (*apple*) are co-occurrent with 10% of the occurrences of word y (*pie*).

- Log-r = -2, *r* = 0.01, e.g. *medal winner* (en), *gare SNCF* (fr)

medal winner: 1% of the occurrences of word x (*medal*) are co-occurrent with 1% of the occurrences of word y (*winner*).

- Log-r = -3, *r* = 0.001, e.g. *earlier offer* (en), *poids trop* (fr)

earlier offer: 0.1% of the occurrences of word x (*earlier*) are co-occurrent with 0.1% of the occurrences of word y (*offer*).

- Log-r = -4, *r* = 0.0001, e.g. *no there* (en), *pas il* (fr)

no there : 0.01% of the occurrences of word x (*no*) are co-occurrent with 0.01% of the occurrences of word y (*there*).

The Log-r value is 0 when two words are strongly associated and have come to function as a single word. Log-r is -4 when there is absolutely no habitual association between the two words. Between 0 and -4 there is a continuum of bigrams with different strengths of association.

2.3. Mathematical Characteristics of Log-r

The characteristics of Log-r can be summarised as follows. First, Log-r is a simple statistic, nothing more than a logarithm of Pearson's *r*. Like frequency of occurrence, it is robust and highly transparent. Second, Log-r is an appropriate statistic for linguistic phenomena. The frequency of words or bigrams is known to be an extremely wide-ranging statistic explained by Zipf's law (Baroni, 2009; Zpf, 1949). It is roughly calculated using the formula: 'occurrence rate (%) = 10/rank'. The most frequent word appears as 10% of the total number of words, the 10th most frequent word appears as 1% of the total, the 100th most frequent word appears as 0.1% of the total and so forth. Using a logarithm makes it easy to handle values and to grasp phenomena intuitively, by enabling us to visualise the

phenomena, as in Wray's diagram shown in Figure 1. Third, Log-r can measure all bigrams in a corpus, because the approximate value of *r* coming from formula (2) does not take a negative value, unlike the definitional value of *r* coming from formula (1). A logarithmic transformation is possible only for positive value. Fourth, Log-r is easy to calculate, because, compared to the definitional formula (1), the approximation formula (2) is simplified. However, there are restrictions when one uses this latter. Caution is necessary, as the difference between definitional formula-based values and approximation formula-based values grows greater as the value of f_{xy} increases and as the scale of the corpus decreases.

It should be finally noted that Log-r is not an entirely novel measure. Among the 82 measures introduced in Patina (2010), Pearson's chi-square test, z-score, Pearson, and Phi share fundamental characteristics with Log-r.

2.4. Comparison with MI

MI is frequently mentioned as a measure of a bigram's strength of association (Ellis, 2012; Evert, 2009; Gries, 2012; Hunston, 2002). The MI definitional formula is (4), and the MI approximation formula is (5). The latter is used in practice.

$$\mathrm{MI} = \log_2 \frac{P_{xy}}{P_x P_y} \tag{4}$$

$$\mathrm{MI} = \log_2 \frac{f_{xy}\, N}{f_x f_y} \tag{5}$$

The essential difference between the MI and Log-r calculation formulas is that in the former (5), the denominator is $f_x f_y$, whereas, in the latter (3), it is square rooted, $\sqrt{f_x f_y}$.

As the formulas clearly show, MI does not simply express the strength of association between x and y. Even if the total number of words in the corpus N is the same and the $f_x : f_y : f_{xy}$ ratio is the same, the value of MI changes depending on the value of f_{xy}.[57] The greater f_{xy}, the smaller the value of MI, and the smaller f_{xy}, the greater the value of MI. On the other hand, in the case of Log-r, if the ratio $f_x : f_y : f_{xy}$ is the same, the Log-r value stays the same regardless of the value of f_{xy}. We illustrate this clearly below on the basis of examples.

3. DATA

Details of the data used in this study can be found in Table 1. Our corpora were English- and French-language newspapers. From the main text of articles of these newspapers, we manually extracted 1.04 million English bigrams that appeared 54 times or more and 400,000 French bigrams that appeared 20 times or more. Then we calculated the occurrence frequency, Log-r, and MI of each. A morphological analysis was not carried out on the data; bigrams were presented in the form that they appeared in texts. We distinguished between upper and lowercase letters, but we did not take into account the presence or not of an apostrophe or hyphen between two words.

[57] The value of MI changes also depending on the size of the corpus, contra Hunston (2002 :73).

Language	No. of Bigrams	Total Number of Words	Newspaper	Corpus Distributor and Name
English	1.04 million (54 tokens or more)	1.1 billion words (Only main text of articles)	*L.A. Times-Washington Post* (1994-1998), *New York Times* (1994-1998), *Reuters Financial News* (1994-1996), *Reuters General News* (1994-1996), *Wall Street Journal* (1994-1996), *Associated Press Worldstream* (1994-1998)	LDC ・North American News Text Corpus ・North American News Text Supplement
French	400,000 (20 tokens or more)	118 million (Only main text of articles)	*Le Monde* (1988, 1994, 1996, 1999, 2000, 2006)	ELRA ・Le Monde

Table 1. Data

4. SCATTERPLOT-BASED COMPARISON OF LOG-R AND MI

4.1. One-dimensional Model

Table 2 shows the Log-r and MI of five English and French bigrams in our database. In these examples, Log-r and MI do not contradict each other in their assessment of the bigrams' strength of association.

Strength of association	English			French		
	bigram	Log-r	MI	bigram	Log-r	MI
+	*lingua franca*	-0.01	22.5	*statue quo*	-0.00	16.7
↕	*apple pie*	-1.01	13.8	*frère aîné*	-1.03	11.9
	medal winner	-2.00	8.0	*gare SNCF*	-2.04	8.1
	earlier offer	-3.00	2.3	*poids trop*	-3.01	2.3
-	*no there*	-4.04	-3.4	*pas il*	-4.01	-5.9

Table 2. One-dimensional Display of Log-r and MI

However, it is easy to find contradictions between the two measures. For example, in English, according to Log-r, *White House* (-0.23) is between *lingua franca* and *apple pie,* whereas according to MI, it is lower (11.1) than *apple pie*. Similarly, in French, according to Log-r, *sans doute* (-0.53) is between *statu quo* and *frère aîné*, whereas according to MI, it is lower (8.8) than *frère aîné*. In both instances, it is not easy to judge which measure is superior using a one-dimensional model.

4.2. Two-dimensional Model

A two-dimensional model brings into relief the differences between Log-r and MI. We placed Log-r and MI on the vertical axes of Figures 2 and 3, respectively, and put $\log(f_{xy})$ on the horizontal axis. We used logarithm for the frequency also, which allowed us to

create a diagram like that of Wray (2012) and to examine the phenomenon through a visual representation.

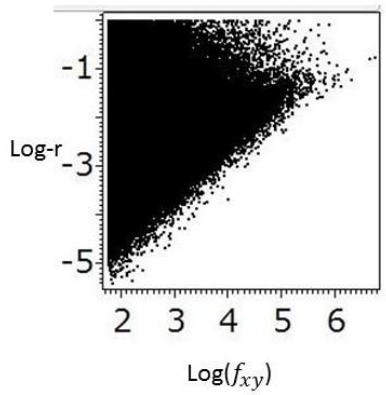


Figure 2. Log-r and Log(f_{xy})

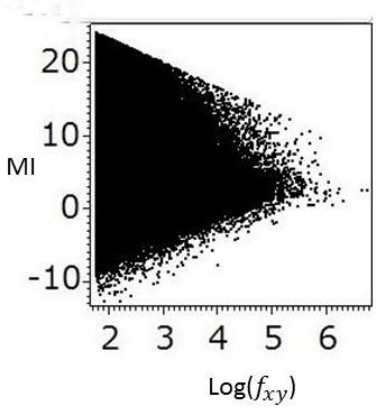


Figure 3. MI and Log(f_{xy})

It is clear that Figures 2 and 3 show variations of the same shape. In the Log-r-based Figure 2, the upper area extends horizontally. In the MI-based Figure 3, it descends toward the right, even though there is no reason to expect that bigrams with higher frequency will be less strongly associated and that bigrams with lower frequency will be more strongly associated. The MI formula results in an unnatural shape. Figure 3 is a transformation of Figure 2.

Although the bottom area of the scatter plot rises up to the right as frequency increases in both graphs, this is a natural increase that reflects actual language use. As the number of words in actual language use, as well as that in the corpus (sample) is limited, the coincidental co-occurrence of the bigrams' constituent elements decreases as their frequency increases (See Table 3 for details regarding this point.) In Figure 1, which we used as our model, the graph forms a square. However, in fact, the difference between *at the end of* in fourth quadrant and *at the home of* in third quadrant should be one of both frequency and strength of association.

4.3. Position of Examples

Next, we will analyse the bigrams positioned on the scatter plots.

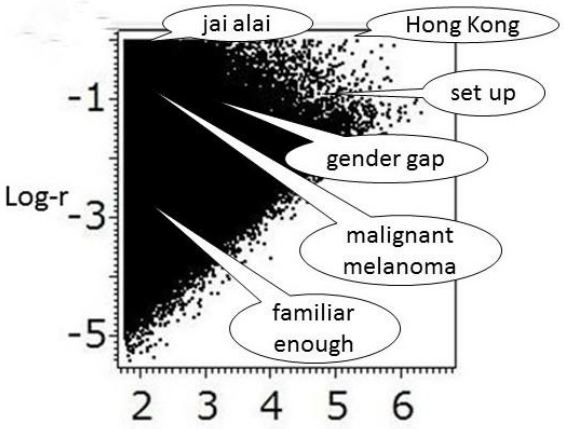


Figure 4. Log-r and Log(f_{xy}) with examples

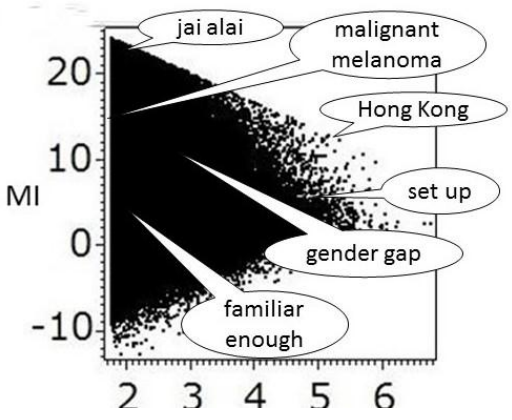


Figure 5. MI and Log(f_{xy}) with examples

A comparison of the placement of bigrams in Figures 4 and 5 reveals more clearly that the MI scatter plot (Figure 5) is a transformation of the Log-r scatter plot (Figure 4). Whereas in Figure 4, *jai alai* (ball game of Basque origin) and *Hong Kong* are located at the

same high position on the y-axis, in Figure 5, the latter is considerably lower than the former. Intuitively, one would think that the two collocations differ in frequency but not of strength of association, and Figure 4 supports this. In the MI-based Figure 5, the placement of *Hong Kong* is below *malignant melanoma* and near *gender gap*. Thus, in the results based on MI, *Hong Kong* is undervalued. It can be seen from the graph shape that the cause of this undervaluation is its high frequency.

4.4. MI's Structural Problem

As we already mentioned, the developers of MI have acknowledged that it overvalues bigrams that have a low frequency.

> Since the association ratio becomes unstable when the counts are very small, we will not discuss word pairs with f(x, y) < 5. (...) For the remainder of this paper, we will adopt the simple but arbitrary threshold and ignore pairs with small counts. (Church & Hanks, 1990: 24)

However, as can be seen in Figure 5, MI has at the same time the structural problem that it undervalues high-frequency bigrams. This problem does not exist for Log-r, as its evaluation of bigrams is not influenced by frequency. Therefore, it appears that, compared to MI, Log-r is more appropriate for accurately describing the strength of association of bigrams.

4.5. The Universality of the Shape of Log-r and MI-based Scatter Plots

Figures 6 and 7 below present scatter plots of 400,000 French bigrams based on the data listed in Table 1. In Figure 6, Log-r and log(f_{xy}) are given on the vertical and horizontal axes, respectively; and in Figure 7, MI and log(f_{xy}) are on the vertical and horizontal axes, respectively. Despite differences in language, number of corpus words, number of bigrams, and the lower limit of the frequency of occurrence, the shapes of the Log-r scatter plots in Figures 6 and 2, as well as that of the MI scatter plots in Figures 7 and 3, are similar to the extent that it is difficult to tell them apart.

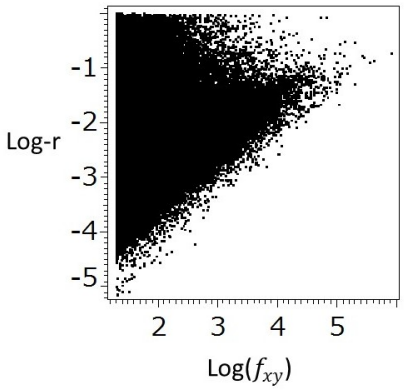


Figure 6. Log-r and log(f_{xy}) in French

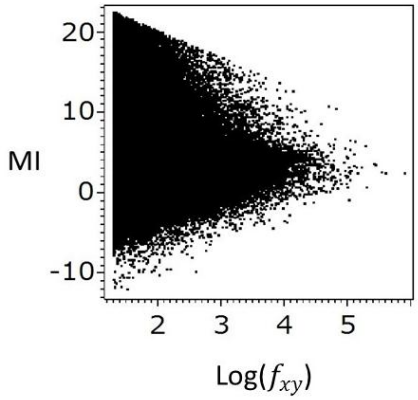


Figure 7. MI and log(f_{xy}) in French

If one uses bigrams extracted from the entirety of a naturally formed large-scale text, the shape of Log-r / log(f_{xy}) scatter plots will be the same, as will that of MI / log(f_{xy}) scatter plots. They can be said to have a universality that transcends individual languages. In other words, the difference between Log-r and MI is universal.

Our analysis of the scatter plots made certain points clear: MI is a measure that merges strength of association and frequency such that, as frequency increases, the MI value decreases. On the other hand, Log-r is a simple measure that only reflects strength of association. By combining Log-r and the logarithm of frequency: log(f_{xy}), it is possible to show clearly the characteristics of bigrams.

5. INFERRING TYPES BASED ON SCATTER PLOT POSITION

Lastly, we will attempt to create a typology of bigrams considering their position in scatter plots. We will do so based on the strength of association between bigram words: Log-r, the logarithm of bigram frequency: $\log(f_{xy})$ and, as well as the vocabulary level of the words constituting the bigram. Below, we will assume that the frequencies of word x and word y: f_x and f_y are the same for the sake of simplicity.

Generally, when given the total number of words in a corpus N and a word's frequency f_x in this corpus, following the approximate formula based on Zipf's law: '$\text{rank} = 10/\text{occurrence rate (\%)}$' (See 2.3 above), the frequency f_x can be converted into its frequency rank within the corpus. It can be assumed that in the case of a large corpus, a word's frequency rank will match its vocabulary level. These calculations enable one to grasp intuitively the word type listed in Table 3 like 'unknown', 'usual', 'functional'.

Table 3 shows the results of a simulation for each part of the scatter plot when the total number of words in the corpus N was assumed to be 1 billion. Figure 8 illustrates this as a scatter plot (as in Figures 2 and 4).

The arrow going from the top left to the bottom right in Figure 8 shows the vocabulary levels of the constituent words. Bigrams comprised of unfamiliar words are found at the top left, and highly familiar words are found around the bottom right. At the very top left are bigrams comprised of words at the 1 million word level (labelled 'unknown' in Table 3) that are decidedly atypical and unknown. At the bottom area appear bigrams comprised of words at the 10 or 100 word level (labelled 'hyper-functional' or 'functional').

According to the above-proposed criteria, as an illustrative example, we show in Table 4 the characteristics of the six English bigrams from Figure 4. The number of +'s indicates the level of each measurement. Both *Hong Kong* and *jai alai* have strongest association between their constituent words. The most significant difference between them consists of their frequency. Similarly, both *set up* and *malignant melanoma* have the same degree of strength of association. Their difference lies in their frequency and their constituent words' vocabulary level (Word type in Table 4). *Set up* is a frequent collocation constituted of 'functional' words, while *malignant melanoma* is an infrequent collocation constituted of 'rare' words.

Log-r	Log(f_{xy})	f_x	Occurrence rate: f_x/N (%)	Rank/ Vocabulary level (x)	Word type (x)	cf. MI
0	2	100	0.00001	1,000,000	Unknown	23.3
-1	2	1000	0.0001	100,000	Rare	16.6
-2	2	10,000	0.001	10,000	Usual	10.0
-3	2	100,000	0.01	1,000	Basic	3.3
-4	2	1,000,000	0.1	100	Functional	-3.3
-5	2	10,000,000	1	10	Hyper-functional	-10.0
0	3	1,000	0.0001	100,000	Rare	19.9
-1	3	10,000	0.001	10,000	Usual	13.3
-2	3	100,000	0.01	1,000	Basic	6.6
-3	3	1,000,000	0.1	100	Functional	0.0
-4	3	10,000,000	1	10	Hyper-functional	-6.6
0	4	10,000	0.001	10,000	Usual	16.6
-1	4	100,000	0.01	1000	Basic	10.0
-2	4	1,000,000	0.1	100	Functional	3.3
-3	4	10,000,000	1	10	Hyper-functional	-3.3
0	5	100,000	0.01	1,000	Basic	13.3
-1	5	1,000,000	0.1	100	Functional	6.6
-2	5	10,000,000	1	10	Hyper-functional	0.0
0	6	1,000,000	0.1	100	Functional	10.0
-1	6	10,000,000	1	10	Hyper-functional	3.3

Table 3- Simulation of Each Part of a Log-r / Log(f_{xy}) Scatter Plot

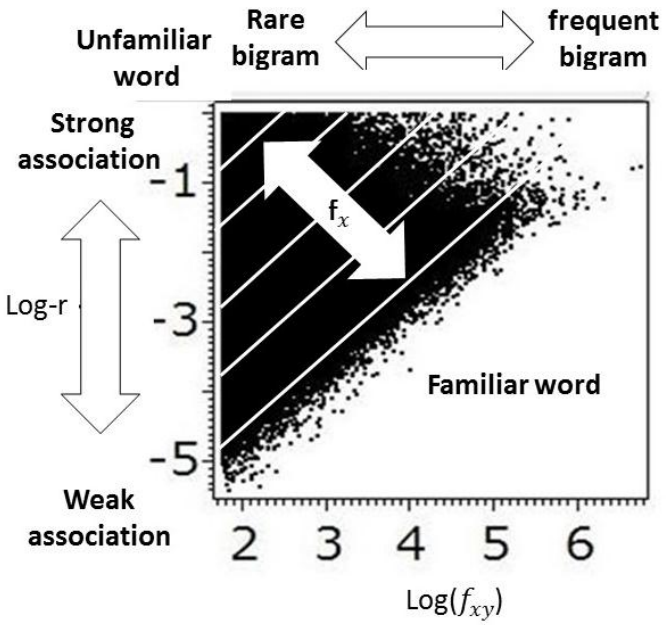


Figure 8. Parts of a Log-r / Log($\boldsymbol{f_{xy}}$)-based Scatter Plot

Bigram	Strength of association: Log-r	Frequency of bigram : Log(f_{xy})	Word type (x)	cf. MI
Hong Kong	+++++ -0.01	+++++ 5.16	Basic	12.82
jai alai	+++++ -0.02	++ 2.18	Unknown	22.63
set up	++++ -1.01	+++++ 4.90	Functional	7.04
malignant melanoma	++++ -1.02	++ 2.04	Rare	16.73
gender gap	++++ -1.20	+++ 3.08	Usual	11.86
familiar enough	++ -2.97	++ 2.05	Basic	3.47

Table 4. Characterisation of Six Bigrams

We can continue to analyse bigrams from these three points of view and make a typological study of collocations. MI does not allow us to describe collocation types.

6. CONCLUSION

In this paper, we proposed a new measure, Log-r, as a straightforward measure for calculating the strength of association between bigram's constituent elements and showed that Log-r is more useful than MI for describing collocation types.

MI measures both frequency and strength of association at the same time, whereas Log-r measures only strength of association. Therefore, even if MI is a practical tool for finding collocations that are infrequent but are composed of strongly associated words, it is not capable of correctly evaluating the strength of association between two words.

By measuring degree of association of bigrams using Log-r, a simple statistic, and combining it with other simple statistics like frequency of occurrence and vocabulary level of constituent words based on Zipf's law, we can describe and explain, through visual representation, different collocation types including those that have been overlooked in the past.

It goes without saying that it is important to know the characteristics of a measure if it is used for measurement of an object of research. To make a comparison between various association measures including t-score and Log-Likelihood Ratio, in relation to the frequency of bigrams, it is necessary to work on the entirety of a naturally formed large-scale corpus of texts and not on a list of predetermined bigrams like adjective + noun.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 26370483. We thank Naohiro Takizawa, Ritsumeikan University (Japan), who helped us a lot to achieve this work.

References

BAKER, P., HARDIE, A., & MCENERY, T., 2006. *A Glossary of Corpus Linguistics*. Edinburgh University Press.

BARONI, M., 2009. Distributions in text, Lduelling, A. & Kitô, M. (eds.), *Corpus Linguistics, An International Handbook*, Mouton de Gruyter, Berlin, 803-822.

BYBEE, J., 2010. *Language, usage, and cognition*. Cambridge University Press.

CHURCH, K., & HANKS, P., 1990. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16(1), 22-29.

ELLIS, N.C., 2012. Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics*, 32,17-44.

EVERT, S., 2009. Corpora and collocations, Lduelling, A. & Kitô, M. (eds.), *Corpus Linguistics, An International Handbook*, Mouton de Gruyter, 1212-1248.

FRANÇOIS, J., & MANGUIN, J.-L., 2006. *Dispute théologique*, *discussion oiseuse* et *conversation téléphonique*: Les collocations adjectivo-nominales au cœur du débat, *Langue Française*, 150, 50-66.

GRIES, S. Th., 2012. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language*, 36(3), 477-510.

GRIES, S. Th., 2013. 50-something years of work on collocations What is or should be next..., *International Journal of Corpus Linguistics*, 18:1, 137-165.

HUNSTON, S., 2002. *Corpora in Applied Linguistics,* Cambridge University Press.

PECINA, P., 2010. Lexical association measures and collocation extraction, *Lang Resources & Evaluation*, 44, 137-158.

WRAY, A., 2012. What Do We (Think We) Know About Formulaic Language? An Evaluation of the Current State of Play, *Annual Review of Applied Linguistics*, 32, 231-254.

ZIPF, G. K., 1949. *Human Behavior and the Principle of Least Effort*, Addison-Wesley.

VARIATION DIATOPIQUE EN PHRASÉOLOGIE SPÉCIALISÉE DANS LE DOMAINE FINANCIER. ÉTUDE COMPARATIVE BASÉE SUR CORPUS

Daniel Gallego-Hernández
Université d'Alicante
daniel.gallego@ua.es

Résumé

Cet article présente une étude basée sur corpus portant sur la variation phraséologique spécialisée dans le domaine de l'économie. Tout d'abord, nous examinons brièvement le concept de phraséologie spécialisée. Ensuite, nous présentons notre étude de cas et analysons la variation de ce type d'unités à partir de deux corpus de comptes annuels : l'un en espagnol d'Espagne et l'autre en espagnol du Mexique. Nous analysons la variation tant au sein de chaque variété diatopique mais aussi entre les deux. Les résultats montrent l'existence de nombreuses variantes dans les deux variétés.

1. INTRODUCTION

Lors des dernières années de nombreux auteurs ont montré la complexité des unités phraséologiques de la langue générale (Corpas, 1996; Ruiz, 1997; Penadés, 2012, parmi d'autres). Nombre d'entre eux ont été particulièrement préoccupés par le figement et pensent que cette propriété ne peut pas toujours être attribuée à ces unités, car la phraséologie présente très souvent des variations sémantiques, morphologiques, morphosyntaxiques, graphiques, etc. (Mogorrón, 2010b).

Par conséquent, le figement peut être étroitement lié au phénomène de la variation. García (2008: 219) parle de *variantes phraséologiques* "quand les modulations des formes portant sur une expression figée sont codées ou institutionnalisées". Ces modulations peuvent être de nature linguistique différente : phonétique, graphique, morphologique, grammaticale, lexicale, etc., et normalement affectent un seul élément de l'expression. À cet égard, les constructions avec variantes sont composées de deux parties : une partie fixe ou invariable qui constitue la base et en facilite l'identification, et une partie variable qui détermine les variantes paradigmatiques (García, 2008: 219).

Parmi les opérations permettant d'introduire des variantes phraséologiques, nous pouvons trouver le remplacement de mots (le plus courant), mais aussi d'autres telles que l'ajout ou la suppression d'éléments ou le changement d'ordre syntaxique. Il est aussi possible de trouver des variations phraséologiques dans les différentes variétés diatopiques d'une même langue.

Mogorrón (2010a), lui, a analysé d'un point de vue contrastif les unités phraséologiques de l'espagnol d'Espagne et du Mexique et montre la complexité et la richesse de la langue espagnole dans son ensemble (Espagne et Amérique latine). Il plaide pour la création de répertoires lexicographiques ou des bases de données tenant compte de l'ensemble de variations diatopiques, ce qui rendra compte de ce qui est commun et idiosyncrasique de cette langue dans les pays où elle est parlée.

De la lecture de ces travaux et de bien d'autres encore sur la phraséologie portant sur les variations qui se produisent dans ces unités, il en résulte que la majorité sont particulièrement axés sur la langue générale. Or, qu'en est-il de la phraséologie spécialisée ? Est-elle également touchée par le phénomène de variation, même si les langues spécialisées sont *a priori* moins sujettes à la variation? Et les variations diatopiques ? Y a-t-il une telle variation entre les différentes variétés linguistiques d'une langue dans un domaine d'expertise particulier ? Y a-t-il des points communs permettant de standardiser la langue ?

Ainsi, dans cet article nous essayons précisément de contribuer à répondre à cette série de questions. Pour ce faire, dans un premier temps, nous nous interrogeons brièvement sur le concept d'*unité phraséologique spécialisée.* Deuxièmement, nous présentons le corpus avec lequel nous avons travaillé, ainsi que la méthodologie utilisée pour l'extraction des variantes phraséologiques spécialisées. Finalement, après avoir analysé les résultats, nous présentons quelques conclusions et ouvrons de nouvelles perspectives de recherche.

2. PHRASÉOLOGIE SPÉCIALISÉE ET VARIATION

Le concept de *phraséologie spécialisée,* comme beaucoup d'autres, n'est pas sans variation terminologique et crée des problèmes à la fois de dénomination et de conceptualisation (Aguado, 2007: 56-58). D'une part, plusieurs étiquettes portant sur ce phénomène ont été proposées : *multi-word terminological phrases, terminological phrasemes, specialized lexical combinations, collocations,* etc. D'autre part, la phraséologie spécialisée peut être comprise différemment: par exemple, des expressions stéréotypées très spécifiques d'une spécialité ne permettant pas d'ambiguïté, ou des combinaisons lexicales contenant une unité terminologique et un verbe ou une catégorie déverbale.

Malgré cette confusion terminologique apparente et mise à part la dénomination utilisée pour porter sur ce type d'unités, nous pouvons trouver pour notre étude quelques définitions qui, dans une certaine mesure, partagent des caractéristiques similaires.

Pour L'Homme (1997 : 16-16), les dénommées *combinaisons lexicales spécialisées* sont caractérisées comme suit :

> il s'agit des combinaisons de mots utilisées en langue spécialisée
> les unités de la combinaison sont liées grammaticalement
> leur emploi est dicté par les spécialistes
> elles comprennent un terme ainsi qu'une autre unité étant généralement polysémique
> le verbe peut être employé avec plusieurs termes

également le terme peut être employé avec plusieurs verbes
un verbe ayant un sens donné peut être utilisé avec toute une classe de termes

Lorente (2001) considère que les unités phraséologiques spécialisées sont des unités de connaissance spécialisée, qui correspondent à des structures syntagmatiques ou de phrases, qui ne sont pas lexicalisées, mais qui ont une certaine tendance à la stéréotypie ou un certain degré de figement. De même, elles contiennent au moins un terme.

De manière analogue, Aguado (2007) affirme que :

il s'agit de structures syntagmatiques comprenant un terme
ces structures comprennent aussi un verbe ou un élément déverbal
elles conservent un certain degré de figement, bien qu'elles admettent l'insertion de plusieurs éléments
elles admettent la commutation de leurs éléments
elles ont un sens spécifique dans le domaine de spécialité
elles sont utilisées fréquemment dans le domaine de spécialité

Selon certaines caractéristiques décrites jusqu'ici (certain degré de figement, le verbe pouvant être utilisé avec différents termes et vice versa, commutation des éléments), il s'ensuit que la variation des unités phraséologiques spécialisées existe et qu'elle peut avoir lieu à cause de plusieurs changements : changements lexicaux soit du verbe soit du terme, ajout d'éléments discursifs parmi les éléments de l'unité, etc.

3. ÉTUDE DE CAS

Le fait que nous pouvons trouver presque plus de variations que de formes figées dans le cas de la phraséologie et que la phraséologie spécialisée semble présenter aussi un certain degré de variation nous mène à étudier de plus près cette dernière. Étant donné que nous ne pouvons pas prétendre étudier toute la phraséologie spécialisée, nous nous penchons ici sur un domaine de spécialité concret.

Le but de cet article est donc, rappelons-le, d'étudier la variation de la phraséologie spécialisée, en particulier dans le domaine de la comptabilité, dans différentes variétés linguistiques ou diatopiques. Pour ce faire, nous avons compilé un corpus espagnol de comptes annuels d'Espagne et du Mexique. Le corpus comprend un total de 40 états financiers relatifs aux exercices 2012 et 2013. 20 comptes proviennent de sociétés cotées à l'Ibex 35 (Espagne) et les 20 autres proviennent des sociétés cotées à la bourse mexicaine BMV. Au total, le corpus contient à peu près 1370255 mots, distribués comme suit: 764486 (IBEX) et 605769 (BMV).

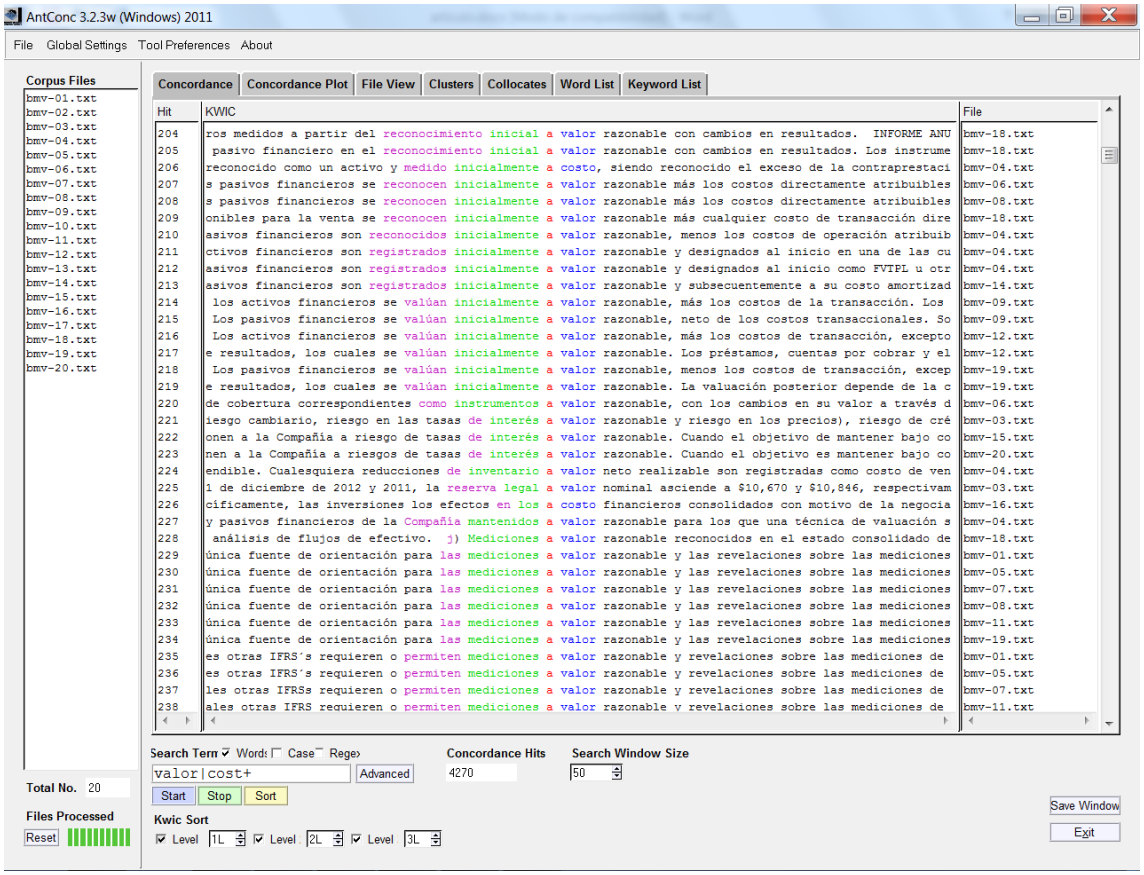
Une fois compilés, nous extrayons une série d'exemples d'unités phraséologiques ou combinaisons lexicales, nous analysons le phénomène de la variation non seulement au sein de chaque corpus mais aussi entre les deux variétés linguistiques :

3.1. Méthodologie d'extraction

Nous nous sommes basés sur deux cas spécifiques. Pour ce qui est du premier cas nous étudions les expressions utilisées par les comptables afin de déterminer ou évaluer le coût ou la valeur des actifs et des passifs des sociétés. Selon les mots-clés les plus fréquents sur les deux corpus, il semblerait qu'il s'agit d'une unité phraséologique assez fréquente dans le domaine de la comptabilité. Quant au deuxième cas, étant donné que la valeur ou le coût des biens peuvent être établis selon plusieurs méthodes, nous analysons les expressions qui introduisent une méthode particulière portant sur l'amortissement ou la dépréciation de la

valeur des biens : la méthode linéaire (*método lineal,* dans le cas de l'espagnol d'Espagne, ou *de línea recta,* dans le cas de l'espagnol du Mexique), celle-ci étant d'après les comptables la méthode la plus utilisée en comptabilité pour calculer la perte de valeur que subissent les actifs pendant les années.

Pour extraire les unités phraséologiques portant sur les notions expliquées ci-dessus, nous avons utilisé le logiciel AntConc (Anthony, 2007) et deux stratégies différentes. Dans le premier cas, nous avons extrait des concordances du corpus IBEX contenant les mots-clés *coste* (coût) ou *valor* (valeur) avec l'expression *coste|valor* afin de lire leurs contextes situés à gauche. Au total, nous avons récupéré 4646 concordances. En ce qui concerne le corpus BMV, étant donné la variation dénominative de type dialectal détectée précédemment dans le terme terme *costo* (au lieu de *coste*), nous avons interrogé le corpus avec l'expression *cost+|valor* et avons récupéré un total de 4270 concordances :



Dans le deuxième cas, nous avons extrait des concordances avec les expressions *lineal** (linéaire), dans le cas du corpus IBEX, et *línea* (ligne), dans le cas du corpus BMV, dans le but de détecter des contextes contentant des unités phraséologiques permettant de déterminer la valeur de l'amortissement linéaire des biens. Dans le corpus BMV, nous avons récupéré un total de 167 concordances, tandis que dans le corpus IBEX, nous avons récupéré 110 concordances :

3.2. Description de la variation

Décrivons maintenant le phénomène de la variation à partir des exemples tirés des corpus dans les deux variétés. Pour ce faire, nous présentons toutes les variantes identifiées pour chaque cas et nous les décrivons d'un point de vue formel. Nous ne tenons pas compte du sens qu'ont les combinaisons des termes *valor* et *coste* avec plusieurs adjectifs : *valor neto* (valeur nette), *valor razonable* (juste valeur), *coste histórico* (coût historique), *valor contable* (valeur comptable), etc.

3.2.1. Cas 1

Le tableau suivant présente les variantes détectées avec les stratégies présentées précédemment pour chaque corpus. La colonne de gauche présente, selon leur fréquence,[58] les unités identifiées dans les comptes annuels des sociétés cotées sur les sociétés de l'Ibex. La colonne de droite recueille celles identifiées dans le corpus BMV :

[58] Pour calculer la fréquence de chaque expression, nous avons interrogé Antconc avec des expressions régulières du type valor*@@a+@coste|valor*@@a+@valor, avec laquelle nous essayons de récupérer toutes les formes du verbe *valorar* (évaluer) accompagnée soit du terme *coste* soit du terme *valor,* ainsi que de la préposition qui se trouve au milieu de deux formes.

IBEX		BMV	
- valorar a (195) - valorar por (78) - registrar a (55) - reconocer por (46) - registrar por (38) - reconocer a (30) - contabilizar a (23) - referenciar a (15) - contabilizar por (11) - medir a (9) - realizar por (9) - determinar en base a (7) - determinar en base al cálculo de (5) - presentar por (5) - reflejar a (5) - registrar como (5)	- calcular a (4) - designar a (3) - presentar a (3) - realizar a (3) - establecer a (2) - reflejar por (2) - designar como a (1) - presentar valorado a (1) - reconocer atendiendo a (1)	- valuar a (158) - medir a (129) - reconocer a (120) - registrar a (91) - designar como a (26) - preparar sobre (la; una) base de (16) - considerar como (13) - designar a (13) - presentar a (13) - determinar con base en el (8) - reconocer por el (7) - evaluar sobre la base de (6) - designar como de (5) - evaluar a (5) - expresar a (5) - valorar a (5)	- calcular a (4) - realizar a (4) - referenciar a (4) - representar por el (4) - valorizar a (3) - calcular con base en el método de (2) - capitalizar a (2) - clasificar como a (2) - contabilizar a (2) - estimar con base en el (2) - valuar con base en el (2) - asignar como a (1) - calcular con base en el (1) - determinar sobre la base de (1) - reconocer bajo el método de (1) - valuar por el (1)

Nous constatons dans ces résultats qu'il existe effectivement différents types de variation intralinguistique et interlinguistique. Pour commencer, dans le cas du corpus IBEX, nous observons non seulement que le verbe qui accompagne *valor* ou *coste* peut varier (*valorar* [évaluer], *registrar* [enregistrer], *reconocer* [reconnaître], *contabilizar* [comptabiliser], etc.), mais aussi la préposition régie par ces verbes (*valorar a|por, registrar a|por, reconocer a|por*). En outre, les contextes de ces expressions nous montrent que ces unités permettent, d'une part, de comprendre différents types de déterminants accompagnant les termes *valor* ou *coste,* comme dans le cas de *contabilizar a valor, contabilizar a su valor, contabilizar al valor*, et, d'autre part, d'inclure entre les éléments de base de l'unité d'autres compléments, comme des adverbes ou des adjectifs, ainsi que d'autres unités terminologiques: *contabilizar al menor valor, se registran inicialmente por, valorar posteriormente a, valorar al cierre a, valorar también a, valorar íntegramente por,* etc. Dans le cas du corpus BMV, même si nous observons des variations lexicales portant sur le verbe accompagnant les termes *valor* ou *costo* (*valuar, medir, reconocer, registrar,* etc.), nous n'apprécions pas, cependant, autant de variation prépositionnelle comme dans le corpus IBEX (le verbe *reconocer* [reconnaître] est accompagné de la préposition *a* et dans une moindre mesure, de la préposition *por*). De même, dans ce corpus, les unités phraséologiques admettent l'inclusion de compléments entre leurs éléments de base: *se valúan al menor de su costo, se pueden medir inicialmente ya sea a valor razonable o al valor,* etc.

Du point de vue interlinguistique, les expressions les plus fréquemment utilisées dans chaque corpus ne coïncident pas. Pour commencer, tandis que dans le corpus IBEX la forme passive est à peine utilisée (*han sido valoradas a valor, son valoradas a coste*) et ce sont les formes impersonnelles qui sont les plus employées (*se valoran a coste*), dans le corpus BMV les formes passives, qui ont une fréquence beaucoup plus élevée dans les textes, se combinent avec les formes impersonnelles. D'autre part, pour ce qui est des différences lexicales, tandis que *valorar a valor* ou *coste* semble être l'unité la plus utilisée dans le corpus IBEX, ce n'est pas le cas dans la variété mexicaine, où cette même expression n'apparaît que cinq fois. Quant à l'expression la plus fréquente dans le corpus BMV, *valuar a valor,* elle n'existe pas dans le corpus IBEX, ce qui nous laisse penser qu'il s'agit effectivement d'une variante diatopique de l'espagnol d'Espagne. Dans le cas du corpus IBEX, nous trouvons ensuite l'expression *valorar por valor,* qui n'apparaît pas non plus dans le corpus BMV, et, dans le cas de ce dernier corpus, l'unité *medir a valor,* qui apparaît aussi dans le corpus IBEX, mais avec une fréquence inférieure. Peut-être que les expressions suivantes des deux corpus contenant les verbes *registrar* et *reconocer* pourraient nous faire croire qu'il s'agit

d'expressions standard car elles sont en effet employées dans le deux variétés linguistique et n'ont donc pas aucune trace de variation diatopique.

3.2.2. Cas 2

Le tableau suivant contient les variantes identifiées dans chaque corpus avec la deuxième stratégie. La colonne de gauche présente, selon leur fréquence,[59] les unités identifiées dans le corpus IBEX. La colonne de droite recueille celles identifiées dans le corpus BMV :

IBEX	BMV
- amortizar de forma lineal (9) - amortizar linealmente (9) - calcular amortización aplicando el método lineal (5) - reconocer de forma lineal (4) - cargar sobre una base lineal (3) - depreciar de forma lineal (3) - imputar de forma lineal (3) - calcular amortización por el método lineal (3) - amortizar distribuyendo linealmente (3) - cargar linealmente (3) - imputar linealmente (3) - realizar amortización linealmente (3) - realizar amortización siguiendo el método lineal (2) - amortizar siguiendo el método lineal (2) - abonar según el método lineal (2) - amortizar según el método lineal (2) - calcular amortización linealmente (2) - registran linealmente (2) - reconocer linealmente (2) - amortización se realiza de forma lineal (1) - hacer amortización de forma lineal (1) - registrar de forma lineal (1) - aplicar de forma lineal (1) - calcular amortización de forma lineal (1) - reconocer amortización de forma lineal (1) - calcular amortización según el método lineal (1) - amortizar por el método lineal (1) - realizar amortización utilizándose el método lineal (1) - calcular amortización usando el método lineal (1) - reconocer distribuyendo linealmente (1) - reconocer repartiendo linealmente (1)	- calcular depreciación conforme al método de línea recta (4) - amortizar en línea recta (3) - reconocer sobre una base de línea recta (3) - reconocer amortización con base en el método de línea recta (3) - calcular depreciación con base en el método de línea recta (3) - reconocer por el método de línea recta (3) - calcular depreciación usando el método de línea recta (3) - calcular depreciación en línea recta (2) - determinar demérito en línea recta (2) - ser registrado con base al método de línea recta (2) - amortizar por el método de línea recta (2) - ser amortizado con base en línea recta (1) - ser depreciado con base en línea recta (1) - depreciar en línea recta (1) - ser depreciado con base en línea recta (1) - reconocer amortización con base al método de línea recta (1) - reconocer amortización en base al método de línea recta (1) - amortizar bajo el método de línea recta (1) - amortizar de acuerdo con el método de línea recta (1) - empleando el método de línea recta (1) - amortiza con base en el método de línea recta (1) - reconocer depreciación con base en el método de línea recta (1) - amortizar mediante el método de línea recta (1) - calcular amortización por el método de línea recta (1) - calcular depreciación por el método de línea recta (1) - amortizar usando el método de línea recta (1) - reconocer depreciación usando el método de línea recta (1) - reconocer amortización usando el método de línea recta (1)

Contrairement au cas précédent, nous pouvons observer qu'il s'agit d'une unité phraséologique utilisée moins fréquemment dans les comptes annuels. Ceci est dû logiquement au fait qu'elle porte sur un mode particulier d'évaluation et non pas sur le fait d'évaluer en général, quelle que soit la méthode utilisée.

Tandis que dans le cas du corpus IBEX nous observons que les expressions *amortizar de forma lineal* et *amortizar linealmente* sont les plus communes, dans le cas du corpus BMV toutes les unités sont plus ou moins utilisées avec une fréquence similaire. Tel qu'il se passait dans le cas précédent, nous voyons que différents cas de variation existent dans les deux corpus. Par exemple, dans le corpus IBEX trois formes essentielles portant sur la

[59] Pour calculer la fréquence de chaque expression, nous avons interrogé Antconc avec tous les éléments formels de chaque unité, à l'exception de leurs verbes : *de forma lineal, con base en línea recta,* etc.

méthode spécifique coexistent (*forma lineal, método lineal* et *linealmente*) auxquelles il faut ajouter une autre forme moins fréquente : *base lineal*. Chacune fonctionne syntaxiquement d'une manière différente : tandis que *linealmente* a une fonction adverbiale et *forma lineal* est toujours précédée *de,* l'unité *método lineal* peut être introduite par différents éléments, tels que les prépositions *por* et *según,* ou les gérondifs *siguiendo, utilizando* o *usando*. En outre, de façon similaire à ce qui se passait avec *valor* et *coste,* ces unités peuvent être employées avec plusieurs verbes tels que, d'une part, *amortizar* et *depreciar,* qui aident à préciser la perte de valeur des actifs, et, d'autre part, *reconocer, registrar, cargar, calcular, imputar, realizar, abonar, aplicar* ou *hacer,* qui nécessitent un autre élément précisant le type d'annotation faite sur les livres comptables. Pour ce qui est du corpus BMV, il n'existe qu'une seule forme portant sur le méthode linéaire : *línea recta,* terme qui est cependant introduit par plusieurs éléments : la préposition *en,* des locutions prépositionnelles *sobre una base de* ou *con base en,* le terme *método de línea recta,* qui lui aussi est précédé des prépositions différentes (*por* et *mediante*), des locutions (*conforme a, con base en, con base a, en base a, bajo* et *de acuerdo con*) et des gérondifs (*usando* et *empleando*).

Si nous nous penchons maintenant sur les différences et les similitudes de chaque corpus, il semblerait qu'il existe plus de différences : le fait que chaque expression utilise une variante diatopique différente portant sur la perte de valeur des actifs (*linealmente* et *lineal,* d'une part, et *línea recta,* d'autre part) oblige à employer une syntaxe différente dans chaque corpus: *linealmente* fonctionne comme un adverbe et ne nécessite aucune préposition mais peut se combiner avec des gérondifs, l'adjectif *lineal* se combine avec l'unité *forme,* substantif qui à son tour est introduit par la préposition *de ;* par contre, le nom *línea* est introduit par la préposition *en* ou les locutions *sobre una base de* et *con base en.* Le seul point commun des deux corpus est, outre l'utilisation de verbes terminologiques *amortizar, calcular, reconocer,* etc., l'introduction du substantif *método,* qui porte explicitement sur le fait que la méthode linéaire (*lineal* ou *línea recta*) est utilisée. Or, dans ce cas, nous trouvons à nouveau plus de différences que de similitudes : tandis que dans le corpus IBEX le mot *método* est précédé des prépositions *según* et *por,* ainsi que des verbes *aplicando, siguiendo, usando* et *utilizando,* dans le corpus BMV il est introduit par des prépositions et des locutions comme *conforme a, con base en, por, con base a, en base a, bajo, de acuerdo con* et *mediante,* ainsi que par les gérondifs *empleando* et *usando.*

3.4. Discussion

Les résultats obtenus dans notre étude de cas nous fournissent certains éléments qui peuvent nous aider à répondre aux questions que nous nous sommes posées au début de ce travail.

A priori on pourrait penser que la variation phraséologique est un phénomène plus fréquent dans la langue générale que dans les langues spécialisées en particulier dans celles ayant un degré élevé de spécialisation et dont les textes sont écrits par des spécialistes et destinés à des spécialistes. Cependant, comme nous avons pu constater, la variation n'a pas seulement lieu au sein des unités phraséologiques spécialisées du fait que les unités terminologiques qu'elles comprennent varient du point de vue de leur dénomination mais aussi au sein des autres éléments de base composant l'unité, comme des prépositions, des locutions, des substantifs, etc.

La variation phraséologique n'existe pas seulement dans une même variété linguistique ou diatopique, mais aussi entre les différentes variétés diatopiques d'une langue. Dans le cadre des comptes annuels, un genre textuel visant à l'internationalité pourrait-on dire, les présidents des sociétés pourraient prétendre à standardiser le langage employé pour rédiger les états financiers de sorte que, par exemple, dans le contexte de la traduction en espagnol, les variations diatopiques soient neutralisées. Dans ce sens, dans certains cas il serait

possible de trouver des éléments phraséologiques en commun entre les variétés diatopiques, tel que nous avons vu avec les variantes phraséologiques *reconocer por el valor* et *registrar a valor* qui, même si elles ne sont pas les plus fréquentes dans les deux corpus, sont utilisées dans les deux variétés linguistiques. Cependant, comme nous l'avons vu dans le deuxième cas, il n'est pas toujours possible d'opter pour une solution intermédiaire qui, dans une certaine mesure, neutralise la variation diatopique des langues. Par conséquent, dans certains contextes, l'unité phraséologique utilisée appartient nécessairement à l'une ou l'autre variété linguistique.

4. EN GUISE DE CONCLUSION

Nous avons souhaité travailler avec la phraséologie spécialisée dans le domaine de la comptabilité afin de décrire les variations affectant certaines unités non seulement au sein d'une même variété linguistique, mais aussi entre différentes variétés diatopiques d'une même langue. En ce sens, nous avons compilé *ad hoc* un corpus spécialisé d'états financiers et nous avons analysé un certain nombre de cas spécifiques. Les résultats nous ont permis d'observer qu'il existe effectivement la variation phraséologique dans le langage spécialisé de la comptabilité et qu'il n'est pas toujours possible de trouver des unités ou des variantes communes aux différentes variétés diatopiques. Or ce travail n'est pas plus qu'une petite étude portant sur un domaine très spécifique, de sorte que les résultats ne sont pas du tout concluants. Ainsi, il conviendrait de continuer à décrire la variation affectant la phraséologie spécialisée en étudiant non seulement plus de cas, mais aussi d'autres variétés diatopiques de l'espagnol, ainsi que d'autres domaines d'expertise.

Bibliographie

AGUADO DE CEA, G., 2007. "La fraseología en las lenguas especializadas." In *Las lenguas profesionales y académicas,* ed. by Enrique Alcaraz Varó, José Mateo Martínez, and Francisco Yus Ramos, 53-65. Madrid: Ariel.

ANTHONY, L., 2007. "AntConc 3.2.1". Laurence Anthony's Homepage.

CORPAS PASTOR, G. 1996. *Manual de fraseología española.* Gredos: Madrid.

GARCÍA-PAGE SÁNCHEZ, M. 2008. *Introducción a la fraseología española. Estudio de las locuciones.* Barcelona: Anthropos.

GONZÁLEZ REY, M. 2012. "De la didáctica de la fraseología a la fraseodidáctica." *Paremia,* 21, 67-84.

L'HOMME, M. 1997. "Méthode d'accès informatisé aux combinaisons lexicales en langue technique." *Meta,* 42 (1), 15-23.

L'HOMME, M. 2000. "Understanding specialized lexical combinations." *Terminology,* 6 (1), 89-110.

LORENTE CASAFONT, M. 2001. "Terminología y fraseología especializada: del léxico a la sintaxis." In *Panorama actual de la terminología,* ed. by Manuel Pérez Lagos, and Gloria Guerrero Ramos, 159-180. Málaga: Comares.

MOGORRÓN HUERTA, P. 2010a. "Estudio contrastivo, lingüístico y semántico de las construcciones verbales fijas diatópicas mexicanas/españolas." *Quaderns de filología. Estudis lingüístics,* XV, 179-198.

MOGORRON HUERTA, P. 2010b. "Analyse du figement et des possibles variations dans les constructions verbales espagnoles." *Linguisticae Investigationes,* 33 (1), 86-151.

PENADÉS MARTÍNEZ, I. 2012. "La fraseología y su objeto de estudio." *Linred,* X, 1-17.

RUIZ GURILLO, L. 1997. *Aspectos de fraseología teórica española.* Valencia: Universitat de València.

DIME CÓMO HABLAS Y TE DIRÉ QUIÉN ERES: LAS UNIDADES DISCURSIVAS CON CARÁCTER FRASEOLÓGICO EN EL DISCURSO POLÍTICO

Henry Hernández Bayter
Université d'Artois
Laboratorio Textes et Cultures

Resumen

Nos proponemos analizar las unidades discursivas con carácter fraseológico movilizadas por el expresidente colombiano Á. Uribe Vélez. Nuestro objetivo es poner de manifiesto ciertas secuencias o fórmulas empleadas por el locutor político y su función en los discursos pronunciados entre 2002 y 2010 durante los Consejos Comunales de Gobierno.

Este estudio se centra sobre todo en ciertas secuencias estereotipadas que contienen aspectos de la religión y que son utilizadas para referirse al enemigo; además de centrarnos en ciertas comparaciones estereotipadas y citaciones que se fijan al interior del corpus de trabajo. Podemos constatar que estas unidades discursivas, en general, constituyen una herramienta primordial para los discursos del expresidente colombiano. Ellas facilitan y dan eficacia a las diferentes estrategias discursivas implementadas por el locutor político.

1. INTRODUCCIÓN

Los Consejos Comunales de Gobierno constituyeron un dispositivo de comunicación innovador en Colombia, durante los dos mandatos del expresidente Álvaro Uribe Vélez, entre 2002 y 2010. Este dispositivo permitió al locutor político desplegar un gran número de secuencias, eslóganes y formulas. Partimos del principio que el discurso político es propicio al empleo de fórmulas (noción empleada por A. Krieg-Planque, 2009) y los discursos del expresidente colombiano no son la excepción. Por el contrario, los discursos pronunciados durante los CCG se caracterizan por la presencia indispensable de fórmulas discursivas que permiten la articulación de la palabra del locutor político. Nuestro objetivo, por lo tanto, es poner de manifiesto el uso de estas fórmulas en los discursos pronunciados durante los CCG, con la ayuda del programa lexicométrico Lexico 3, y la descripción y análisis de dichas unidades desde un punto de vista discursivo. Hemos decidido dicho título para nuestra investigación ya que nos interesa destacar la manera cómo Álvaro Uribe Vélez habla durante sus discursos y qué lo caracteriza. Además, nos interesa analizar la función que dichas formulas obtienen particularmente en la comunicación política uribista.

Lo que nos interesa, en la presente investigación, es saber cuál es el papel de las unidades discursivas, bajo forma de fórmula o eslogan, utilizadas por el expresidente colombiano en los 277 discursos, que constituyen el corpus de estudio, pronunciados durante los CCG durante sus dos mandatos. Nuestro procedimiento de investigación se centrara, primero que todo, en la presentación del marco teórico y metodológico. Más adelante, nos focalizaremos en el análisis y descripción de la unidades discursivas que nos interesan y su función en los discurso del expresidente colombiano. Finalmente, presentaremos una síntesis del presente trabajo.

2. ENFOQUE TEÓRICO Y METODOLÓGICO

Presentamos en esta primera parte el enfoque teórico y metodológico que orienta el presente trabajo de investigación. Adoptamos así ciertos aspectos del análisis del discurso político apoyándonos en algunas nociones del campo de la fraseología y teniendo como metodología el uso del software de lexicometría, Lexico 3.

Para la definición del discurso político, tendremos cuenta de los aspectos analizados por E. Coseriu:

"La política puede entenderse como:

1. Terminología relativa a instituciones políticas.
2. Modo de emplearse los signos lingüísticos en la política y los significados que tienen según las distintas ideologías.
3. Conjunto de procedimientos propios de los políticos, es decir, el empleo del lenguaje en discursos o textos políticos, su finalidad y su función." (E. Coseriu, 1994:11-32)

Para nuestro trabajo, definimos el discurso político como toda utilización que de la lengua se hace en una situación de comunicación dentro del ámbito político, por locutores políticos. Dicha utilización puede ser escrita u oral utilizando un dispositivo de comunicación preciso, ya sea la televisión, la radio, la prensa, internet, una conferencia, meeting, entre otros. Por lo tanto, el discurso político corresponde a todo discurso pronunciado en una situación de política que pone en juego un mensaje político entre interlocutores políticos igualmente. Como lo precisa E. Coseriu, los elementos discursivos empleados en el discurso político corresponden a unidades discursivas que se adaptan al contexto político. Es decir, el discurso político no produce sus propias unidades, su propia terminología, por el contrario los locutores políticos extraen las unidades discursivas de otros dominios y las adaptan de acuerdo a su propósito, al mensaje y al destinatario.

En cuanto a la fraseología, se sabe de antemano que los locutores políticos recurren a unidades discursivas que pertenecen, en cierta medida, al dominio de la fraseología, por lo general para hacer referencia a la memoria colectiva de los interlocutores. Estas unidades poseen, por ende, aspectos tales como un cierto grado de fijación, una frecuencia de uso, una estructura polilexical y finalmente un alto contenido cultural. Como lo precisa D. Maingueneau (2015:79), las unidades discursivas que contienen características fraseológicas suponen, en el discurso político, todo un análisis y compresión de la memoria colectiva que debe ser decorticada para ser entendida, tanto por quien las utiliza como por quien las recibe.

La fraseología es entendida aquí, en su sentido más amplio, como el estudio lingüístico de un conjunto de unidades (secuencias, unidades fijas, colocaciones, formulas, eslóganes, expresiones idiomáticas, proverbios, entre otros) que poseen una frecuencia y un contexto de utilización, un grado de fijación, tanto a nivel de la estructura, como del significado y a nivel pragmático-discursivo. Dichas unidades hacen parte de un continuo fraseológico que

comprende el espectro entre las unidades libres de la lengua y aquellas que se encuentran completamente fijadas. Así retomamos la noción de continuo empleada por A. Rey, 1977, G. Gross, 1996, I. González Rey, 2002, y más recientemente por M. Pecman, 2004.

Para nuestro estudio utilizaremos igualmente la noción de fórmula que nos permitirá definir las unidades discursivas que nos interesan. Así, definiremos estas formas de expresión estereotipadas desde el punto de vista de P. Fiala y de A. Krieg-Planque. Por ende, una formula contiene un aspecto repetitivo y frecuente y un alto grado de circulación (P. Fiala, in P. Charaudeau et D. Maingueneau, 2002: 274). A. Krieg-Planque destaca el aspecto discursivo-pragmático, y por lo tanto la importancia del contexto y la adquisición de un significado especifico en dicho contexto de cada una de las formulas.

> "*[…] un usage particulier, ou une série d'usages particuliers, par lesquels la séquence prend un tournant, devient un enjeu, est repérée, commentée, cesse de fonctionner sur le mode «normal», des séquences qui nomment paisiblement et s'utilisent sans que l'on s'en rende compte.*" (A. Krieg-Planque, 2009: 85).

Estos aspectos teóricos nos permiten proponer la denominación de Unidades Discursivas con Carácter Fraseológico, UDCF, que pueden ser definidas, caracterizadas y analizadas a partir de ciertos métodos lexicométricos propuestos por Lexico 3 (A. Salem, 2003). Se trata, por consiguiente, de unidades discursivas ya que corresponden a un cierto contexto de utilización dentro del ámbito político y en un contexto frástico y temático particular dentro del corpus. Esta característica pragmático-discursiva en puesta de manifiesto gracias al estudio del inventario de los segmentos repetidos, de las concordancias y del análisis distribucional. En primer lugar, los segmentos repetidos nos dan cuenta de segmentos que tienden a aparecer siempre en bloque dentro del corpus. Así vemos que estos segmentos tienen una frecuencia de empleo. Luego, las concordancias y el análisis distribucional nos permiten estudiar el contexto de utilización de ciertos segmentos. En consecuencia, dichas unidades discursivas poseen ciertas características, más o menos marcadas, del dominio fraseológico, con carácter fraseológico: son más o menos fijadas, más o menos frecuentes, etc.

La lexicometría, como metodología, nos permite, como lo indicamos anteriormente, destacar, analizar y caracterizar las UDCF.

El software Lexico 3 propone un conjunto de métodos lexicométricos: en primer lugar, el inventario de segmentos repetidos permite dar cuenta de las formas simples que se atraen lexicalmente y aparecen frecuentemente juntas dentro del corpus; en lo que concierne a las concordancias, éstas nos permiten, a partir de una forma de base, identificar los elementos colocativos de dicha forma, pero también los elementos contextuales inmediatos, es decir el contexto frástico; finalmente, el inventario distribucional permite constatar los contextos más frecuentes de una forma y de su expansión colocativa hacia la derecha, poniendo de realce ciertas secuencias imprevistas igualmente.

3. ANÁLISIS Y DESCRIPCIÓN DE LAS UDCF

« La tradición social proporciona a la gente, en menor medida que antes, formas de expresión estereotipadas, […]. Las formulas y ritos convencionales de antes se siguen efectivamente utilizando [...]» (N. Elias, 1987:34).

Analizamos y describimos en esta parte de nuestra investigación lo que N. Elias llama formas de expresión estereotipadas o formulas y ritos convencionales que hacen parte de toda sociedad. Es bien sabido que el discurso político privilegia el uso de formas convencionales que facilitan la comunicación. Así pondremos a consideración ciertas secuencias estereotipadas fijas o que se fijan en el corpus, algunas comparaciones fijas y ciertas citaciones que se fijan igualmente en un contexto dado.

3.1. Secuencias fijadas que resaltan aspectos religiosos

En primer lugar, como lo explica H.M. Gauger, 1998 el aspecto religioso en las UDCF corresponde a un proceso de metaforización. En este caso, podemos afirmar que la religión está siempre presente en la estructura de las UDCF y que no hace referencia a lo religioso directamente sino a una característica idiomatizada que busca ilustrar algo más. Por lo tanto, podemos decir que la religión y las referencias religiosas en las UDCF representan un dispositivo idiomático muy interesante y de gran frecuencia.

Analizamos aquí algunas secuencias encontradas en el corpus y que contienen ciertos aspectos religiosos. Podemos afirmar que el discurso del expresidente colombiano recurre a secuencias estereotipadas que hacen referencia a las costumbre religiosas del país. Lo religioso concentra en sí mismo, por lo general, los valores sociales de los colombianos, como: la familia, la moral y la tradición, entre otros.

Para empezar, la UDCF *Semana Santa* se encuentra utilizada frecuentemente en los discursos del ex-presidente colombiano. La UDCF da la posibilidad al locutor político de acentuar ciertos aspectos de dicho evento religioso, de gran importancia para el mundo de habla hispana, en particular. La Semana Santa constituye un momento de reflexión y de unión alrededor de la religión. Este evento religioso permite así un momento de comunión y de identificación como miembros de un grupo o colectivo. La UDCF permite al locutor político poder dirigirse a un solo cuerpo unido por el evento religioso. Por otro lado, la Semana Santa representa un periodo de descanso y de vacaciones, para los menos crédulos. Podríamos decir que esta secuencia estereotipada permite al locutor político transmitir el siguiente mensaje: La Semana Santa es un momento de comunión y yo (presidente de la república) estoy en completa comunión con ustedes. A pesar de ser periodo de vacaciones, yo (Álvaro Uribe Vélez) continuó trabajando sin descansar por el bienestar del país. Si interrogamos las concordancias de la secuencia, obtenemos la siguiente figura que constituye un extracto de las concordancias completas de la UDCF Semana Santa:

Cauca) "los días sábado , en la víspera de la **Semana Santa** , hemos podido adelantar este *consejo Comunitario
que aquí la Semana Santa es muy bella , aquí la **Semana Santa** invita a concentrarse en Nuestro Señor y
hemos querido hacer en este sábado , víspera de la **Semana Santa** : para provocar otra reflexión en los *colombianos
aís empieza su recogimiento anual con motivo de la **Semana Santa** . un país creyente , con vocación , que respeta

La UDCF que verbaliza la referencia a la fiesta religiosa da la posibilidad de poner en marcha una estrategia discursiva concreta y particular. Para el expresidente colombiano acudir a la UDCF representa una forma de economía de la lengua, ya que la UDCF condensa en sí misma un significado discursivo muy importante. Por consiguiente, podemos afirmar que las UDCF poseen igualmente una característica de condensación significativa. A. Uribe Vélez trata de hacer pasar el siguiente mensaje: si la Semana Santa es una semana de reflexión y descanso, mientras que el presidente trabaja sin parar, la fiesta religiosa da la ocasión al locutor político de presentarse sí mismo como un hombre trabajador y siempre atento a las necesidades de los colombianos, sin importar la época de vacaciones, sino poniendo de manifiesto el hecho de reflexionar y comunicar las preocupaciones con el presidente. El locutor político se pone en analogía con Dios, ya que mientras durante la Semana Santa se reza a Dios y se le pide favores o se arrepiente de algo,

A. Uribe Vélez pide a los colombianos que le cuenten a él sus necesidades. Así asistimos a una comunión entre Dios y el locutor político que le legitima.

Invocar la dimensión religiosa de las UDCF permite al locutor político un número importante de aspectos: En primer lugar, le permite poner de manifiesto la tradición católica-cristiana que une a los colombianos. Lo que da una sensación de identidad nacional y de comunión por medio de lo religioso. Luego, utilizar esta UDCF da la posibilidad al locutor político de presentarse a sí mismo como creyente y practicante. Así el locutor político pone en relación sus acciones y decisiones con las del mundo divino. Es decir, crea un puente entre Dios mismo y los colombianos, él como persona, que ejecuta directamente los planes del todopoderoso. Además, el locutor político presenta una imagen de garantía y de moral, ya que la legitimidad le viene directamente de Dios. Finalmente, el aspecto religioso le da la ventaja de poner en marcha planes que no pueden ser cuestionados ya que son legitimados por medio de la religión.

3.2. Secuencias fijadas que hacen referencia al enemigo o adversario

Si las secuencias que hacen referencia a la religión tienen un lugar importante en los discursos del expresidente colombiano, también lo tienen otro tipo de secuencias: las que hacen referencia al enemigo. Comenzaremos aquí con las secuencias que remiten al enemigo y analizaremos, en particular, la secuencia estereotipada « *culebra de la violencia* ». Dicha UDCF remite al campo semántico de la violencia. Ella permite al locutor político designar a sus adversarios y de mostrar una imagen negativa de éstos a los ciudadanos. « *culebra de la violencia* » hace referencia a los grupos armados ilegales, en particular a las Farc. El uso del campo semántico del reino animal, en particular la serpiente o culebra, da una imagen negativa poniendo de manifiesto los aspectos negativos característicos de este animal. De hecho, el significado metafórico de la secuencia remite a la serpiente como animal oscuro y peligroso. Remite igualmente a la imagen bíblica de la serpiente que encontramos en El Génesis. El objetivo principal del locutor político es de llamar la atención de la ciudadanía haciendo un llamado a los saberes colectivos referentes a la Biblia y al mundo de la experiencia cotidiana de los colombianos. El uso de la UDCF tiene como objetivo permitir una recepción eficaz del mensaje. Así el locutor político busca crear un frente que encarne la lucha contra las Farc como enemigo, no solamente del expresidente A. Uribe Vélez, sino también de todos los colombianos.

Si analizamos las concordancias de esta UDCF:

on los espacios de paz que les ofrecieron los gobiernos para acrecentar su propósito criminal . la	***culebra de la violencia*** . por eso , compatriotas , mucho cuidado : si le aflojamos a esta
no repetición . por eso , no le podemos aflojar el garabato de madera firme que hemos puesto en la	***culebra de la violencia*** , sino que le tenemos que agregar otro en la cola , para acabar definitivamente
o lo dicen de frente , pero que soslayadamente lo dejan entrever . en lugar , le hemos puesto a la	culebra de ***la violencia*** una horqueta , un garabato de naranjo . no vamos a permitir , compatriotas

Se trata de utilizar la UDCF como un acto de habla que busca hacer reaccionar a los ciudadanos colombianos, que, hasta el gobierno de A. Uribe Vélez, sufrieron de la violencia de forma pasiva, sin actuar, y como una cierta forma de destino trágico. Se trata así de despertar a una ciudadanía pasiva que se había resignado a ser víctima de la situación del país. El locutor político incita a los colombianos a que hagan parte activa

en la guerra contra el terrorismo. La secuencia connotada y metafórica permite mostrar el lado negativo de las Farc, como lo indicamos anteriormente, y alertar a la población. Por lo tanto, da la posibilidad al locutor político de implementar una estrategia de persuasión: convencer a los colombianos de que son capaces de afrontar a los grupos terroristas. Por ende, el uso de la UDCF y de ciertos utensilios de la vida rural colombiana permite al expresidente vehicular un mensaje de manera directa, eficaz y de economizar su palabra, pero también hacer reaccionar a los colombianos rápidamente. Afirmamos aquí que se trata de una economía del lenguaje ya que el locutor político no tiene que crear nuevas formas o unidades discursivas, simplemente acude al imaginario de los colombianos y verbaliza ciertas imágenes que sirven a su propósito.

3.3. Comparaciones fijas

Analizaremos estructuras comparativas fijadas o que se fijan o estereotipadas que no tienen, en ningún caso, una simple función comparativa. Por el contrario, estas unidades son utilizadas con el objetivo principal de intensificar o ponderar, como lo precisa G. Corpas Pastor que las clasifica dentro de las locuciones adjetivales. (G. Corpas Pastor, 1996 :121).

Las comparaciones fijas pueden ser tratadas como UDCF gracias a su contenido, al contexto de utilización, a la frecuencia y al contenido histórico-cultural que vehiculan. De esta manera, funcionan como una estructura que facilita la recuperación de una información importante en el discurso, permiten intensificarla y/o acompañar un mensaje implícito y facilitar su transmisión. Además, estas UDCF corresponden a la función poética del lenguaje.

Para ilustrar, analizaremos dos UDCF comparativas fijas que presentan una misma estrategia discursiva: construir una imagen del adversario. Estas unidades contrastan un referente prototípico, un animal y un grupo opuesto a la política del expresidente. Las unidades son: *« como el pavo real »* y *« como el chavarrí »* (el *chavarrí* es un ave del norte de Colombia, a menudo utilizado para proteger a las otras aves en los criaderos gracias a su corpulencia y su carácter. Podemos constatar que el objetivo principal de estas comparaciones estereotipadas no es solamente comparar dos elementos, sino también intensificar el contenido del mensaje. En este caso, el objetivo es transmitir una imagen de agresividad y de cobardía de los adversarios al gobierno de A. Uribe Vélez.

¿Qué prefieren, mis compatriotas del Casanare? el Gobierno no se puede poner ***como el pavo real****, pero tampoco* ***como el chavarrí****. El Gobierno tiene que ser un guerrero de todos los días por los superiores intereses del país.* *CCG -197* *¿Qué por qué son tan bravos para calumniar y tan cobardes cuando se les enfrenta?* *Eso sí, para calumniar* ***son unos pavos reales: esponjados****. Y cuando se les enfrenta, se vuelven como un pájaro que hay en la Costa:* ***el chavarrí****. Le meten un grito y se echa.* *Entonces, a la hora de la calumnia* ***son un pavo real: esponjado, engreído****. Y a la hora que se les enfrenta, un* ***chavarrí acobardado****, buscando por ahí una raíz para esconderse echado.*
CCG – 197

Las características prototípicas utilizadas por el locutor político para su estrategia argumentativa son: las dos aves tienen un tamaño consecuente y tienen una apariencia imponente. El pavo real con su plumaje coloreado, esponjado y elegante; y el chavarrí por su tamaño es un ave que asusta a los predadores. Se trata de intensificar la imagen del gobierno del locutor político, por un lado: su gobierno no es un gobierno de apariencias,

que busca solamente mostrar una imagen imponente y pasajera; por el contrario, su gobierno representa un gobierno permanente que no cambia y que está siempre disponible para los colombianos.

Además, las comparaciones son utilizadas para describir al adversario. En un juego de espejos, el adversario como las dos aves son pretenciosos y cobardes. De hecho, se trata de una estrategia de desacreditación del enemigo por medio de una imagen negativa de un elemento de la naturaleza conocido por todos los colombianos. Ahora, hay que aclarar que las propiedades negativas no vienen de la memoria colectiva de los colombianos, pero de la interpretación del expresidente y de su estrategia de comunicación. Los adversarios se muestran bravos y pretenciosos al criticar al gobierno, pero cuando se trata de dar la cara y argumentar, desaparecen y se esconden. De esta manera, la oposición se ve restringida y casi borrada de la política colombiana. El objetivo del expresidente sería gobernar sin ninguna crítica y completamente solo.

3.4. Citaciones que se fijan

> « Se diferencian por […] tener un origen conocido. Se trata de enunciados extraídos de textos escritos o de fragmentos hablados puestos en boca de un personaje, real o ficticio[60]. » (G. Corpas Pastor, 1996: 143).

Como lo precisa G. Corpas Pastor, las citaciones son UDCF bajo forma de enunciados que retoman textos o fragmentos de textos de otra persona. Su característica fraseológica reside en el contenido cultural muy importante y depende del contexto de utilización. Por consiguiente, tienen como función establecer puntos comunes entre el discurso del locutor y la palabra de otro locutor, ausente, con el objetivo de reforzar sus ideas y decir que la información vehiculada es verdadera porque alguien más lo ha dicho anteriormente. Las fuentes principales de citaciones son los textos culturalmente marcados.

La citación no es solamente un medio estilístico, pero también una estrategia discursiva que permite al locutor mostrar que posee un cierto bagaje cultural y que él puede compartirlo con el auditorio. Por otro lado, si analizamos la estructura de las citaciones, éstas representan lo que D. Maingueneau llama una toma de distancia entre el locutor y lo que él dice, que se reconoce por el uso de comillas (D. Maingueneau, 1991: 134). Esta toma de distancia permite al locutor político legitimar su discurso dándole un grado de verosimilitud

Las citaciones que analizaremos aquí provienen del dominio del marxismo, del maoísmo y de la revolución china. Corresponden a citaciones de Mao Tse Tung o Mao Zedong tomadas de su libro « De la contradicción » 1937 et de Deng Xiaoping o Teng Hiao-Ping o Teng Hsiao-Ping: « *¡qué importa que el gato sea blanco o negro, lo importante es que cace ratones! »* Deng Xiao Ping pronunció esta frase con respecto a la condición de China, dirigida entonces por un régimen dictatorial. Hace referencia al cambio de régimen e indica que todo es relativo y que lo importante es la mejoría de la economía china.

Nosotros no podemos condenar a los pobres a que sigan siendo pobres, lo que hay que abrir en este país es el camino de que la gente se reivindique.
Ten-Siao-Pin, *el promotor de esa nueva gran revolución china, de él se destaca aquella frase que desató tantas consecuencias en el modelo de desarrollo chino:* ***no importa que el gato sea pardo o blanco, lo que importa es que cace ratones.***
CCG - 127

[60] Ils ou elles se différentient par… avoir des origines inconnues. Il s'agit d'énoncés extraits de textes écrits ou de fragments de textes oraux mis dans la bouche d'un personnage, réel ou fictif.

*Y en el debate él contesta con gran sentido común**: "no importa que el gato sea pardo o blanco, lo que importa es que cace ratones"**. Llevan casi 17, 20 años en China recibiendo flujos de inversión de 67 mil, 55 mil, 60 mil millones de dólares al año.*
CCG - 157

Podemos constatar que la UDCF es utilizada en el dominio de la economía, pero podemos indicar igualmente que el expresidente colombiano trata de construir una imagen de sí mismo de hombre "ilustrado", una imagen de "sabio", pero también una imagen de "profesor". Trata de hacer aparecer sus conocimientos en el dominio de la política, pero también busca subrayar que lo que pudo funcionar en China puede también servir para Colombia. Así, lo que se necesita es alertar a la población sobre la situación para poder luchar contra la pobreza y no marginar a las personas pobre por ser pobres. Por lo tanto, el locutor político se vuelve un guia moral para los colombianos, quien les enseña lo que deben y no deben hacer. Es decir, A. Uribe Vélez personifica una estrategia pedagógica por medio de sus discursos.

4. SÍNTESIS

En conclusión, podemos decir que el análisis de los discursos políticos presentado aquí, con la ayuda de un software lexicométrico, permite la exploración de ciertas unidades discursivas con carácter fraseológico utilizadas por A. Uribe Vélez. Hemos podido subrayar que el locutor político recurre a un número importante de secuencias estereotipadas, privilegiando el uso de aspectos religiosos que le permiten hacer un llamado a la memoria colectiva y a la tradición de los colombianos.

Además, hemos podido corroborar el empleo de ciertas comparaciones fijas que tienen como función intensificar el contenido del mensaje y resaltar ciertos aspectos que deben ser asimilados directamente por el público, gracias a un llamado a la experiencia de la vida cotidiana. Finalmente, el uso de ciertas citaciones adquiere un significado particular en este estudio, ya que permiten establecer un grado de legitimidad por medio de las palabras de un enunciador, ausente.

Bibliografía

CORPAS PASTOR, G., 1991. *Manual de fraseología española.* Madrid: Gredos Biblioteca Románica Hispánica.

COSERIU, E., 1994. Lenguaje y política. *Política, lengua y nación*. Madrid: Fundación Friedrich, Ebert, p. 11-32.

ELIAS, N., 1987. *La soledad de los moribundos*. México: Fondo de cultura económica, p, 34.

FIALA, P., 2002. Dictionnaire d'analyse du discours. Sous la direction de Patrick Charaudeau et Dominique Maingueneau. Paris: Éditions du Seuil, p. 274.

GAUGER, H., 1998. Elementos religiosos en las expresiones fijas del español. *Festschrift für Heinrich Bihler zu seinem 80. Geburtstag, hrgs. Von D. Brisenmeister und A. Schonberger.* Berlín: Domus Editoria Europea. [661], p. 893-903.

GONZÁLEZ-REY, I., 2002. *La phraséologie du français*, Linguistique et didactique. Toulouse : Presses universitaires du Mirail.

GROSS, G., 1996. *Les expressions figées en français. Noms composés et autres locutions.* Paris : Editions Ophrys.

KRIEG-PLANQUE, A., 2009. *La notion de « formule » en analyse du discours. Cadre théorique et méthodologique.* Besançon: Presse Universitaires de Franche-Comté.

MAINGUENEAU, D., 2015. Argumentation et scénographie. *Discours et effets de sens : argumenter, manipuler, traduire.* Arras: Artois Presses Université, p. 79.

MAINGUENEAU, D., 1991. *L'analyse du discours. Introduction aux lectures de l'archive.* Paris : Hachette.

PECMAN, M., 2004. *Phraséologie contrastive anglais-français : Analyse et traitement en vue de l'aide à la rédaction scientifique.* Th : Ling. : Nice, Université de Nice – Sophia Antipolis. Directeur de thèse: Henri ZINGLÉ.

REY, A., 1977. *Le lexique: images et modèles du dictionnaire à la lexicologie.* Paris: Libraire Armand Colin, p. 188-200.

SALEM, A. et al. 2003. *Lexico 3. Outils de statistique textuelle. Manuel d'utilisation,* SYLED – CLA2T, Université de la Sorbonne Nouvelle-Paris 3.

SALEM, A., 1988. Approches du temps lexical, statistique textuelle et séries chronologiques. *Revue Mots*, vol. 17, p. 105-143.

KORPUSBASIERTE INTRA- UND INTERLINGUALE KOLLOKATIONEN

Zita Hollós
Karoli Gaspar University of the Reformed Church in Hungary,
Faculty of Humanities
Institute of German and Netherlandic Studies, Department of
German Language and Literature
hollos.zita@kre.hu

Abstract

The following article describes different types of collocations and analyses them exemplarily in the corpus- and databased bilingual syntagmatic dictionary 'KOLLEX'.

Based on different aspects of the dictonary's concept in HOLLÓS (2004) the first and second sections shortly overview this german-hungarian learner's collocations dictionary. Then the next chapter presents the typology of collocations and shows its concrete realisation with a focus on the language learner. Though 'KOLLEX' is designed for active language production and therefore contains to 2262 polysemous words almost 50 000 collocations, due to its german register of more than 10.000 cotextpartners it can be used as an appropriate translation tool as well. With the new webbased userinterface and dynamic datastructure it can be extended, as a result mono- and bilingual collocations dictionary ressources can grow up on the database of KOLLEX.

1. ZIELSETZUNG

Folgender Beitrag hat die Zielsetzung den integrativen, in HOLLÓS (2004, 71f.) bereits in groben Zügen entworfenen Kollokationsbegriff weiter zu präzisieren, sowie mit Beispielen aus dem KOLLEX (ung. SZÓKAPTÁR) (HOLLÓS 2014c) zu untermauern. Dazu ist es unabdingbar, den integrativen, lernerlexikographisch orientierten Kollokationsbegriff und das korpus- und datenbankbasierte *Deutsch-ungarische KOLLokationsLEXikon* (HOLLÓS 2014b) kurz zu charakterisieren. Nach der Vorstellung dieses neuen Wörterbuchtyps werden die verzeichneten korpusbasierten Wortverbindungen und Kollokationen sowie die wichtigsten zahlenmäßigen Angaben über KOLLEX vorgestellt.[61]

Bei der Untersuchung der Kollokationen wird die Aufmerksamkeit einem stiefmütterlich behandeltem Kollokationstyp gewidmet, nämlich Kollokationen mit Adverbien. Sie entsprechen von den allgemein bekannten Strukturtypen den folgenden

[61] Frühere Publikationen wurden verschiedenen ausgewählten Aspekten der Wörterbuchkonzeption – wie der äußeren und inneren Selektion, dem Datenangebot – und damit auch der Korpusbasiertheit anhand von DTW und CCDB gewidmet, siehe dazu die Publikationsliste auf www.kollex.hu.

zwei: ADV+VERB oder ADV+ADJ. Während bei meinen letzten Untersuchungen der Kollokationen das Augenmerk zunächst dem intralingualen Aspekt, z.B. Syntax/Morphosyntax (HOLLÓS 2014a) oder dem interlingualen Aspekt wie der Interferenz (HOLLÓS 2013)[62] galt, richtet sich die Aufmerksamkeit diesmal mit Hilfe der korpusermittelten Daten von KOLLEX auf beide Aspekte. Es ist ein Versuch, in die Untersuchung eines komplexen phraseologischen Gegenstandes intra- und interlinguale Aspekte zu integrieren, um dadurch neue Erkenntnisse über dieses Phänomen, genauer über diese Kollokationstypen zu gewinnen. Außerdem sollte die integrative Methode für die Untersuchung der Kollokationen vertieft werden. Integrativ soll heißen, dass das Phänomen mit einer Methodenkombination aus der Kollokationsforschung (Strukturtypen), Korpuslinguistik (statistische Signifikanz), kontrastiver Linguistik (Interferenz) und Syntax (Valenz) näher eingegrenzt wird. Vorrangiges Ziel ist es, den intralingualen und den interlingualen Kollokationsbegriff in Anlehnung an HOLLÓS (2004) in Bezug auf einen Strukturtyp ADV+VERB weiter zu spezifizieren. Weiterhin werden intralinguale ADV+VERB-Kollokationen der Lemmastrecke G ausgewählt und auf ihren Status als intralinguale Kollokation geprüft. Danach wird eine mögliche, auf der Datenbank von KOLLEX basierende, automatisch erstellte lexikographische Präsentation vorgestellt. Mit dieser exemplarischen Bestandsaufnahme soll ein Schritt in Richtung einer phraseologisch orientierten Systematik im Bereich der Kollokationen getan werden. Als kleines Resümee werden die Tendenzen anhand der bisherigen Analysen zusammengefasst. Gleichzeitig wird für die Anwendung der erarbeiteten Kollokationstypologie für die zweisprachige Lernerlexikographie plädiert.

Zum Schluss werden Möglichkeiten der Onlinestellung, der Weiterentwicklung und der Erstellung weiterer (Teil-)Wörterbücher von KOLLEX skizziert.

1.2. Experiment 1[63]

Um die praktische Relevanz der Kollokationen für Fremdsprachenlerner bei der Sprachproduktion an eigener Haut zu spüren und die Notwendigkeit korpusbasierter (digitaler) Kollokationswörterbücher zu demonstrieren, ist ein kleines Experiment notwendig.

Vielen ist die folgende Situation wahrscheinlich bereits bekannt:

Das **neugeborene Baby** schreit viel, und man weiß häufig nicht, warum. Vielleicht will es **gestillt**, frisch **gewickelt** oder **hochgenommen** werden. Es muss regelmäßig gefüttert, gebadet und **trockengelegt** werden. Man freut sich, wenn es von Tag zu Tag **pummeliger** wird, oder anfängt zu **strampeln** und zu **lallen**. Meistens **nuckelt** es gern und wenn es beginnt zu **fremdeln** oder zu **zahnen**, weint es mehr. Wenn es endlich beginnt zu **krabbeln**, wissen die Eltern: bald ist nichts mehr in der Wohnung in Sicherheit. Es gibt nichts Schöneres auf der Erde als ein **kerngesundes Baby**!

Im oben aufgeführten Beispieltext liegt die Problematik nicht bei der Grammatik, sondern vor allem bei der Wortwahl, d.h. welche Adjektive und Verben der Deutschlerner

[62] In diesem Beitrag wurde gezeigt, wie Interferenz anhand von konkreten Beispielen, den sogenannten Interferenzkandidaten im KOLLEX berücksichtigt wurde. Dazu war die kurze Vorstellung der Interferenz auf mehreren sprachlichen Ebenen, vor allem auf der morphosyntaktischen, semantischen und stilistischen Ebene unumgänglich (vgl. u.a. FORGÁCS 2007, HEINE 2006, RÉDER 2006).

[63] Der leicht veränderte und ergänzte Text stammt aus KOLLEX (ung. SZÓKAPTÁR) (HOLLÓS 2014c, 954f.).

in Verbindung mit „Baby“ wählt. Die hervorgehobenen Wörter stehen nämlich in enger Verbindung zum Wort „Baby” und bilden zusammen typische Wortverbindungen/Kollokationen. Wenn es Schwierigkeiten gibt, dann ist oft die Rede von einer „Wortlücke” oder von einer „Wortunsicherheit”. Ob einem ein authentischer Text in einer Fremdsprache über dieses Thema ohne Nachschlagewerk, z.B. für das Deutsche KOLLEX oder FWD gelingt, kann leicht getestet werden.

Sie haben zwei Möglichkeiten:

1.Wenn Sie Deutsch als Fremdsprache sprechen, formulieren Sie selber schriftlich oder mündlich einen ähnlichen Text zum Thema *Baby*, ohne die obigen Vorlage!

2. Wenn Sie deutscher Muttersprachler sind führen Sie das selbe “Experiment” in Ihrer bestbeherrschten Fremdsprache aus!

1.2. Experiment 2 – mit Hilfe eines Nachschlagewerkes

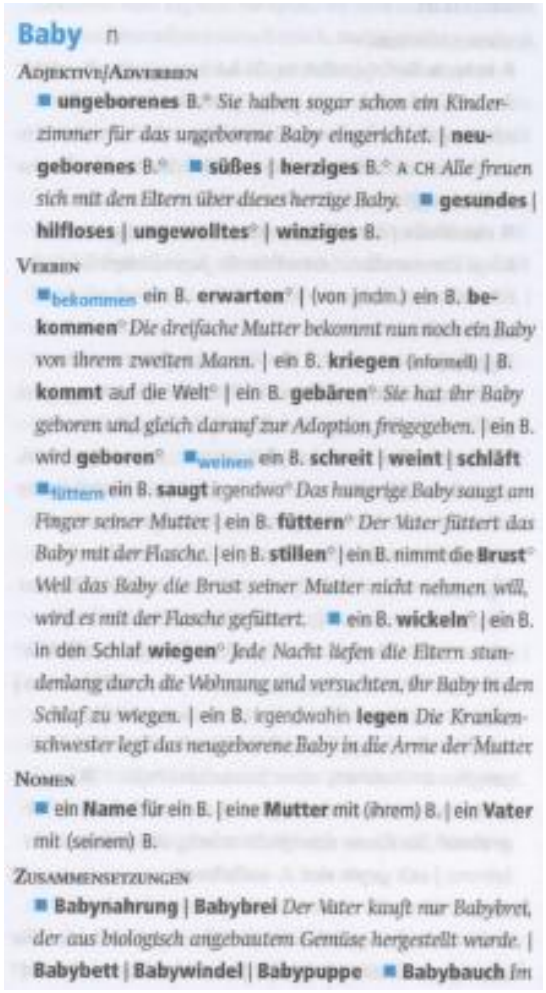
Jetzt greifen Sie zu einem der eben gewähnten Wörterbücher und versuchen Sie einen anspruchsvolleren Text zum Verhalten der Babys bzw. zum Umgang mit Babys zu gestalten. Hier gibt es auch zwei Wege, wenn Sie für das Deutsche als Fremdsprache ein Nachschlagewerk brauchen.

Erste Möglichkeit, mit Hilfe des Wörterbuchartikels zu **Baby** aus KOLLEX:

Baby ['be:bi] das, des Babys, die Babys <fn>
(kis)baba, csecsemő, bébi
ADJ ált jel **neugeboren** újszülött • csak jel **ungeboren** születendő, meg nem született • biz **kerngesund** makkegészséges • **munter** eleven, vidám • biz **pummelig** dundi, duci • **rosig** rózsás • csak jel **langersehnt** várva várt • **süß** édes, aranyos
VERB **gebären**[AKK] szül • **stillen**[AKK] (meg)szoptat • **adoptieren**[AKK] örökbe fogad • **wickeln**[AKK] (be)pelenkáz • **baden**[AKK] (meg)fürdet • **ein B. bekommen**[AKK] *állapotos* babája lesz, babát vár • **trockenlegen**[AKK] tisztába tesz • **anziehen**[AKK] felöltöztet • **ein B. erwarten**[AKK] *állapotos* babát vár • **pudern**[AKK] behintőporoz • **lallen**[NOM] gőgicsél, gügyög • **strampeln**[NOM] rugdalódzik • **fremdeln**[NOM] *idegenekkel szemben* félénk(en viselkedik) • **krabbeln**[NOM] *négykézláb* mászik • **nuckeln**[NOM] cumizik • **schreien**[NOM] bőg, ordít • **zahnen**[NOM] jön a foga, fogzik • **füttern**[AKK] megetet • **hochnehmen**[AKK] *karba* felvesz • **sich**[Dat] **ein B. wünschen**[AKK] babát szeretne

1. Abbildung: wa_1 aus KOLLEX (HOLLÓS 2014c, 56)

Zweite Möglichkeit mit Hilfe des einsprachigen Kollokationenwörterbuchs FWD:



Baby n

ADJECTIVE/ADVERBIEN

■ **ungeborenes** B.° *Sie haben sogar schon ein Kinderzimmer für das ungeborene Baby eingerichtet.* | **neugeborenes** B.° ■ **süßes** | **herziges** B.° A CH *Alle freuen sich mit den Eltern über dieses herzige Baby.* ■ **gesundes** | **hilfloses** | **ungewolltes**° | **winziges** B.

VERBEN

■bekommen ein B. **erwarten**° | (von jmdm.) ein B. **bekommen**° *Die dreifache Mutter bekommt nun noch ein Baby von ihrem zweiten Mann.* | ein B. **kriegen** (informell) | B. **kommt** auf die Welt° | ein B. **gebären**° *Sie hat ihr Baby geboren und gleich darauf zur Adoption freigegeben.* | ein B. wird **geboren**° ■weinen ein B. **schreit** | **weint** | **schläft** ■füttern ein B. **saugt** irgendwo° *Das hungrige Baby saugt am Finger seiner Mutter* | ein B. **füttern**° *Der Vater füttert das Baby mit der Flasche.* | ein B. **stillen**° | ein B. nimmt die **Brust**° *Weil das Baby die Brust seiner Mutter nicht nehmen will, wird es mit der Flasche gefüttert.* ■ ein B. **wickeln**° | ein B. in den Schlaf **wiegen**° *Jede Nacht liefen die Eltern stundenlang durch die Wohnung und versuchten, ihr Baby in den Schlaf zu wiegen.* | ein B. irgendwohin **legen** *Die Krankenschwester legt das neugeborene Baby in die Arme der Mutter*

NOMEN

■ ein **Name** für ein B. | eine **Mutter** mit (ihrem) B. | ein **Vater** mit (seinem) B.

ZUSAMMENSETZUNGEN

■ **Babynahrung** | **Babybrei** *Der Vater kauft nur Babybrei, der aus biologisch angebautem Gemüse hergestellt wurde.* | **Babybett** | **Babywindel** | **Babypuppe** ■ **Babybauch** *m*

2. Abbildung: wa_2 aus FWD (HÄCKI BUHOFER et al. 2014, 70)

Wenn jedoch diese bestbeherrschte Sprache das Englische ist, können Sie unter zwei Kollokationswörterbüchern (vom Oxford- oder Macmillan-Verlag) wählen. Ersteres, das OCDSE (CROWTHER et al. 2002) war das erste neukonzipierte, korpusbasierte Kollokationswörterbuch für das Englische.

Wenn diese bestbeherrschte Fremdsprache jedoch eine andere ist, stellt sich die Frage, ob es hierfür ein ein- oder zweisprachiges Print- oder digitales Wörterbuch für Kollokationen existiert, das Sie zur Rate ziehen könnten. Wenn nicht, kann der entsprechende Text über *Babys* mit Hilfe eines einsprachigen Lernerwörterbuches formuliert werden?

2. ALLGEMEINE VORSTELLUNG VON KOLLEX[64]

Wie eben bei den kleinen „Experimenten" hoffentlich ersichtlich wurde, reicht es nicht aus, Wörter zu lernen, man muss sie auch kombinieren können, d.h. die Wörter so verbinden, wie es für eine andere Sprache typisch ist, damit es „wirklich deutsch" klingt. Beim Verfassen eines Textes oder bei der Übersetzung in eine andere Sprache spielt nicht

[64] Die allgemeine Einführung in die Wörterbuchkonzeption, sowie die konkreten Zahlen übernehme ich leicht verändert aus KOLLEX (ung. SZÓKAPTÁR) (HOLLÓS 2014c, 954-956).

nur die Grammatik eine wichtige Rolle, sondern es kommt vor allem darauf an, ob wir natürlich und in einem angemessenen Stil formulieren, z.B. mit Hilfe typischer Wortverbindungen. Leider gibt es keine genauen Regeln dafür, wie, was oder warum etwas miteinander verbunden bzw. zusammen verwendet wird, uns bleibt lediglich die Aufzählung und Systematisierung der Wortverbindungen in einem speziellen Wörterbuch. Deshalb entstand das KOLLEX, ein korpusbasiertes Lernerwörterbuch der Kollokationen, in dem typische und häufige Wortverbindungen (sog. Kollokationen, z.B. **geschr** *ein frischvermähltes Paar* oder *ein vielversprechender Anfang,* **gespr** *die Schule schwänzen* oder *eine Frage aufwerfen, emsig arbeiten* oder **geschr** *effizient arbeiten* und *hundertprozentig sicher*) auf solche Weise zusammengestellt sind, dass den einzelnen Lemmata die Wörter anhand von Kollokationskategorien zugewiesen werden. Die 2262 Lemmata, meist mehrdeutige Wörter (Substantive, Verben, Adjektive und Adverbien – die sog. Basen), decken den Wortschatz von *Zertifikat Deutsch* ab. Zu diesen Lemmata werden die typischen, aus realen Texten stammenden Partner (sog. Kollokatoren) nach der statistischen Wahrscheinlichkeit ihres gemeinsamen Vorkommens aufgelistet. Darüber hinaus wurden sie mit weiteren, aus modernen ein- und zweisprachigen Wörterbüchern stammenden Partnern ergänzt. Auf diese Weise sind im KOLLEX 10 313 verschiedene, ebenfalls mehrdeutige Partner zu den einzelnen Lemmata aufgeführt.

2.1. Kollokationen und Wortverbindungen im KOLLEX

Das KOLLEX beinhaltet nahezu 50 000 typische Wortverbindungen und Kollokationen dieser Art, die nach ihren jeweiligen Kategorien unter den jeweiligen Wortbedeutungen verzeichnet sind. Bei den ***Substantiven*** sind in erster Linie Adjektive und Verben (z.B. *ein lesenswertes* ***Buch***, **gespr** *das* ***Buch*** *verschlingen*) zu finden, ***Verben*** und ***Adjektive*** stehen vor allem in Verbindung mit Adverbien (z.B. *behindertengerecht* ***bauen***, ÖKON *kostengünstig* ***bauen*** und *greifbar* ***nahe***).

Weiterhin zeichnet sich das Wörterbuch dadurch aus, dass zu den Wortverbindungen Valenzrealisierungen angegeben sind: etwa eine Reihe von Substantiven, von denen eins obligatorisch zusammen mit dem Verb erscheinen muss, z.B. treiben **Akk** <hat> *vmilyen tevékenységet* **űz, folytat**; **foglalkozik** vmivel Akk Sport Handwerk Viehzucht Ackerbau Spionage Unfug). Im Falle von Adjektivlemmata kommen auch noch weitere Wortverbindungskategorien, d.h. typische Substantive (sog. Basen) vor, z.B. breit +SUBS Fluss Schrank Bett Hüfte usw. Wenn sie adverbial gebraucht werden, dann stehen dort typische Verb- und Adjektivbasen (oder häufig attributiv gebrauchte Partizipien), z.B. breit +VERB grinsen lächeln lachen.

Unter den Wortverbindungen gibt es auch größere Wortkombinationen, die – auch wenn sie aus mehreren Wörtern bestehen – eine Einheit bilden und dennoch keine idiomatischen Wortverknüpfungen sind, z.B. *js* ***Hoffnung*** *bleibt unerfüllt*, *IRGENDWIE eingerichtetes* ***Haus***. Sie befinden sich in der letzten Kategorie, nach der Abkürzung KOMB. Ihre Zahl liegt über 2600. Die Gesamtzahl der Wortverbindungen im KOLLEX beträgt demnach etwa 62 000. Sowohl Anfänger als auch Fortgeschrittene finden in Produktionssituationen mit großer Wahrscheinlichkeit den fehlenden Partner zu einem konkreten Wort.

KOLLEX ist ein echtes Lernerwörterbuch, weil der Fokus auf der Sprachproduktion liegt. Außer den erwähnten Eigenschaften wird bei den typischen Wortverbindungen/Kollokationen auf grammatische, strukturelle, stilistische oder semantische o.ä. Abweichungen zwischen den beiden Sprachen (sog. Interferenzerscheinungen) geachtet. Auf solche wird mit einem Warnsymbol (☝) – mehr als 8000 Mal – hingewiesen. Diese Unterschiede bleiben nämlich während der Sprachrezeption

leicht unbemerkt, obwohl sie bei der Formulierung in der Fremdsprache eine äußerst wichtige Rolle spielen.

2.2. Wichtigste Zahlen über KOLLEX

KOLLEX und die dazugehörige webbasierte Datenbasis enthält **61 617 Wortverbindungen, Kollokationen und Kombinationen**.

Darunter sind:

- 48 757 Wortverbindungen/Kollokationen (primär aus dem Korpus gewonnene sememspezifische Kollokationen und freie Wortverbindungen, angeordnet nach der Signifikanz des Zusammenkommens im Korpus, Valenzangaben inklusive)
  - in den SUBS-, ADJ- und VERB-Zonen der Substantivartikel
  - in der ADV-Zone der Verb- und Adjektiv-/Adverbartikel

- 2661 (mehrgliedrige) Kombinationen in den KOMB-Zonen aller Wörterbuchartikel

- 8590 (ein- oder mehrgliedrige) Wortverbindungen/Kollokationen mit Partnern als Valenzrealisierungen, die hinter den grau hinterlegten Valenzangaben wie Akk aufgelistet sind

- 1609 (ein- oder mehrgliedrige) Wortverbindungen/Kollokationen mit (Basis)-Partnern in den Subzonen +SUBS, +VERB und/oder +ADJ, aber ausschließlich bei Adjektivartikeln

Es enthält weiterhin **10 313 deutsche Partner/Kollokatoren** im Nachspann, im *Register der deutschen Partner*, mit Angaben des sememspezifischen Vorkommens. Das Wörterbuch hat insgesamt **2262 Wörterbuchartikel:**

- Substantivartikel
- Verbartikel
- Adjektivartikel
- Adverbartikel

Um auch dem kontrastiven Aspekt Rechnung zu tragen gibt es im KOLLEX **8378 interlinguale Kollokationen,** die mit Warnsymbol „□“ markiert sind, also solche interferenzverdächtige Wortverbindungen, die bei der Sprachproduktion mit großer Wahrscheinlichkeit fehlerhaft produziert werden.

3. INTRA- UND INTERLINGUALE KOLLOKATIONEN

Die Bestimmung und Benennung der jeweiligen Kollokationskategorie resultiert auf der obersten Typologieebene aus dem Verhältnis der wortwörtlichen Übersetzung der Wortverbindung zur äquivalenten Wortverbindung im Ungarischen. Wenn sie miteinander in Großem und Ganzen übereinstimmen, also in Bezug auf die Struktur, Syntax und Semantik – mit Ausnahme der grundlegenden sprachtypologischen Unterschiede – annähernd gleich sind, dann können sie freie Wortverbindungen oder intralinguale Kollokationen sein. Was jeweils der Fall ist und wie intralinguale Kollokationen erfasst und für die lexikographische Praxis definiert werden können, wird weiter unten dargelegt und

nachgewiesen. Wenn die deutsche mit der ungarischen Wortverbindung nicht übereinstimmt, liegt eine interlinguale Kollokation vor. Sie lässt sich auch mit der intralingualen Kollokation kombinieren, wodurch ein Kollokations-Mischtyp *intra- und interlinguale Kollokation* als Resultat entsteht. Die anhand der statistischen Kookkurrenzanalyse mit Einbeziehung des kontrastiven Aspektes entstandenen vier Kategorien, darunter die drei Kollokationskategorien, sind dementsprechend die folgenden:

freie Wortkombination
Kollokation $_{\text{intra}}$
Kollokation $_{\text{inter}}$
Kollokation $_{\text{intra+inter}}$

In den nächsten Unterkapiteln gehe ich den beiden Haupttypen der Kollokationen, den intralingualen und den interlingualen Kollokationen nach.

3.1. Intralinguale Kollokationen

Im Folgenden werden zwei Hypothesen genannt, die für die Feststellung des intralingualen Kollokationsstatus aufgestellt wurden:

H1: Wenn sich ein Bestandteil einer sogenannten typischen, anhand einer Kookkurrenzanalyse signifikanten Wortverbindung, die einem der sechs definierten Strukturtypen entspricht, nämlich der affine (Kollokator-)Partner in einer bestimmten Bedeutung – in Bezug auf den semantisch autonomen (Basis-)Partner – durch seine sehr begrenzte Kombinierbarkeit auszeichnet, dann gilt die Wortverbindung **intralingual als kollokationsverdächtig**.

H2: Wenn das sehr begrenzte Kollokationspotenzial, das etwa 5 bis 10 (Basis-)Partner bedeutet, im IDS-Korpus oder CCDB empirisch, durch eine reziproke, vom (Kollokator-)Partner her durchgeführte Kookkurrenzanalyse nachweisbar ist, dann gilt die Wortverbindung von ihrem Status her als **intralinguale Kollokation**.
⇒ In diesem Fall kann das ganze Kollokationspotenzial des Kollokators dargestellt, also alle (Basis-)Partner angegeben werden.

Diese Hypothesen müssen an einer größeren Menge von konkreten Kollokationen aus KOLLEX und mit Hilfe der CCDB erprobt und verifiziert werden. Die gesammelten, potenziellen adverbialen Kollokatoren aus den Registerstrecken G und H im KOLLEX wurden überprüft und nur die beschränkt kombinierbaren als Kandidaten für intralinguale Kollokationen beibehalten. Als Ergebnis der Selektion und deren Kontrolle sind z.B. von den 595 adverbialen Kollokatoren mit G zuerst 79, dann 60 geblieben; gleichzeitig ist die Zahl der intralingualen Kollokationen von 139 auf 96 gesunken.

Zu diesen potenziellen Kollokatoren wurden gleichzeitig ihre aus der Kookkurrenzdatenbank (CCDB) gewonnenen weiteren signifikanten Basis-Partner tabellarisch gesammelt und jene markiert, die auch in der KOLLEX-Datenbasis vorhanden waren. Nicht nur die Basis, sondern auch das zugehörige Semem wurde dabei angegeben. Auf diese Art und Weise kann man Selektionsfehler später beheben, die durch die riesigen zu bewältigenden Datenmengen leicht entstehen konnten. Wenn der Kollokator im KOLLEX verzeichnet, aber in der CCDB nicht vorhanden war, wurde dies ebenfalls vermerkt. Diese Praxis ermöglicht relativ objektive Aussagen über die begrenzte Kombinierbarkeit der Kollokatoren für die weitere theoretische Arbeit.

Nach diesen Selektions- und Bearbeitungsvorgängen entstand eine komplexe Tabelle der Kollokatoren und eine der intralingualen Kollokationen mit ihren jeweiligen Übersetzungen.

Der folgende vereinfachte Ausschnitt aus der zweiten Tabelle, die der Kollokationen, zeigt nicht die ganze Komplexität der Analyseergebnisse, gibt aber einen guten Überblick über deren Endergebnis für die Kollokatoren von *geruhsam* bis *gewerblich*:

Deutsche intralinguale Kollokation (Kollokator **Basis**)	***inter***	***Ungarische Übersetzung der intralingualen Kollokation***
GERUHSAM **spazieren**		békésen **sétál(gat)**
GESCHÄFTSMÄSSIG **kühl**		hivatalosan **hűvös, kimért** *kissé elutasító*
GESCHLIFFEN **reden**		*formailag tökéletesen* kifinomultan / csiszolt stílusban **beszél**
GESCHRAUBT **reden**		erőltetetten / modorosan **beszél**
GESCHWOLLEN **reden**		fontoskodvan **beszél**
GESPENSTISCH **leer**		*félelmetesen* kísértetiesen **kihalt, néptelen**
gespenstisch **leise**		kísértetiesen **halk** *alig hallhat*
gespenstisch **still**		kísértetiesen **csendes** *zajtalan*
gespenstisch **still**		kísértetiesen **csendes** *békés*
GESTELZT **reden**		mesterkélten / természetellenesen **beszél**
GESTOCHEN **scharf**	x	minden részletében **éles** *teljesen kivehető kontúrokkal*
sich^{Akk} GESUNDHEITSBEWUSST **ernähren**	x	*egészségére figyelve* egészségesen / egészségtudatosan **táplálkozik**
gesundheitsbewusst **leben**		egészségtudatosan **él**
jn GEWALTFREI **erziehen**		erőszakmentesen / erőszak nélkül **(fel)nevel** vkit
sich^{Akk} GEWANDT **bewegen**		ügyesen **mozog**
in etw^{Akk}/auf etw^{Akk} gewandt **steigen**		ügyesen / fürgén **beszáll** vmibe, **felszáll** vmire
(jm/an etw^{Akk}) etw^{Akk} GEWERBLICH **vermieten**	x	(kis)üzemi / (kis)ipari célokra **kiad, bérbe ad, kölcsönöz** (vkinek) vmit

1. Tabelle: Auszug aus den Kollokationsanalysen der verbalen Kollokationen der Registerstrecke G, inter = interlinguale Kollokation

Dank den Erfahrungen des KOLLEX-Projekts und den zahlreichen konkreten Analysen anhand der intralingual kollokationsverdächtigen Wortverbindungen der G- und H-Registerstecken konnten die zwei Grundtypen der Kollokationen für die zweisprachige Lernerlexikographie wie folgt näher bestimmt werden (in Anlehnung an HOLLÓS 2004, 71f.):

> **Kollokation$_{intra}$**: Zu den intralingualen Kollokationen zählen solche typischen (i.S.v. statistisch signifikanten, also nachweisbaren) Wortverbindungen eines bestimmten Strukturtyps (von den sechs definierten Strukturtypen), die aus mindestens zwei lexikalischen Einheiten bestehen, und innerhalb derer einer der Kombinationspartner als semantisch autonome ***Basis*** gilt, die zur Semantisierung des von ihr abhängigen ***Kollokators*** beiträgt (in Anlehnung an HAUSMANN 1984, 1985), der mit Ergebnissen einer angemessenen reziproken statistischen Kookkurrenzanalyse (z.B. in den IDS-Korpora mit COSMAS oder CCDB) vom Kollokator her bestimmt werden kann und somit den intralingualen Kollokationsstatus der Wortkombination (d.h. die statistisch signifikante Häufigkeit des „Zusammen-Auftretens" der Kombinationspartner) festlegt. Die Verknüpfbarkeit der Basis mit dem Kollokator lässt

sich nicht (allein) mit semantischen Regeln begründen, sondern ihre eingeschränkte Substituierbarkeit ist (auch) auf die pragmatisch eingespielte Sprachnorm zurückzuführen.

Kollokation$_{inter}$: Zu den interlingualen Kollokationen zählen solche spezifischen Wortverbindungen eines bestimmten Strukturtyps (von den sechs definierten Strukturtypen), die nur mit Hilfe einer kontrastiven, in diesem Fall einer deutsch-ungarischen Analyse zu gewinnen sind und bei der Sprachproduktion die häufigsten lexikalischen und (morpho)syntaktischen Interferenzquellen darstellen, weil sie aus der Muttersprache nicht problemlos und fehlerfrei ins Deutsche übersetzt werden können und deshalb meistens nicht erratbar sind. Ihre Klassifizierung geschieht in Anlehnung an FORGÁCS (2007), nach der den Kollokationen angepassten Äquivalenztypologie für Phraseme.

3.2. Probeartikel zu den Wortverbindungen/Kollokationen im sekundären Wörterverzeichnis der „Kollokatoren“

Die folgenden Testartikel wurden aus der Datenbasis anhand der Basis-Artikel automatisch generiert und stellen die Rohversion für die weitere Redaktion der „Kollokator“-Artikel dar:

galant <mn>
ADV + Verb
kissé régi, galant auffordern↑ --- gálánsan felkér

ganztags <mn>
ADV + Verb
ganztags unterrichten↑ --- egész nap tanít oktat
ganztags unterrichten↑ --- egész nap tanít
ganztags arbeiten↑ --- teljes munkaidőben/nyolc órában dolgozik
ADV + Adj
ganztags berufstätig↑ --- teljes munkaidőben/nyolc órában dolgozó/hivatását gyakorló/munkaviszonyban álló
ganztags beschäftigt↑ --- teljes munkaidőben/nyolc órában foglalkoztatott/alkalmazott

gastfreundlich <mn>
ADJ + Subs
csak jel, gastfreundlich Haus↑ --- vendégszerető család/ház
gastfreundlich Ort↑ --- vendégszerető város/falu (lakossága)
ADV + Verb
gastfreundlich aufnehmen↑ --- (nagy) vendégszeretettel/vendégszeretőn befogad/felvesz

gebetsmühlenhaft <mn>
ADV + Verb
gebetsmühlenhaft wiederholen↑ --- robot módjára (meg)ismétel

gebietsweise <mn>
ADV + Verb
gebietsweise regnen↑ --- egyes helyeken/elszórtan esik (az eső)
gebietsweise schneien↑ --- helyenként esik a hó/havazik

gebührenfrei <mn>
ADJ + Subs
gebührenfrei Anruf↑ --- díjmentes (telefon)hívás
gebührenfrei Nummer↑ --- díjmentesen hívható (telefon)szám
gebührenfrei Leistung↑ --- díjmentes/ingyenes juttatás(ok)/szolgáltatás(ok)
gebührenfrei Studium↑ --- díjmentes/államilag finanszírozott tanulmányok/képzés
ADV + Verb
gebührenfrei umtauschen↑ --- díjmentesen (át)vált
gebührenfrei umtauschen↑ --- díjmentesen/illetékmentesen kicserél/becserél

gebührenfrei einzahlen↑ --- díjmentesen/illetékmentesen befizet
gebührenfrei parken↑ --- díjmentesen parkol

3. Abbildung: automatisch generierte Wörterbuchartikel, $wa_{3\text{-}9}$ aus der Datenbasis von KOLLEX

Natürlich enthalten nicht alle nur intralinguale Kollokationen, sondern auch freie Wortverbindungen sowie interlinguale Kollokationen. Letztere können manuell, anhand der Basis-Artikel markiert werden. Das Programm verbindet automatisch den „Kollokator" im neuen Wörterbuchartikel mit dem im Basis-Artikel. Gleichzeitig werden sie mit Verweiskennzeichnungen in Form von ↑-Symbolen versehen. Damit entsteht ein automatisch generiertes, bidirektionales Verweisungssystem.

Die obigen Wörterbuchartikel müssen für das zweite Kollokator-Wörterverzeichnis noch manuell ergänzt, verfeinert und evtl. mit den Verben (oder Adjektiven, Substantiven) aus der CCDB ergänzt werden, falls sie für die jeweilige Benutzergruppe eine relevante Wortschatzgröße darstellen. Die Verweise werden in beiden Anwendungen immer automatisch generiert. Dadurch werden die Kollokationen des Basis- und Kollokatorverzeichnisses miteinander verknüpft und stehen somit für die elektronische Version von KOLLEX auch als Hyperlinks zur Verfügung.

Es ist noch zu überlegen, welche sekundären Wörterverzeichnisse von den „Kollokatoren" her für verschiedene Benutzergruppen sinnvoll sind. Das kann jedoch m.E. ohne große Mengen von automatisch generierten Wörterbuchartikeln nicht entschieden werden, deshalb bedarf es weiterer theoretischer und praktischer Untersuchungen.

4. KURZES RESÜMEE

Zusammenfassend lassen sich aufgrund der intra- und interlingualen Kollokationen zur Registerstrecke G aus KOLLEX folgende Analyseergebnisse formulieren:

1. Von den potenziellen adverbialen Kollokatoren (595) ist nur etwa ein Zehntel ein echter Kollokator, also ein Kollokator mit beschränktem Kollokationspotenzial und damit ein Bestanteil einer intralingualen Kollokation.

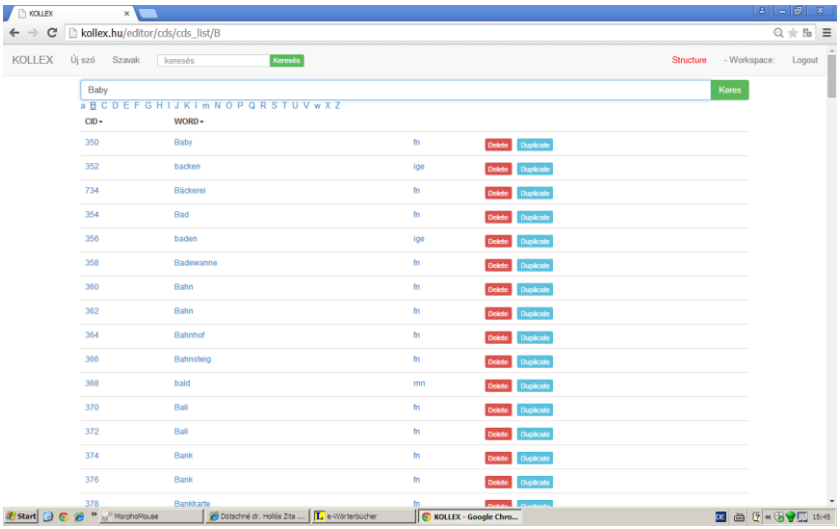
2. Unter den intralingualen Kollokationen (96) ist nur etwa ein Viertel (27) auch interlingual markiert, gehört also zu den intra- und interlingualen Kollokationen.

3. Bei den untersuchten echten Kollokatoren gibt es nur ganz wenige, deren Übersetzung auch im Ungarischen ein echter Kollokator ist und wenn ja, bildet er mit anderen Verben eine intralinguale Kollokation. Deshalb scheinen in zwei Sprachen äquivalente intralinguale Kollokationspaare sehr selten zu sein, was – in aller Deutlichkeit – für ihre gesonderte Kodifizierung in lexikographischen Nachschlagewerken spricht.

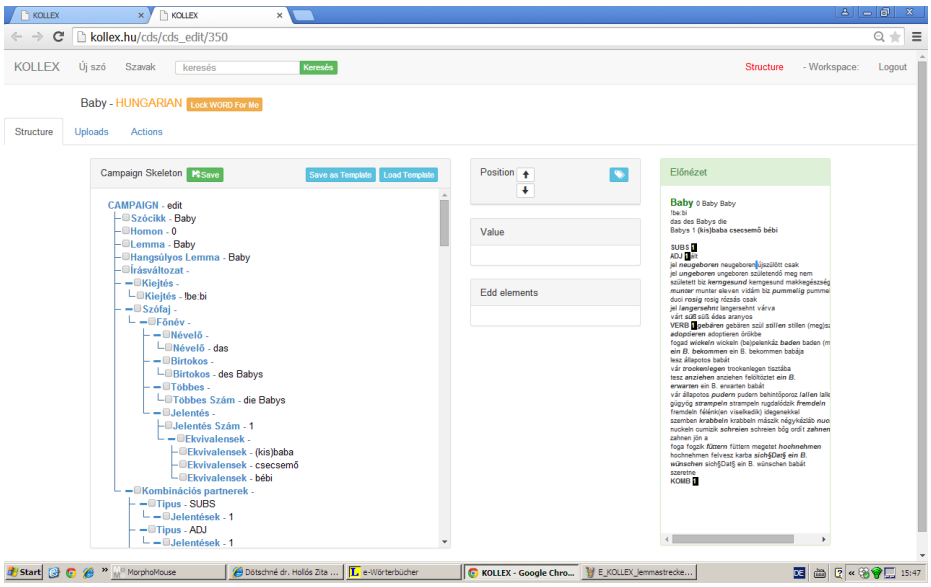
5. AUSBLICK: KOLLEX & CO.

Abschließend wird die Onlinestellung der Datenbasis[65] anhand von zwei Screenshots und die Erstellung weiterer (Teil-)Wörterbücher von KOLLEX kurz diskutiert.

Folgende Screenshots zeigen den Anfang der Lemmastrecke B, dessen erstes Lemma **Baby** ist und den dazugehörigen Wörterbuchartikel, der am Anfang des Beitrags behandelt wurde, aus der webbasierten Datenbasis von KOLLEX:



4. Abbildung: Auszug aus dem webbasierten User-Interface der KOLLEX-Datenbasis zur Lemmastrecke B



5. Abbildung: Auszug aus dem webbasierten User-Interface der KOLLEX-Datenbasis zum Lemma Baby

[65] Die ganze Datenbasis von KOLLEX bereits mit dynamischer Datenstruktur ist passwortgeschützt auf der Webseite von KOLLEX für die Projektmitglieder seit Frühling 2015 zugänglich.

Die Layout-Formate als „Node Types“ für die Benutzerschnittstelle im Web können in der webbasierten Anwendung beliebig gewählt werden. Sie werden automatisch auch für den Druck oder für die Onlinepublikation für alle Daten gleichen Typs übernommen.

Zuletzt möchte ich kurz die möglichen, in der Datenbasis bereits angelegten weiterführenden Erweiterung- und Entwicklungsmöglichkeiten ansprechen. Auf der Materialgrundlage von KOLLEX ist die sukzessive Erstellung weiterer Online-Wörterbuchresourcen möglich, sowie gedruckt oder online ein- und zweisprachige Lernerwörterbücher und/oder (Teil-)Wörterbücher von SZÓKAPTÁR/KOLLEX.

In deutsch-ungarischer Relation sind folgende Print- und Online-„Wörterbücher“ vorstellbar:

1. Digitales zweisprachiges Lernerwörterbuch: **WEB-SZÓKAPTÁR**
2. Multimediales zweisprachiges Lernwörterbuch: **KIDS-SZÓKAPTÁR**
3. Digitale (Teil-)Wörterbücher von SZÓKAPTÁR/KOLLEX:
   3.1. Wörterbuch der intralingualen (deutschen) Kollokationen
   3.2. Wörterbuch der interlingualen Kollokationen
   3.3. Wörterbuch der intra- und interlingualen Kollokationen

Im zweisprachigen Bereich sind die Möglichkeiten nahezu „unbegrenzt“:

- Deutsch-**spanisch**es KOLLokationsLEXikon
- Deutsch-**portugiesisch**es KOLLokationsLEXikon
- Deutsch-**englisch**es KOLLokationsLEXikon
- Deutsch-**französisch**es KOLLokationsLEXikon
- ...

Ob KOLLEX & CO. Zukunftsmusik bleibt, hängt nicht zuletzt von der Wissenschafts-Community ab.

Literatur

(Korpora)

CCDB = Belica, Cyril: Kookkurrenzdatenbank CCDB**.** Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. 2001 ff., Institut für Deutsche Sprache, Mannheim. [online] Internet-Adresse: http://corpora.ids-mannheim.de/ccdb [am 1. Juni2015].

DTW = Korpus „Deutscher Wortschatz“. [online] Internet-Adresse: http://wortschatz.informatik.uni-leipzig.de/ [am 1. Juni2015].

IDS-KORPUSZ = Korpora des Instituts für Deutsche Sprache, Mannheim. [online] Internet-Adresse: http://www.ids-mannheim.de/cosmas2/win-app/install/ [am 1. Juni2015].

(Bücher)

FORGÁCS, E. 2007: *Kontrastive Sprachbetrachtung*. Szeged: 2007.

HOLLÓS, Z. 2004: *Lernerlexikographie: syntagmatisch. Konzeption für ein deutsch-ungarisches Lernerwörterbuch*. Tübingen: Niemeyer. (Lexicographica. Series Maior 116).

HOLLÓS, Z. 2014c: *SZÓKAPTÁR: Német–magyar SZÓkapcsolatTÁR. Korpuszalapú kollokációs tanulószótár. KOLLEX: deutsch-ungarisches KOLLokationsLEXikon. Korpusbasiertes Wörterbuch der Kollokationen. Deutsch als Fremdsprache*. Szeged: Grimm Kiadó.

RÉDER , A. 2006: *Kollokationen in der Wortschatzarbeit*. Wien: Präsens 2006.

(Sammelbände)

CROWTHER, J., DIGNEN, S., LEA, D. eds. 2002. *Oxford Collocations Dictionary for students of English*. Oxford: University Press.

HÄCKI BUHOFER, A. DRÄGER, M., MEIER, S., ROTH, T. (Hrsg.) 2014: *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen: Francke.

(Buchkapitel)

HAUSMANN, F. J. 1985: Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: H. Bergenholtz/J. Mugdan, (Hrsg.) *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 22.-30.6.1984*. Tübingen: Niemeyer. 118-129. (Lexicographica. Series Maior 3)

HOLLÓS, Z. 2014a: Syntagmatik im KOLLEX: Die lexikographische Darstellung grammatischer Syntagmatik in einem zweisprachigen Kollokationslexikon für Deutschlerner. In: M. J. Domínguez Vázquez, F. Mollica, M. Nied Curcio, (Hrsg.) 2014. *Zweisprachige Lexikographie zwischen Translation und Didaktik*. Berlin/Boston: de Gruyter. 113-129. (Lexikographica. Series Maior 147)

HOLLÓS, Z. 2014b: Ein Stiefkind der Kollokationsforschung. Kollokationen mit Adverbien. In: I. Dringó-Horváth, J. Fülöp, Z. Hollós, P. Szatmári, A. Szentpétery-Czeglédy, E. Zakariás (Hrsg.) 2014. *Das Wort – ein weites Feld. Festschrift für Regina Hessky*. Budapest: L´Harmattan. 25-39.

(Zeitungsartikel)

HAUSMANN, F. J. 1984: Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts 31*, 395-406.

HOLLÓS, Z. 2013: Interferenzkandidaten in zweisprachigen Lernerwörterbüchern, insbesondere im deutsch-ungarischen Kollokationslexikon KOLLEX. *Lexicographica 29*, 92-116.

(Internetquellen)

HEINE, A. 2006: *Ansätze zur Darstellung nicht- und schwach idiomatischer verbonominaler Wortverbindungen in der zweisprachigen (Lerner)-Lexikografie Deutsch-Finnisch (Beschreibung eines Forschungsvorhabens).* [online] Internet-Adresse: <http://www.linguistik-online.de/27_06/heine_a.html> [am 1. Juni 2015].

KOLLEX, 2014: *SZÓKAPTÁR/KOLLEX*. [online] Internet-Adresse: <www.kollex.hu> [am 1. Juni 2015].

MIT BEDACHT: KORPUSLINGUISTISCHE UNTERSUCHUNGEN ZU STRUKTUREN *[PRÄPOSITION + SUBSTANTIV]* MIT ADVERBIALER FUNKTION

Herbert J. Holzinger
Instituto Interuniversitario de Lenguas Modernas Aplicadas (IULMA)
Universitat de València
herbert.holzinger@uv.es

Abstract

In der vorliegenden Arbeit wird das Präpositionalgefüge *mit Bedacht* im Deutschen Referenzkorpus einer Analyse unterzogen. Dabei stellt sich heraus, dass *Bedacht* ein phraseologisch gebundenes Wort (unikales Element) ist, das fast ausschließlich in bestimmten Phrasemen auftritt, hauptsächlich in *mit Bedacht*. Obwohl es als phraseologische Einheit im Prinzip stabil sein müsste, lassen sich zwischen Präposition und Substantiv lexikalische Elemente zur Intensivierung einschieben (*mit viel/großem/besonderem Bedacht*), welche die interne Festigkeit des Phrasems durchbrechen. Untersucht wird auch die externe Festigkeit, die darin besteht, dass das adverbiale Phrasem vornehmlich mit bestimmten Verben (*wählen auswählen, aussuchen*) und Verbgruppen (Handlungsverben und Kommunikationsverben) kookkurriert.

Die aus der Korpusanalyse abgeleiteten Ergebnisse sind sowohl für die einsprachige als auch für die mehrsprachige Lexikografie von Bedeutung.

1. EINLEITUNG

Im Rahmen der Forschungsgruppe FRASESPAL werden unter Projekt Nr. FFI2013-45769-P Strukturen des Typs *[Präposition + Substantiv]* des heutigen Deutsch analysiert und

mit dem Spanischen kontrastiert. Der vorliegende Beitrag gliedert sich in dieses Projekt[66] ein, indem eine bestimmte dieser Strukturen detailliert untersucht wird: *mit Bedacht*.

Es scheint sich zunächst um ein ganz reguläres Präpositionalgefüge zu handeln, in dem *mit* in der in Grammatiken und sonstigen Nachschlagewerken dokumentierten modalen Funktion auftritt und aufgrund der Unzählbarkeit von *Bedacht* kein Artikel steht. Allerdings stellt sich nun die Frage nach der Bedeutung des Substantivs *Bedacht*. Ein Blick in einschlägige Wörterbücher soll darüber Auskunft geben.

Wahrig (2011, CD-Rom) verwendet als Bedeutungsparaphrase „Überlegung, Ruhe, Besonnenheit, Umsicht" und führt folgende Beispiele an: *etwas mit ~ tun; mit (gutem) ~ vorgehen; ohne ~*.

Das DWDS bietet keine Definition des Substantivs, sondern zitiert *mit, voll, ohne Bedacht* und definiert „mit, voll, ohne Überlegung, Umsicht, Sorgfalt". Dazu kommt die als „gespreizt" markierte Wendung *auf etw. Bedacht nehmen* mit der Bedeutungsumschreibung „auf etw. bedacht sein".

Im Duden online findet man folgende Darstellung:

- **ohne Bedacht** (ohne jede Überlegung, unbesonnen, überstürzt: sie reagierte ohne Bedacht)
- **mit Bedacht** (mit einer bestimmten Überlegung; überlegt, besonnen: mit Bedacht auswählen)
- **auf etwas [keinen] Bedacht nehmen** (auf etwas [nicht] bedacht sein, [nicht] achten: darauf müssen wir besonderen Bedacht nehmen)

Während laut Wahrig *Bedacht* eine spezifische, beschreibbare Bedeutung hat, wird sowohl im DWDS als auch im Duden dem Wort keine eigene, unabhängige Bedeutung zugeordnet. Es werden nur Verbindungen angegeben, in denen es vorkommt. Daraus muss der Wörterbuchbenutzer schließen, dass das Wort nur in den angegebenen festen Wendungen, d.h. also Phrasemen, auftritt. Was in diesen Wörterbüchern nur implizit angedeutet ist, bringt die Printversion *Duden Universalwörterbuch* explizit zum Ausdruck durch die Markierung „nur in den gehobenen Fügungen u. Wendungen".[67]

Laut dieser Markierung handelt es sich also um ein so genanntes, aus der Forschungstradition bekanntes **unikales Element** (cf. Holzinger 2013, 2012), ein Wort, das nicht frei in semantisch und syntaktisch regulären Verbindungen der Sprache vorkommt, sondern nur in bestimmten festen Verbindungen, also Phrasemen.

[66] Das genannte Forschungsprojekt fügt sich als „Wortverbindungsmuster kontrastiv" ein in das von Kathrin Steyer am Institut für Deutsche Sprache entwickelte Gesamtkonzept der usuellen Wortverbindungen.

[67] Der Sonderstatus derartiger Wörter wird in den Wörterbüchern ebenso unterschiedlich wie unzuverlässig behandelt (vgl. dazu Holzinger 2013: 55).

Alternative Bezeichnungen für *unikales Element* sind u.a. **phraseologisch gebundenes Formativ** (PGF) und **unikales Wort** (Sailer/Trawiński 2006). Dobrovol'skij/Piirainen (1994) stellen eine Liste mit 188 Wörtern zusammen, die das „Zentrum" der PGF bilden, da sie „dem „durchschnittlichen" Muttersprachler bekannt sind" (S. 457). In der *Sammlung der unikalen Wörter des Deutschen* (Sailer/Trawiński 2006) werden 445 Kandidaten für unikale Wörter aufgelistet.[68] *Bedacht* ist in keiner der beiden Listen enthalten, auch nicht in der umfangreichen Auflistung von Fleischer (1997: 37ff).

Ich verwende die Benennung **phraseologisch gebundenes Wort** (PGW) und die damit zusammenhängenden Termini *phraseologisch gebundener Gebrauch* und *ungebundener* (=„freier") *Gebrauch*.

Um diese vermutete „Unikalität", also das ausschließliche Auftreten des fraglichen Wortes in einem Phrasem überprüfen zu können, war man lange Zeit hauptsächlich auf Introspektion, auf sprachliche Intuition angewiesen, sowohl auf die eigene als auch die von eventuell befragten Informanten. Diese bekanntlich nicht immer zuverlässige Methode kann heute durch Korpusanalyse ersetzt werden, die zu empirisch fundierten Ergebnissen führt.

2. KORPUSANALYSE

Als Untersuchungsbasis dienen die Vorkommen im Deutschen Referenzkorpus (DeReKo) des Instituts für Deutsche Sprache, in dem über das Such- und Analysesystem COSMAS II auf Texte mit insgesamt über 25 Milliarden Wörtern (Stand 15.09.2014) zugegriffen werden kann.

2.1. Gebundener und ungebundener Gebrauch

Verschiedene gezielte Suchanfragen[69] an die öffentlichen Korpora brachten am 11.6.2015 als erste Annäherungswerte für das Substantiv ***Bedacht*** insgesamt 9352 Treffer, die in Tabelle 1 nach gebundenem und ungebundenem Gebrauch aufgeschlüsselt sind.

	Treffer	Prozent
Bedacht	9352	100
mit Bedacht	7335	78
ohne Bedacht	100	1
Bedacht nehmen	1050	11
gebundener Gebrauch insgesamt	8485	91
ungebundener Gebrauch	867	9

Tabelle 1: Vorläufige Verteilung des Substantivs *Bedacht*.

Diese ersten Ergebnisse bestätigen zunächst einmal die in den Wörterbüchern angegebenen Phraseme, es verbleibt aber auch, zumindest vorläufig, ein nicht

[68] Online zugänglich unter http://english-linguistics.de/codii/codiibw/de/list-complete.xhtml.

[69] Alle in dieser Studie verwendeten Suchanfragen wurden zwischen 30. April und 22. Juli 2015 durchgeführt.

unbedeutender Rest von 9% des Vorkommens in ungebundenem, also „freiem" Gebrauch. *Bedacht* ist nicht absolut fest an die drei genannten Phraseme gebunden und demnach nicht absolut „unikal" in dem Sinn, dass es nur in einem bzw. mehreren (einander bedeutungsähnlichen) Phrasemen vorkommt, sondern auch ungebunden.

Neuere Untersuchungen zur Festigkeit[70] haben gezeigt, dass die phraseologische Gebundenheit der traditionell als unikal eingestuften Wörter nicht immer absolut fest ist. Der Stabilitätsgrad dieser phraseologischen Einheiten variiert stark und erreicht nur in bestimmten Fällen 100%, wie etwa bei *in Windeseile*, laut einer korpusbasierten Analyse von Stumpf (2014) im DeReKo.[71]

Innerhalb des gebundenen Gebrauchs ist unter den Treffern von *mit Bedacht* die als Präposition fungierende Fügung *mit Bedacht auf* (ca. 100 Vorkommen) enthalten, die aber, ebenso wenig wie das Funktionsverbgefüge *Bedacht nehmen* Gegenstand dieser Untersuchung ist.

Auffällig ist, dass *ohne Bedacht* bei Weitem seltener auftritt als *mit Bedacht*. Dazu kommt der Umstand, dass es in den meisten Fällen (74 von 100) in der verneinten Form *nicht ohne Bedacht* gebraucht wird und somit semantisch als Litotes seinem Pendant *mit Bedacht* entspricht, bzw. es sogar noch verstärkt.

Wie beschrieben, ist *Bedacht* in mehrere Phraseme integriert, tritt aber auch, bedeutend weniger häufig, in unabhängiger Verwendung auf. Folgende Korpusbelege für ungebundenen Gebrauch zeigen, dass dieser unauffällig weil semantisch gleich wie im Präpositionalgefüge ist.

PNW Lebendspenden erfordern besonderen **Bedacht** und eine besondere Umsicht, vor allem
E96 Regierenden erwarten dürfen: Umsicht, **Bedacht,** aber auch Weitblick und - falls nötig
BRZ06 Und langsam, wie in sorgsam abwägendem **Bedacht,** schüttelte er den Kopf. Der kleine

2.2. Interne Festigkeit von *mit Gebrauch*

Obwohl es sich um ein Phrasem handelt, also eine im Prinzip invariable Einheit mit interner Festigkeit (Fixiertheit), können zwischen Präposition und Substantiv so genannte lexikalische Füller eingeschoben werden, die durch entsprechende Suchanfragen an das Korpus ermittelt werden können.

Eine Suchanfrage an alle öffentlichen Korpora des Archivs W (mit Neuakquisitionen) ergab, dass innerhalb einer Spanne von fünf Wörtern zwischen *mit* und *Bedacht* in fast 700 Fällen die Festigkeit durch den Einschub von Adjektiven durchbrochen ist. Das Phrasem

[70] Vgl. dazu etwa Stumpf (2014); Holzinger (2013, 2012); Söhn (2003).

[71] Meinen Untersuchungen nach ist der Stabilitätsgrad von *in Windeseile* zwar sehr hoch, erreicht aber nicht ganz 100%. Eine Suche im DeReKo am 11.6.2015 ergab ca. 98% *in Windeseile*; ca. 1% *mit Windeseile*; ca. 1% ungebundene Verwendung. Ein Blick in die Korpora des DWDS zeigt, dass im Kernkorpus, das Texte aus dem gesamten 20. Jahrhundert enthält, *mit Windeseile* fast die Hälfte der Vorkommen ausmacht. Im Zeit-Korpus (Texte dieser Zeitung ab 1946) schrumpft dieser Prozentsatz auf 4% (Holzinger 2013: 57). Es könnte sich um eine Entwicklungstendenz handeln, in der *in Windeseile* den Gebrauch von *mit Windeseile* immer mehr zurückdrängt.

ist also modifiziert, es liegt aber dennoch ein gebundener Gebrauch vor. Deshalb muss die obige Tabelle 1 revidiert werden. Der gebundene Gebrauch erreicht nun 98%.

	Treffer	Prozent
Bedacht	9352	100
mit Bedacht	7335	78
mit X Bedacht	695	7
ohne Bedacht	100	1
Bedacht nehmen	1050	11
gebundener Gebrauch insgesamt	9180	98
ungebundener Gebrauch	172	2

Tabelle 2: Revidierte Verteilung des Substantivs *Bedacht*.

In der Folge soll die Struktur *mit X Bedacht* näher analysiert werden. Mithilfe einer Kookkurrenzanalyse (vgl. dazu ausführlich Steyer 2004) kann man feststellen, welche Lexikoneinheiten mit welcher Frequenz zwischen Präposition und Substantiv eingeschoben werden und wie stark diese Bindung ist. Die Kookkurrenzanalyse ist ein mathematisch-statistisches Verfahren, das signifikante Regelmäßigkeiten in Wortkombinationen aufdeckt und als Maß für deren Affinität oder Kohäsion angesehen werden kann.[72] Als Indikator der Affinität dient der so genannte LLR-Wert (LLR steht für *log likelyhood ratio*). Die wichtigsten lexikalischen Füller für *mit X Bedacht* sind in Tabelle 3 festgehalten.

	LLR
viel	1185
sehr viel	43
mehr	180
großem	642
grossem	48
größtem	88
besonderem	315
äußerstem	149
vollem	47
klugem	25

Tabelle 3: Lexikalische Füller.

Diese Adjektive mit der stärksten Affinität zum Phrasem dienen primär der Intensivierung und können als „usuelle Wortverbindungen" (Steyer 2013) angesehen werden. Bedeutungsspezifische Adjektive sind belegt, aber äußerst selten und als eher auffälliger Sprachgebrauch einzustufen, wie die folgenden Korpusbelege zeigen:

R97 Ausgereiftes war auf dem Treffen zu hören. Reinhard Gagel hatte **mit** provokativem **Bedacht** die Wortschöpfung Improvisiakum parallel zu

[72] Vgl. dazu http://www1.ids-mannheim.de/kl/projekte/methoden/ka.html.

T99 Leere Bierflaschen zum Beispiel. Die stellt die Gefährtin **mit** nonchalantem **Bedacht** stets auf dem Küchenfußboden vor der Balkontür
U94 **'Mit** bösartigem **Bedacht** hat der Republikaner-Führer Schönhuber den Juden

Als lexikalische Füller können nicht nur Einzelwörter auftreten, sondern auch syntaktische Muster (Steyer 2013) des Typs *mit X und Bedacht*. Dabei sind die wichtigsten Vertreter *mit Ruhe und Bedacht, mit Sorgfalt und Bedacht, mit Vorsicht und Bedacht, mit Umsicht und Bedacht*. In diesen Musterbildungen wird jeweils ein semantisches Merkmal, das in *Bedacht* vorhanden ist, besonders hervorgehoben. Dabei fällt auf, dass sowohl *Ruhe*, als auch *Sorgfalt* und *Umsicht* in den eingangs erwähnten Wörterbuchdefinitionen als Quasi-Synonyme zur Bedeutungsumschreibung verwendet werden. *Vorsicht* konnte in zahlreichen Beispielen aus dem Korpus als Bedeutungselement isoliert werden. Im Folgenden werden repräsentative Korpusbelege angeführt:

WDD11 Aber um Editwars zuvorzukommen, schlage ich vor das ganze **mit** Ruhe und **Bedacht** anzugehen - also erstmal ein bisschen warten,
PBY Die Schritte für die Zukunft können und müssen wir **mit** Sorgfalt und **Bedacht** planen, Schritt für Schritt die Weichen
DPA13 bedrohten Länder, die USA, Japan und Südkorea, haben bislang **mit** Vorsicht und **Bedacht** auf die andauernden Provokationen reagiert,
RHP13 den Ball an die Hand. Schiedsrichter Fabian Vollmar, die Partie **mit** Umsicht und **Bedacht** leitend, wies auf den Punkt.

Diese Musterbildungen sind relativ variabel: *Bedacht* kann auch die erste Position einnehmen. Hierbei handelt es sich dann allerdings nicht mehr um eine interne Komponente im engeren Sinn, sondern um eine quasi interne Konstituente, die durch *und* mit *Bedacht* koordiniert wird.

NUN97 Vatikan nähert sich dem heißen Thema **mit Bedacht und** Vorsicht. Eher unpolitisch

Auch Dreierstrukturen sind durchaus geläufig:

RHZ11 Es sei **mit** Ruhe, **Bedacht** und Weitsicht vorzugehen.

Beide Elemente können durch Adjektive modifiziert werden:

SOZ07 **Mit** viel Ruhe und gebührendem **Bedacht**

2.3. Externe Festigkeit von *mit Bedacht*

Unter der externen Festigkeit soll die bevorzugte Kombinatorik verstanden werden, also die Tatsache, dass bestimmte Phraseme besonders häufige Kookkurrenzpartner haben. Eine Kookkurrenzanalyse zeigt eine auffällig starke Bindung des Verbs *wählen* an das adverbiale Gefüge *mit Bedacht*, gefolgt von den zwei bedeutungsähnlichen Verben *auswählen* und *aussuchen*.

	LLR	Treffer
wählen	16.432	2543
auswählen	2.921	470
aussuchen	570	154
formulieren	321	71
vorgehen	250	118
angehen	178	68
handeln	120	77
sprechen	65	92
sagen	60	87
tun	55	76
herangehen	35	10

Tabelle 4: Kookkurrierende Verben von *mit Bedacht*.[73]

Man kann einen Schritt weiter gehen und die Frage stellen: was wählen wir mit Bedacht? Auch hier gibt es klare Präferenzen. Die stärkste Bindung besteht zu *Worte* (LLR 3218), dann folgen *Ort* (LLR 659), *Zeitpunkt* (LLR 434) und *Termin* (LLR 428). Dazu kommen spezifischere Bezeichnungen für *Ort* wie *Standort, Tagungsort, Veranstaltungsort*; statt *Worte* findet man auch *jedes Wort* oder den allgemeineren Ausdruck *Formulierung*.

Hinzuweisen ist auch auf typischen Gebrauch bestimmter Verbformen, wie etwa des Partizips II von *wählen*, das die stärkste Affinität aufweist, in Beispielen wie:

Der Zeitpunkt ist/war mit Bedacht gewählt.

Aus semantischer Perspektive lassen sich die kookkurrierenden Verben in eine Gruppe 'überlegtes menschliches Handeln' eingliedern, die man wie folgt unterteilen kann:

1) *wählen, auswählen aussuchen*;

2) Verben der Kommunikation: *sprechen, sagen, schreiben, formulieren*;

3) Planung oder Beginn einer Handlung: *angehen, herangehen*;

4) Ablauf einer Handlung: *vorgehen, tun, handeln*.

Um zu überprüfen, ob die starke Bindung von *mit Bedacht* an die Verben *wählen/auswählen/aussuchen* auch bei im Prinzip synonymen Adverbialverbindungen besteht, wurde eine Kontrastanalyse (Steyer 2013: 139) durchgeführt, bei der *Bedacht* durch die in den Bedeutungsparaphrasen der Wörterbücher angegebenen Quasi-Synonyme (*Überlegung, Ruhe, Besonnenheit, Umsicht*) ersetzt wurde, um die Kookkurrenzverhalten auf Gemeinsamkeiten und Unterschiede untersuchen zu können.

[73] Treffer nach Suchanfrage an alle Korpora (8786 Treffer), 30. April 2015;

LLR nach Suchanfrage an öffentliche Korpora (7357 Treffer), lemmatisiert.

	Kookkurrenzen mit LLR-Wert					
mit Bedacht	wählen 16432	auswählen 2921	aussuchen 570	formulieren 321	vorgehen 250	angehen 178
mit Überlegung	ausführen 58	töten 45	angehen 34	geschehen 32	leben 26	tun 19
mit Ruhe	entgegensehen /blicken, .. 266	angehen 114	begegnen 110	herangehen 85	tun 75	meistern 73
mit Besonnenheit	reagieren 155	begegnen 57	handeln 53	führen 40	agieren 34	bewältigen 34
mit Umsicht	leiten 379	führen 252	walten 83	angehen 72	lenken 72	agieren 60
mit Sorgfalt	auswählen 345	behandeln 208	pflegen 178	prüfen 148	zubereiten 143	vorgehen 122

Tabelle 5: Kontrastanalyse.

Ein relativ ähnliches Kookkurrenzverhalten wie *mit Bedacht* zeigt lediglich *mit Sorgfalt* (das nur im DWDS als Quasi-Synonym angeführt ist); es hat *auswählen* als häufigsten Kookkurrenzpartner**;** *aussuchen* (LLR 104) und *wählen* (LLR 85) zeigen ebenfalls starke Affinität. Demgegenüber haben *mit Umsicht* und *mit Überlegung* nur das Verb *wählen* als schwachen Kookkurrenzpartner (der bereits außerhalb der Tabelle liegt) und sowohl *mit Ruhe* als auch *mit Besonnenheit* haben *wählen* nicht unter ihren statistisch relevanten Kookkurrenzpartnern.

Allgemein kann eine semantisch-pragmatische Spezialisierung der einzelnen Adverbialgefüge auf bestimmte Verben bzw. Domänen festgestellt werden, die in diesem Rahmen nicht vertiefend analysiert, sondern nur in Hinweisen dargestellt werden kann. Wünschenswert wäre eine noch tiefer gehenden Analyse in Form einer Kontrastanalyse mit den entsprechenden Adjektivadverbien (*bedachtsam, überlegt, ruhig, besonnen, umsichtig, sorgfältig*), um Unterschiede zwischen Wortverbindung und potenziellem monolexematischem Äquivalent herauszuarbeiten, was allerdings den Rahmen dieses Artikels sprengen würde. Hier kann nur allgemein festgestellt werden, dass man zwar Übereinstimmungen, aber auch beträchtliche Unterschiede durch semantische Spezialisierungen findet. Die Ergebnisse sind oft durchaus überraschend, wie z.B. im Fall von *mit Überlegung*, das zum einen vornehmlich in negativen Kontexten auftritt (*ein Verbrechen/einen Mord mit Überlegung begehen*) und zum anderen stark auf den Fußball zugeschnitten ist, ebenso wie die Einwortentsprechung *überlegt* (*den Ball mit Überlegung/überlegt ins lange Eck schieben*).

3. LEXIKOGRAFISCHE DARSTELLUNG

Ausführlichkeit und Tiefe einer lexikografischen Darstellung variiert in Abhängigkeit vom angestrebten Adressatenkreis. Aus dem bisher Dargelegten können folgende Bedeutungskomponenten von *mit Bedacht* isoliert werden, die für die lexikografische Darstellung[74] von Bedeutung sind.

[74] Die lexikografische Darstellung von Mehrworteinheiten wird in Steyer (2013: 289-336) ausführlich behandelt. Vgl. auch allgemein zur aktuellen Lexikografie Engelberg/Lemnitzer 2008.

Den semantischen Kern, der stets vorhanden ist, kann man mit 'überlegtes Handeln' umschreiben. Hinzu kommen als weitere Komponenten 'ruhig', 'sorgfältig', 'umsichtig' und 'vorsichtig'. Je nach Kotext (sprachliche Umgebung) und Kontext (situative Einbettung) kann eine Bedeutungskomponente priorisiert werden. Im folgenden Beleg liegt die Betonung auf 'vorsichtig'.

NUZ03 Von dieser doppelten Vorsicht konnte das Leben abhängen. **Mit** äußerstem **Bedacht** näherte er sich den glitschigen Steinen

Neben der Semantik sind natürlich auch die usuellen Wortverbindungen entsprechend zu berücksichtigen.

Der eingangs zitierte Dude-onlinen stellt eine durchaus brauchbare Ausgangsbasis dar, die durch einige Ergänzungen (hier in spitze Klammern gesetzt) und eine Umstellung in der Mikrostruktur relativ problemlos verbessert werden kann:

Bedacht

<meist in den folgenden Wendungen>

- <sehr häufig> **mit Bedacht** (mit einer bestimmten Überlegung; überlegt, besonnen: mit Bedacht auswählen / <vorgehen>)
- <selten> **ohne Bedacht** (ohne jede Überlegung, unbesonnen, überstürzt: sie reagierte ohne Bedacht; <oft verneint: nicht ohne Bedacht>)
- <selten> **auf etwas [keinen] Bedacht nehmen** (auf etwas [nicht] bedacht sein, [nicht] achten: darauf müssen wir besonderen Bedacht nehmen)

In eine detailliertere Darstellung von *mit Bedacht* müssten die in dieser Studie erarbeiteten Resultate einfließen.

- Bedeutungsumschreibung: mit Überlegung handeln, etwas meist ruhig, sorgfältig, umsichtig und vorsichtig tun;
- Intensivierung: mit viel, großem, besonderem, Bedacht;
- Häufige Partner: *mit Ruhe und Bedacht, mit Sorgfalt und Bedacht, mit Umsicht und Bedacht, mit Vorsicht und Bedacht*;

- Meist in Verbindung mit den Verben *wählen, auswählen, aussuchen*: seine Worte mit Bedacht wählen; der Ort/Zeitpunkt/Termin war mit Bedacht gewählt;

- In Verbindung mit Kommunikationsverben: ich sage das mit Bedacht; einen Text mit Bedacht formulieren;

- In Verbindung mit Handlungsverben: mit Bedacht vorgehen, handeln; mit Bedacht an etwas herangehen.

4. SPANISCHE ENTSPRECHUNGEN

Im Spanischen existiert keine direkte Entsprechung, was auch nicht zu erwarten war, da *mit Bedacht* ein relativ großes semantisches Spektrum abdeckt und sehr kontextabhängig ist.

Die Suche nach Übertragungsmöglichkeiten[75] geht daher von den einzelnen prototypischen Verwendungen im Deutschen aus. Je nach Art des Verbs und des Ko- und Kontextes treten andere Übersetzungsmöglichkeiten in den Vordergrund.

Die in diversen zweisprachigen Wörterbüchern vorgeschlagenen Möglichkeiten sind weder eine ausreichende noch immer eine verlässliche Hilfe. Sowohl *pons-online* als auch *Langenscheidt Handwörterbuch* schlagen an erster Stelle *deliberadamente* vor, das aber, ähnlich wie *mit Überlegung*, aber im Gegensatz zu *mit Bedacht*, stark negativ konnotiert ist und daher in den wenigsten Fällen eine akzeptable Entsprechung darstellt. Alle hier vorgeschlagenen Äquivalente sind in spanischen Korpora[76] verifiziert.

In Verbindung mit den wichtigsten Kookkurrenzpartnern *wählen, auswählen, aussuchen* können im Spanischen im Allgemeinen (*elegir, seleccionar) cuidadosamente / conscientemente / a conciencia / con buen criterio* verwendet werden, wie etwa für den folgenden Korpusbeleg:

WPD11 Gedenktafel für das Dreiländertreffen Der Tagungsort war **mit Bedacht** gewählt. Der weite Blick ins Land und vor allem die wenige

Für Beispiele wie das folgende, in dem *mit Bedacht* durch *sehr viel* intensiviert ist, bieten sich Entsprechungen wie *después de mucho reflexionar / después de largas reflexiones; después de una larga reflexión / tras (larga) reflexión*.

VDI10 Wir haben unsere acht Modellregionen **mit** sehr viel **Bedacht** aus 130 Bewerbungen ausgewählt.

[75] Für zahlreiche Hinweise und Vorschläge bedanke ich mich bei meiner Kollegin Cecilia López Roig.
[76] Verwendet wurden folgende Korpora:
CREA http://corpus.rae.es/creanet.html

Corpus del español http://www.corpusdelespanol.org/

Wird die Bedeutungskomponente ‘vorsichtig’ besonders hervorgehoben, kann man spanische Entsprechungen verwenden wie: *con prudencia / con sensatez / con precaución*.

NKU11 das Gold irgendwohin einschicken soll. "Die Konsumenten sollten **mit Bedacht** auswählen, wem sie ihr Gold anvertrauen", sagt Dünkelmann.

Einige Beispiele mit Kommunikationsverben (*formulieren, sagen, schreiben*) zeigen teils gleiche, teils andere Äquivalente. Im folgenden Beleg stellt *a conciencia* eine gute Lösung dar:

M06 ist im Grundgesetz verankert. Das haben unsere Verfassungsväter **mit Bedacht** so formuliert. Dem Schwachen wird z.B. im

Im folgenden Beispiel steht die Bedeutungskomponente ‘vorsichtig’ im Vordergrund, daher kann man im Zusammenhang mit Kommunikationsverben *con cautela / con precaución* verwenden.

WDD11 Da das ein heikles Thema ist, das **mit Bedacht** formuliert werden sollte, bitte ich mal

Steht die Bedeutungskomponente ‘sorgfältig’ im Vordergrund, bietet sich *con esmero* als Äquivalent an:

T04 um ihre Liebhabersammlung zu komplettieren. Sie formulieren **mit Bedacht** eine Suchanfrage: Nicht an Google, sondern an alle

Die Kombination „sauber und mit Bedacht“ im folgenden Beispiel lässt sich treffend mit *de forma clara y cautelosa* übertragen.

WDD11 Man muss da sehr sauber und **mit Bedacht** formulieren, wenn man nicht provozieren will

Für die im Deutschen hochfrequente Wortverbindung „ich sage das mit Bedacht“ ist im Spanischen *a conciencia* eine übliche Entsprechung.

PBW Die zweite Anmerkung ist die, meine Damen und Herren -ich sage das **mit Bedacht** Nach meiner festen Überzeugung bemüht sich die

Für Handlungsverben (*vorgehen, angehen, herangehen, handeln, ...*) kann im Allgemeinen verwendet werden: *(actuar) con sensatez / de forma sensata / prudente / responsable / con prudencia / con sentido común*.

Tritt die Komponente ‘vorsichtig’ in den Vordergrund, sind folgende Äquivalente treffend: *con cautela / de forma prudente*.

RHZ01 Wir wollen, dass **mit Bedacht** vorgegangen und Krieg vermieden wird. Unsere Gedanken sind prudente

Die folgende Kombination „ruhig und mit Bedacht“ kann folgendermaßen übertragen werden: *con serenidad y sensatez / de forma tranquila y sensata*.

> RHZ13 Wir sollten die Sache ruhig und **mit Bedacht** angehen, anstatt jetzt voreilige Schlüsse zu ziehen,

Eine Zusammenstellung der vorgeschlagenen Äquivalente, die durchaus nicht als vollständig anzusehen sind, macht deren lexikalische und strukturelle Vielfalt bewusst und legt Zeugnis ab von der komplexen mentalen Arbeit, die ein Übersetzer verrichten muss.

concienzudamente	de forma clara y cautelosa
conscientemente	de forma prudente
cuidadosamente	de forma responsable
	de forma sensata
a conciencia	de forma tranquila y sensata
con cautela	después de mucho reflexionar
con buen criterio	después de largas reflexiones
con (mucho) cuidado	después de una larga reflexión
con precaución	tras (larga) reflexión
con prudencia	
con sensatez	prudente y concienzudamente
con sentido común	
con seriedad y sensatez	

5. SCHLUSSFOLGERUNGEN

Korpusanalysen sind heute als Basis für Lexikologie und Lexikografie unerlässlich. Sie eröffnen nicht nur Einsichten in die Semantik, sondern decken auch zu sprachlichen Gewohnheiten gewordene bevorzugte Ausdrucksweisen auf, also usuelle Wortverbindungen. Man erkennt Wortkombinationen mit übersummativer Bedeutung, also Phraseme, die bislang aufgrund ihrer Unauffälligkeit allein mithilfe der Intuition nur schwer und unsicher als solche identifiziert werden konnten.

Hinsichtlich der Fixiertheit stellt sich heraus, dass einerseits die interne Festigkeit des hier untersuchten Phrasems nicht absolut ist, dass aber andererseits eine gewisse externe Festigkeit vorliegt, die darin zu sehen ist, dass das Phrasem in überraschend hohem Maß mit bestimmten Verben (*wählen auswählen, aussuchen*) und Verbgruppen (Handlungsverben und Kommunikationsverben) kookkurriert.

Zusammenfassend kommt man zur Einsicht, dass insgesamt ein breites Spektrum sprachlicher Fixiertheit angenommen werden muss, ausgehend vom Zentrum der Phraseologie mit opaken, nicht motivierten, festen Idiomen, bis hin zur Peripherie, deren Auslotung eben erst begonnen hat und die noch so manche neue Erkenntnisse verspricht.

Literatur

BROCKHAUS WAHRIG, 2011. *Deutsches Wörterbuch*. Gütersloh: wissenmedia.

BURGER, H., 2010. *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt Verlag.

DEUTSCHES REFERENZKORPUS (DeReKo). [online] https://cosmas2.ids-mannheim.de/cosmas2-web/ [22.07.2015].

DOBROVOL'SKIJ, D. / PIIRAINEN, E. (1994). Sprachliche Unikalia im Deutschen: Zum Phänomen phraseologisch gebundener Formative, *Folia Linguistica 27*(3,4), 449-473.

DUDEN ONLINE. [online] http://www.duden.de/woerterbuch [15.02.2013].

DUDEN DEUTSCHES UNIVERSALWÖRTERBUCH, 2011. Mannheim (u.a.): Dudenverlag.

DWDS. *Digitales Wörterbuch der deutschen Sprache*. [online] <http://www.dwds.de> [22.07.2015].

ENGELBERG, St., LEMNITZER, L., 2008. *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

FLEISCHER, W., 1997. *Phraseologie der deutschen Gegenwartssprache*. Tübingen: Niemeyer.

HELBIG, G. /BUSCHA, J., 2001. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Berlin (u.a.): Langenscheidt-Verlag Enzyklopädie.

HOLZINGER, H.J., 2013. Unikale Elemente: Eine Herausforderung für Lexikologie und Lexikografie. *Aussiger Beiträge* 7, pp. 53-66.

HOLZINGER, H.J., 2012. *Unikale Elemente*: Apuntamentos sobre as palabras ligadas fraseoloxicamente do alemán actual. *Cadernos de Fraseoloxia galega* 14, pp. 165-173. [online] http://www.cirp.es/pub/docs/cfg/cfg14.pdf [22.07.2015].

KISS, T. et al., 2014. *Ein Handbuch für die Bestimmung und Annotation von Präpositionsbedeutungen im Deutschen*. [online] http://www.linguistics.ruhr-uni-bochum.de/bla/014-kiss_etal2014.pdf2 [22.07.2015].

LANGENSCHEIDT HANDWÖRTERBUCH SPANISCH. 2006. Berlin (u.a.): Langenscheidt.

PONS ONLINE. [online] http://es.pons.com/traducción [22.07.2015].

SAILER, M. / TRAWIŃSKI, B., 2006. Die Sammlung unikaler Wörter des Deutschen. Aufbauprinzipien und erste Auswertungsergebnisse. In: A. Häcki Buhofer/H. Burger (Hrsg.): *Phraseology in Motion I. Methoden und Kritik. Akten der Internationalen Tagung zur Phraseologie (Basel 2004)*. Hohengehren: Schneider Verlag, 439-450.

SCHRÖDER, J., 1990. *Lexikon deutscher Präpositionen*. Leipzig. Verlag Enzyklopädie.

SÖHN, J.-Ph., 2003. *Von Geisterhand zu Potte gekommen. Eine HPGS-Analyse von PPs mit unikaler Komponente*. Magisterarbeit. Universität Tübingen. [online] http://publikationen.ub.uni-frankfurt.de/volltexte/2008/111147 [22.07.2015].

STEYER, K., 2013. *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr.

STEYER, K. 2004: Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: K. Steyer (Hrsg.): *Wortverbindungen – mehr oder weniger fest*. Institut für Deutsche Sprache. Jahrbuch 2003. Berlin/New York: de Gruyter.

STUMPF, S., 2014. *Mit Fug und Recht*. Korpusbasierte Erkenntnisse zu phraseologisch gebundenen Formativen. *Sprachwissenschaft* 39, 1: 85-114.

ZIFONUN, G./HOFFMANN, L./STRECKER, B., 1997. *Grammatik der deutschen Sprache. Band 3*. Berlin/New York: de Gruyter.

FRASEOLOGISMOS DE COLOR EN ESPAÑOL Y RUSO: ESTUDIO DE FRASEOLOGISMOS SIN ANÁLOGOS EN LA OTRA LENGUA

Anastasia Kovaleva
Universidad de Málaga
akovaleva@uma.es

Resumen

El estudio del aspecto contrastivo de la fraseología siempre ha provocado cierto interés en los investigadores dada la gran variedad de vías de investigación que se presentan en este ámbito. En los últimos años, dentro de este enfoque, se destaca la tendencia a aplicar los avances de la lingüística de corpus con el fin de investigar sobre las similitudes y las diferencias entre fraseologismos de lenguas distantes, basándose en una comparación del funcionamiento de los fraseologismos, desde el punto de vista pragmático-discursivo, realizada a través de la aplicación de técnicas y herramientas de PLN (Procesamiento de Lenguaje Natural), con especial referencia al uso de corpus. Nuestro trabajo se enmarca, precisamente, en estos nuevos enfoques de los estudios fraseológicos basados en corpus. Siguiendo estas tendencias, el presente trabajo se centra en un estudio de fraseologismos acromáticos (unidades fraseológicas con blanco o negro entre sus componentes) y fraseologismos cromáticos (unidades fraseológicas con colores básicos entre sus componentes) en el par de lenguas español-ruso. Se definen y presentan fraseologismos de color y se estudia el grupo de los fraseologismos que disponen de equivalentes fraseológicos que no denotan colores en la otra lengua. Se trata de los fraseologismos del tipo: *черный монах* ("monje negro"), *белая горячка* (resultado del síndrome de abstinencia del alcohol), *увидеть белый свет в клеточку* ("ver la luz blanca en cuadraditos"), *зеленый змей* ("serpiente verde") en ruso y *quedarse en blanco* y *mangas verdes* en español. Frecuentemente estas unidades presentan dificultades para traducción ya que reflejan realidades u objetos propios de una de las dos culturas.

El estudio se realiza de acuerdo a los principios y técnicas de investigación propias de la fraseología computacional, con especial referencia a la metodología de corpus. La base para el análisis de los fraseologismos acromáticos y cromáticos la constituyen los estudios de G. Corpas Pastor dedicados a esta rama de investigación fraseológica. Los datos se obtienen mediante la consulta a corpus y sistemas en línea que proporcionan más información que los diccionarios bilingües y monolíngües. Se utiliza el CREA (Corpus de referencia del español actual), el CNR (Corpus Nacional Ruso) y el corpus basado en la red Internet y gestionado por el programa WebCorp, herramienta de gestión de corpus que resulta imprescindible para analizar las combinaciones que se caracterizan por su baja frecuencia de aparición en el discurso. La aplicación de la metodología de corpus para el análisis de los resultados obtenidos ofrece una información fraseológica contrastiva relativa a los dos universos fraseológicos más completa dado que este método proporciona datos acerca de uso y el funcionamiento pragmático-discursivo de los fraseologismos de color.

1. INTRODUCCIÓN

El estudio del aspecto contrastivo de la fraseología siempre ha provocado cierto interés en los investigadores dada la gran variedad de vías de investigación que se presentan en este ámbito. En los últimos años, dentro de este enfoque, se destaca la tendencia a aplicar los avances de la lingüística de corpus con el fin de investigar sobre las similitudes y las diferencias entre fraseologismos de lenguas distantes, basándose en una comparación del funcionamiento de los fraseologismos, desde el punto de vista pragmático-discursivo, realizada a través de la aplicación de técnicas y herramientas de PLN (Procesamiento de Lenguaje Natural), con especial referencia al uso de corpus. Nuestro trabajo se enmarca, precisamente, en estos nuevos enfoques de los estudios fraseológicos basados en corpus. Siguiendo estas tendencias, el presente trabajo se centra en un estudio de fraseologismos acromáticos (unidades fraseológicas con blanco o negro entre sus componentes) y fraseologismos cromáticos (unidades fraseológicas con colores básicos entre sus componentes) en el par de lenguas español-ruso. Se definen y presentan fraseologismos de color y se estudia el grupo de los fraseologismos que disponen de equivalentes fraseológicos que no denotan colores en la otra lengua. Se trata de los fraseologismos del tipo: *черный монах* ("monje negro"), *белая горячка* (resultado del síndrome de abstinencia del alcohol), *увидеть белый свет в клеточку* ("ver la luz blanca en cuadraditos"), *зеленый змей* ("serpiente verde") en ruso y *quedarse en blanco* y *mangas verdes* en español. Frecuentemente estas unidades presentan dificultades para traducción ya que reflejan realidades u objetos propios de una de las dos culturas.

El estudio se realiza de acuerdo a los principios y técnicas de investigación propias de la fraseología computacional, con especial referencia a la metodología de corpus. La base para el análisis de los fraseologismos acromáticos y cromáticos la constituyen los estudios de Corpas Pastor (2013) dedicados a esta rama de investigación fraseológica. Los datos se obtienen mediante la consulta a corpus y sistemas en línea que proporcionan más información que los diccionarios bilingües y monolíngües. Se utiliza el CREA (Corpus de referencia del español actual), el CNR (Corpus Nacional Ruso) y el corpus basado en la red Internet y gestionado por el programa WebCorp, herramienta de gestión de corpus que resulta imprescindible para analizar las combinaciones que se caracterizan por su baja frecuencia de aparición en el discurso. La aplicación de la metodología de corpus para el análisis de los resultados obtenidos ofrece una información fraseológica contrastiva relativa a los dos universos fraseológicos más completa dado que este método proporciona datos acerca de uso y el funcionamiento pragmático-discursivo de los fraseologismos de color.

2. LOS FRASEOLOGISMOS CROMÁTICOS DE COLOR ROJO

En el presente estudio dividimos los fraseologimos de color en cromáticos y acromáticos. Dicha división se basa en dos teorías importantes sobre las formas de expresar los conceptos de colores en distintas lenguas. De acuerdo con la primera teoría desarrollada por Berlin y Kay (1969), existe una relación entre categorías de colores y los patrones neurológicos encargados de la percepción del color y se establecen once colores relevantes para todas las lenguas: seis colores básicos o fundamentales, ubicados en cuatro grupos (*stages*): blanco y negro; rojo; verde o amarillo; azul; y otro grupo adicional que reúne los colores no básicos que existen en todas las culturas: rosa, marrón, violeta (morado) y gris. Por otro lado, Küppers (1972) en su teoría fijó ocho colores elementales que coinciden con ocho combinaciones posibles de percepción del órgano de la vista. Son tres colores primarios (el rojo, el verde, el azul), tres secundarios (el amarillo, el azul cian, el magenta) y

dos colores acromáticos, el blanco que corresponde a la combinación de los tres primarios, y el negro, que representa la ausencia de los tres.

Tomamos estas teorías de base para nuestra investigación y dividimos los fraseologismos en dos grupos: fraseologismos cromáticos (FC) y fraseologismos acromáticos (FA), tomando en cuenta también el componente de color que contengan. A continuación ofrecemos algunos ejemplos de cada para el par de lenguas español-ruso:

(1) Fraseologismos cromáticos: **E.** *al rojo vivo*, *príncipe azul, pasarlas moradas, humor amarillo, mangas verdes* | **R.** *провести красную линию* ("trazar la línea roja"), *зеленей горькой полыни* ("más verde que artemisia amarga"), *голубая кровь* ("sangre azul"), *синий чулок* (lit. 'media azul', "[mujer] marisabidilla").

(2) Fraseologismos acromáticos: **E.** *mirlo blanco, oveja negra*, *de punta en blanco*, *parecerse en el blanco de los ojos, estar negro por algo,* | **R.** *не видеть света белого* ("no ver la luz blanca"), *шито белыми нитками* (lit. 'cosido con hilos blancos', "[estar] poco disimulado"), *белая кость* ("hueso blanco"), *на чёрный день* ("para un día negro").

Cuando nos referimos a los fraseologismos que no tienen análogos en la otra lengua: se trata de los fraseologismos del tipo: *черный монах* ("monje negro"), *белая горячка* (resultado del síndrome de abstinencia del alcohol), *увидеть белый свет в клеточку* ("ver la luz blanca en cuadraditos"), *зеленый змей* ("serpiente verde"), *красная цена* ("precio rojo"), *красный петух* ("gallo rojo") en ruso *y quedarse en blanco*, *alerta roja, marea roja y mangas verdes* en español.

El objetivo de este estudio es realizar un análisis empírico con la ayuda de la metodología de corpus y ofrecer los datos completos sobre la naturaleza y funcionamiento de los fraseologismos cromáticos con componente de color rojo que no cuentan con equivalentes o análogos en la otra lengua dada la ausencia del fenómeno o acción expresados por el fraseologismo o por la ausencia de equivalente fraseológico en la otra lengua.

Por otro lado, dicha categoría de unidades fraseológicas presenta ciertas dificultades a la hora de traducción. Generalmente, estas dificultades están relacionadas con falta de datos sobre la semántica y el uso de estas unidades en diccionarios a los que se recurre a la hora de traducirlos. Precisamente este último aspecto sirve de punto de partida para nuestra investigación: existe una necesidad de rellenar algunos vacíos de información acerca del funcionamiento y naturaleza de estos fraseologismos.

3. LA METODOLOGÍA DE ANÁLISIS

La metodología de análisis que seguimos supone tres fases: la fase de localización y selección de los fraseologismos en las fuentes lexicográficos bilingües y monolingües, la fase de estudio de cada fraseologismo desde una perspectiva pragmático-discursiva con metodología de corpus, contrastando los resultados y haciendo resúmenes generales de esta categoría de fraseologismos, su estructura y funcionamiento. Para realizar el análisis de dichas unidades, seguimos el modelo contrastivo propuesto por Corpas Pastor (2013), que se basa en el análisis de ejemplos procedentes de corpus y de Internet mediante el uso de

programas y herramientas de gestión de la red, adaptándolo a la específica de las unidades fraseológicas que nos ocupan. De esta manera, primero se ofrece una interpretación inicial y etimología del fraseologismo o fraseologimos seleccionados de los repertorios lexicográficos disponibles. Posteriormente se estudia la forma de cada fraseologismo en el corpus, se detectan las tendencias de uso generales, los patrones combinatorios, los casos de variación y las variantes posibles que se encuentran para la unidad seleccionada entre los resultados de los corpus y mediante el programa de gestión de corpus WebCorp. En el último lugar se analizan casos de modificaciones semánticas, formas poco comunes del fraseologismo en cuestión y se estudia valor pragmático de fraseologismo.

Para este estudio nos basaremos en dos corpus monolingües en línea y un sistema de gestión de la red como corpus:

a) el CREA-*Corpus de Referencia del Español Actual* (http://www.corpus.rae.es/creanet.html);

b) el CNR-*Национальный корпус русского языка* [Corpus Nacional Ruso] (http://ruscorpora.ru/index.html);

c) WebCorp (http://www.webcorp.org.uk/live/).

Dadas las limitaciones de espacio no será posible ofrecer todos los datos de estudio dentro de este trabajo, por lo tanto, a continuación exponemos nuestra propuesta metodológica de una forma resumida y ofrecemos análisis de fraseologismos que aparecen en los textos relacionados con el mundo de economía: *красная цена* ("precio rojo") ruso y *en números rojos* en español.

4. EL ESTUDIO DE LOS FRASEOLOGISMOS

4.1 El estudio de la secuencia [КРАСНАЯ ЦЕНА] ("precio rojo")

El FSRLJ recoge el fraseologismo *красная цена* ("precio rojo") bajo la entrada para el sustantivo, definido como: "el precio más alto que se puede ofrecer por algo, lo máximo que se puede obtener por algo o alguien". En este caso existe una aparente relación entre este fraseologismo y el color rojo, sin embargo, tras una búsqueda en diccionarios esta conexión no encuentra confirmación ya que, según nos consta, esta unidad se formó sobre el concepto "bonito, agradable"[77].

Entre los diccionarios bilingües consultados tan solo dos (el DFRE y el metadiccionario electrónico) incluyen *красная цена* ("precio rojo") y ofrecen como equivalente de traducción *el mejor precio, el precio máximo*. Además el metadiccionario dispone de ejemplos de uso del fraseologismo ruso, extraídos de periódicos digitales del año 2006. Ambos ejemplos corresponden a la acepción mencionada anteriormente. De esta manera, el fraseologismo en cuestión no guarda ninguna relación con el color rojo.

[77] El significado de la palabra "rojo" es posterior al significado original de dicha palabra "bonito" (Fasmer, 1996).

El fraseologismo *красная цена* ("precio rojo") tiende a aparecer en textos periodísticos y literarios, y su uso se guía por los siguientes aspectos principales: 1) puede aparecer en todas las formas posibles de sustantivo de número singular, 2) anteriormente no tenía formas plurales, 3) se combina con sustantivos que definen cantidades de dinero y aparecen en formas de caso acusativo, lo que podemos apreciar en los siguientes ejemplos:

6: него... один из... волжских пароходов, которому **красная цена** 60000 рублей.
10: Марка семикопеечного достоинства. Всей то книге **красная цена** пятак.
25: Хмельницкого, 79) во время действия акции "**красная цена**" удимился, что хоть она действует по 17
15: STONEX (только до 30 июня 2014г.): Смотрите **красные цены** в нашем прайс-листе!!! А также
22: выгодные и привлекательные акции, устанавливаем «**красные цены**» и значительные скидки.

Tabla 2. Concordancias alineadas a 1-IZQ para la secuencia [КРАСНАЯ ЦЕНА] (WebCorp)

Los resultados del análisis indican la coexistencia de un fraseologismo polisémico *красная цена* ("precio rojo") que denomina un precio alto o bien una oferta o bien una colocación homónima que define una marcación de productos o mercancías. La mayoría de los ejemplos corresponden al primer significado que denota el precio más alto para un producto o un servicio. En estos casos, la secuencia entra a formar parte de cláusulas en función de complemento directo (locución nominal) o bien como atributo de un sintagma nominal. Cuando funciona como locución nominal, suele colocarse con a) verbos que relacionados con el área de compra y venta (предлагать, действовать, установить); y b) sustantivos relacionados con el mismo tema (товары, продукты, скидки, предложения). El patrón verbal puede ser transitivo: 1. VB + [КРАСНАЯ ЦЕНА], por ejemplo, "искать красную цену"; o 2. VB + SN +PREP. + [КРАСНАЯ ЦЕНА], por citar un ejemplo, "заказать окно по красной цене". Los sustantivos que suelen coaparecer con la secuencia en función de objeto directo y formar colocaciones concatenadas son sustantivos que hacen referencia a cantidades de dinero *(пятак* "moneda de cinco", *4 доллара* "4 dolares", *копейка* "una copeca", *50 тысяч* "50 mil"), nombres de algunos productos y mercancías (*окна* "ventanas", *квартиры* "apartamentos"), sustantivos que denominan tipos y categorías de productos (*продукты, товары, ассортимент*), servicios *(лечение* "tratamiento médico"). Las formas verbales que junto con el fraseologismo forman una colocación corresponden a la tercera y la segunda persona singular y plural del presente de indicativo, no encontramos formas del futuro y contamos con tan solo un ejemplo con la forma del pасado de indicativo, como se observa en las líneas de concordancia que se presentan a continuación:

в сентябре мы предлагаем вам заказать окна по **красным ценам**!
Акция **«Красная цена»** действует до 30 сентября!
одеждуфирмы Didriksons! Скидка действует и на **красные цены**.
В интернете множество сайтов пестрят **красными ценами** на эту модель: "Машинка

Tabla 3. Líneas de concordancias que muestran colocaciones para la secuencia [КРАСНАЯ ЦЕНА] en función de complemento directo

Los resultados extraídos de internet a través de WebCorp no solo ofrecen ejemplos más usuales para las unidades en cuestión, sino también proporcionan otros ejemplos de uso que no aparecen en diccionarios y, por lo tanto, complementan el significado de la unidad.

Algunos de los ejemplos encontrados atestiguan unos cambios en la semántica del fraseologismo en cuestión. Así por ejemplo, el fraseologismo *красная цена* ("precio rojo") sufre ciertas transformaciones semánticas que se basan en el cambio de contexto de uso de esta unidad (mundo de publicidad, negocios), lo cual se refleja en la estructura del fraseologismo:1) en numerosas ocasiones el fraseologismo aparece marcado por comillas o bien se marca solo la palabra rojo: "во время действия акции "красная цена"" , "Также ищите на сайте специальные "красные" цены" ("también buscad en la página web precios rojos"); 2) aparece en las combinaciones imposibles anteriormente (*Опубликована красная цена* "precio rojo publicado", *красная цена действует до* "precio rojo está vigente hasta") 3) aparece en formas de plural que corresponden a distintos casos (anteriormente este fraseologismo no se aplicaba en forma de plural*): Охота на красные цены* ("caza por precios rojos")*, заказать окна по красным ценам* "encargar ventanas por precios rojos"*, предлагает красные цены на новые окна* "ofrece precios rojo para ventanas nuevas" y etc.).

La valoración semántica negativa de este fraseologismo se observa en numerosos ejemplos ya que generalmente se aplica para desaprobar o bien despreciar el valor que se le tienen a algo otras personas, se trata tanto de objetos, como de ideas abstractas, tales como, amistad, actividad, trabajo e incluso se aplica para referirse a algunas personas:

надо думать, оценил ситуацию — и себя: **<u>красная цена</u>** ему, А сама ни разу на занятиях не была.
Я ее оглядел с ног до головы и думаю: тебе **<u>красная цена</u>** в базарный день — пять десят тугриков,
Собственно — **<u>красная цена</u>** всей вашей деятельности…

Tabla 4. Líneas de concordancias que muestran valoración negativa para la secuencia [КРАСНАЯ ЦЕНА]

4.2 El estudio de la secuencia [EN NÚMEROS ROJOS]

La unidad española *en números rojos* que se define en el DRAE como: *"s*aldo negativo en una cuenta bancaria o en una contabilidad empresarial". Se trata de una unidad monosémica que se combina con sustantivos que denotan cantidades de dinero (*millones de euros*), pertenecen al mundo de economía (*cuentas, pasado ejercicio*, *saldo*, *caja, banco* etc) o bien caracterizan un período de tiempo (*mes, año, temporada*). También se suele colocar con los verbos *estar, salir, entrar, quedarse, cerrar, encontrarse, mantenerse.* El patrón verbal puede ser intransitivo: 1. VB + [EN NÚMEROS ROJOS], por ejemplo, "quedarme en números rojos"; o transitivo: 2. VB + SN + [EN NÚMEROS ROJOS], por ejemplo, "deja al equipo en números rojos". Lo cual se observa en los siguientes ejemplos extraídos con WebCorp:

: (0,25 %), el resto de sectores cerraron **<u>en numeros rojos</u>**, entre los que destacaron el de empresas de
10: del campo. Esta gran inversion deja al equipo **<u>en numeros rojos</u>**, pero se tiene esperanzas en que futuras
8: cada uno lo que consuma. Puesto que estabamos **<u>en numeros rojos</u>** en las cuentas de la comunidad, con el objeto de
6: Vivo 2012 Contáctanos Top El Planeta entra hoy **<u>en</u>**

numeros rojos: ya hemos consumido nuestro capital natural para
11: que la clase empresarial inutil que mantiene **en numeros rojos** a las empresas sí merece cobrar mas alla del

Tabla 5. Concordancias alineadas a 1-IZQ para la secuencia [EN NUMEROS ROJOS] (WebCorp)

Las formas verbales que entran en colocación con el fraseologismo suelen estar restringidas a las formas de tercera persona singular del presente o tercera persona singular y plural del pasado de indicativo, o bien son infinitivos. En ciertas ocasiones el hablante intenta definir el hecho de quedarse en números rojos como consecuencia de abusos de bancos o empresas. En estos casos se usan las formas verbales de tercera persona singular con sujeto que constituye el sustantivo *cuenta.*

La valoración de este fraseologismo es negativa, en numerosas ocasiones los anunciados que disponen de este fraseologismo entre sus componentes se caracterizan por las funciones de los sujetos agentes de la acción que, por norma general, se consideran como los culpables de la situación, suelen ser políticos y empresarios, así como ciudadanos que se encuentran en una situación complicada por la pérdida de empleo o la crisis económica, lo cual podemos observar en las siguientes líneas de concordancia:

3: ransformación de la compañía Nokia vuelve a estar **en 'numeros rojos'**, aunque con "signos de mejoría" Ha generado
42: luego no es por mi culpa que la Comunidad este **en numeros rojos**, eso es para los politicos que por desgracia son
27: dos cuestiones a resolver: La primera: estoy **en numeros rojos** y necesito vender, acabo fichar a Uche...no se
38: aunque ponga de su voluntad, consiguen meter **en numeros rojos** a la empresa, les meten una carga de trabajo del

Tabla 6. Líneas de concordancias que muestran colocaciones para la secuencia [EN NÚMEROS ROJOS]

La deixis en primera persona se observa en los casos cuando el hablante se refiere a sí mismo como víctima de una error de bancario o como en el siguiente fragmento, de un error de la aseguradora, por lo cual, en el discurso toda la culpa recae sobre la aseguradora y el banco[78].

> *llevo un cabreooo!!! sabeis q pasa si estas una semana en numeros rojos? me han girado el seguro de mi hermano a una cuenta q tengo... q me ingresan la nomina pero desvio el dinero a otra...*

En algunas ocasiones presenciamos la combinación del fraseologismo crómatico de color rojo con otros fraseologismos:

[78] http://www.webcorp.org.uk/live/view.jsp?query=en+numeros+rojos&url=http%3A%2F%2Fforos.vogue.es%2Fviewtopic.php%3Ft%3D124201 [Fecha de consulta: 21/06/2015].

El negocio bancario español entra de lleno en números rojos. Bancos y cajas apenas ganaron 82 millones de euros en 2011.[79]

Este titular de un portal de noticias de economía tiene como el objetivo llamar la atención de los lectores a través de la intensificación semántica del fraseologismo *en números rojos* con la locución adverbial *de lleno* que no solo cumplimenta el significado de otro fraseologismo, indicando la gravedad de la situación, sino también transmite una valoración negativa más fuerte.

5. CONCLUSIONES

En este trabajo hemos estudiado los fraseologismos cromáticos, unidades fraseológicas con componente de color rojo que no disponen de análogos en la otra lengua mediante la aplicación de una metodología de análisis basada en corpus. En líneas generales, podemos decir que ambas lenguas presentan un número considerable de variantes y formas. Los fraseologismos estudiados están formados por componentes del léxico usual de la lengua general, es decir, dentro de las unidades estudiadas no se hallan componentes que pertenezcan al lenguaje arcaico o formas anticuadas que ya se encuentran fuera de las costumbres de uso de la lengua actual.

Desde una perspectiva semántica las UF estudiadas se alejan del concepto de color y obtienen un significado nuevo, en el caso del fraseologismo *красная цена* ("precio rojo") la referencia al color es aparente ya que realmente esta unidad no hacía referencia al color rojo.

Asimismo, los datos proporcionados por WebCorp ofrecen nuevas formas y modelos de uso de fraseologismo. Así, por ejemplo, anteriormente no se había registrado el uso de las formas plurales del fraseologismo *красная цена* ("precio rojo") o la existencia de combinaciones imposibles anteriormente (*опубликована красная цена* "precio rojo publicado", *красная цена действует до* "precio rojo está vigente hasta", предлагает красные цены "ofrece precios rojos").

Finalmente, conviene destacar la importancia del estudio en el campo de la investigación fraseológica con la ayuda de la metodología de corpus ya que trabajos de este tipo escasean en la fraseología contrastiva. Asimismo, apenas encontramos estudios de fraseología contrastiva ruso-española o dedicados a los fraseologismos de color en este ámbito. Por tanto, este trabajo podría considerarse como una de las primeras aproximaciones a la fraseología contrastiva con corpus en el par de lenguas ruso-español. Esperamos que esta vía de investigación sea un campo fructífero para futuros estudios.

[79] http://www.libremercado.com/2012-03-26/el-negocio-bancario-espanol-entra-de-lleno-en-numeros-rojos-1276454102 / [Fecha de consulta: 06/2015]

Bibliografía

ALEFIRENKO, N. AND SEMENENKO N., 2009. *Fraseologuija i Paremiologuija* [Фразеология и паремиология]. Moscú: Flinta, Nauka.

ARSENTIEVA, E., 2006. Fraseologuija y fraseografia v sopostavitelnom aspecte: (na materiale russkogo y anglijskogo yazikov [Фразеология и фразеография в сопоставительном аспекте: (на материале рус. и англ. яз.)]. Kazán: Universidad de Kazán.

BARANOV, A. AND DOBROVOL'SKIJ, D., 2008. *Aspecti teorii fraseologuii* [Аспекты теории фразеологии]. Moscú: Znak.

BERLIN, B. AND KAY, P., 1969. *Basic Color Terms: their Universality and Evolution*. Berkeley y Los Angeles: University of California Press.

CORPAS PASTOR, G., 1996. *Manual de fraseología española*. Biblioteca Románica Hispánica. Manuales. Madrid: Gredos.

CORPAS PASTOR, G., 2008. *Traducir con corpus: los retos de un nuevo paradigma*. Fráncfort: Peter Lang.

CORPAS PASTOR, G., 2013. Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. En I. Olza and E. Manera, ed. 2013. Fraseopragmática, Berlín: Frank & Timme. pp. 335-373.

CORPAS PASTOR, G., 2014. El fraseólogo internauta: cómo pasarlo pipa en la red. En J. Sevilla Muñoz, ed. 2014. Fraseología y paremiología: enfoques y aplicaciones. Madrid: Instituto Cervantes pp. 133-152.

DMITRIEV D., 2003. *Tolkovij slovar' russkogo jazika* [Толковый словарь русского языка]. Moscú: Astrel. [TSRJAZ]

FASMER M., 1996. *Etimologicheskij slovar´russkogo jazika* [Этимологический словарь русского языка]. San Petersburgo.

FEDEROV, A., 2008. *Fraseologicheskij slovar' russkogo literaturnogo jazika* [Фразеологический словарь русского литературного языка]. Moscú: Astrel. [FSRLJ]

FELITSINA, V., [1990] 1994. *Russkie fraseologismi*. [Русские фразеологизмы]. Moscú: Russkij yazik. [RF]

GUILLÉN MONJE, G., 2004. *Fraseología contrastiva ruso-española: análisis del corpus bilingüe de somatismos*. Tesis doctoral. Universidad de Granada.

KÜPPERS, H., 1971. *Fundamentos de la teoría de los colores*. Barcelona: Gustavo Gili SA.

Metadiccionario bilingüe español-ruso/ ruso-español [online] Available at: http://www.diccionario.ru/ [Accessed 1 June 2015].

MOLINER, M., 2007. *Diccionario de uso del español*. 3ª ed. Madrid: Gredos. [online]. Available at: http://www.diclib.com/ [Accessed 25 May 2015]. [DUE].

MOLOTKOV, A., 1967. *Fraseologicheskij slovar' russkogo jazika* [Фразеологический словарь русского языка]. Moscú: Russkij jazik. [FSRJ]

OZHEGOV, S. AND SHVEDOVA, N., 1992. *Tolkovij slovar' russkogo yazika* [Толковый словарь русского языка]. Moscú: Az. [TSRJ]

REAL ACADEMIA ESPAÑOLA (2001): *Diccionario de la lengua española,* 22ª ed. Madrid: Espasa. [DRAE]

SECO, M., RAMOS, G. AND ANDRES, O., 1999. *Diccionario del español actual*. Madrid: Aguilar.

SECO, M., RAMOS, G. AND ANDRES, O., 2004. *Diccionario fraseológico documentado del español actual*. Madrid: Aguilar.

SIEFRING J., [2004]1994. *Oxford dictionary of idioms*. Nueva York: Oxford University Press. [OD]

STEPANOVA, M., 2010. *Schkol'nij frazeologicheskij slovar'* [Школьный фразеологический словарь]. Rostov del Don: Fenix. [SFS]

SZAlEK, J., 2005. Los colores y su semántica en las expresiones fraseológicas españolas. In: Studia Romanica Posnanienska. 32. pp. 87-96.

TELIA, V., 1996. *Russkaja fraseologuija* [Русская фразеология]. Moscú: Shkola "Yazyki russkoj kultury".

TIMOFEEVA, L., 2008. *Acerca de los aspectos traductológicos de la fraseología española*. Tesis doctoral. Alicante: Universidad de Alicante.

TUROVER, G. AND NOGUEIRA, J., 2000. *Gran Diccionario Ruso-español* [Большой русско-испанский словарь]. Madrid: Rubiños-1860.

VARELA, F. AND KUBARTH, H., 1994. *Diccionario fraseológico del español moderno*. Madrid: Gredos. [DFEM]

WERESCHAGIN, E AND KOSTOMAROV W., 1990. *Lengua y cultura: estudios de lengua y geografía en la enseñanza de la lengua rusa como extranjera* [Язык и культура: Лингвострановедение в преподавании русского языка как иностранного]. Moscú: Shkola "Russkiy Jazik".

ZAGORSKAYA, N., 2010. *Gran diccionario español-ruso* [Большой испанско-русский словарь]. Moscú: Drofa. [GDER]

ZHOLOBOVA, A., 2005. *Fraseologicheskie edinitsi biblejskogo proishoghdenija v anglijskom, ispanskom i russkom jazikah* [Фразеологические единицы библейского происхождения в английском, испанском и русском языках]. Tesis doctoral. Kazán: Universidad de Kazán.

A TIROS Y *A BALAZOS*: ANÁLISIS CONSTRUCCIONAL

Belén López Meirama
Universidade de Santiago de Compostela
belen.meirama@usc.es

Resumen

Pretendemos analizar dos realizaciones de la construcción fraseológica [*a* + $S_{plural/acción\ violenta}$] estableciendo, a través de datos extraídos de varios corpus del español, su grado de fijación interna y externa. Nuestra intención es trascender lo estrictamente morfosintáctico para centrarnos en aspectos semánticos y, en la medida de lo posible, también pragmáticos.

1. INTRODUCCIÓN

Esta comunicación se enmarca en un proyecto de investigación[80] cuyo propósito es integrar en la fraseología los presupuestos de la Gramática de Construcciones (Goldberg 1995, Taylor 2014) y la teoría de las «combinaciones usuales de palabras» (CUP) a través del análisis de la estructura [preposición + sustantivo] del alemán y el español. Las CUP (Steyer 2013) o *constructional idioms* (Taylor 2014) son patrones construccionales recurrentes que se componen de ciertos constituyentes fijos y otros que, aunque son casillas vacías (*free slots*), están sometidos a ciertas restricciones en su combinatoria semántica. La metodología empleada en el proyecto para analizar las CUP se basa en la convicción de que la frecuencia de uso de las construcciones propicia la creación de patrones o modelos combinatorios, de forma que los candidatos para ser considerados CUP solo pueden extraerse de corpus sometidos a análisis estadísticos.

En este contexto, para seleccionar nuestro objeto de estudio hemos partido de la construcción [*a* + $S_{plural/acción\ violenta}$], conocida en la fraseología del español y de gran productividad: *a golpes*, *a palos*, *a empujones*, *a pisotones*, *a cuchilladas*, *a pedradas, a mordiscos*, *a balonazos*, *a codazos*, *a guantazos*, *a pelotazos*, *a tortazos*, etc.

Tras ejecutar la orden de búsqueda 'a + sustantivo' en el *Corpus del español* (DAVIES) para determinar las combinaciones usuales más frecuentes con este esquema,

[80] *Combinaciones fraseológicas del alemán de estructura [PREP. + SUST.]: patrones sintagmáticos, descripción lexicográfica y correspondencias en español*, proyecto impulsado por el equipo de investigación FRASESPAL y financiado por el MEC (FFI2013-45769-P).

comprobamos que las secuencias *a tiros* y *a balazos* se encuentran entre las de mayor frecuencia absoluta: presentan 70 y 52 ocurrencias, respectivamente, en textos del siglo XX. Por ello las seleccionamos, con el objetivo de llevar a cabo un análisis de corpus que permita describir sus características más relevantes. En particular, analizaremos cuál es su grado de fijación siguiendo las pautas del proyecto de investigación mencionado e intentaremos dibujar sus contextos de uso prototípicos.

Para ello, contamos con los datos de tres corpus del español, DAVIES, CREA y CORPES. Nos serviremos de los tres para abastecernos de ejemplos reales y para realizar búsquedas, pero nos limitaremos al CORPES para obtener datos estadísticos. Con el objetivo de acotar el número de ocurrencias y así simplificar los análisis, hemos seleccionado dos subcorpus, eligiendo aquellos con frecuencias (absolutas y normalizadas) más altas en alguna de las dos realizaciones. Tales subcorpus son los correspondientes al español de España y de México:

País	Frecuencia absoluta		Frecuencia normalizada	
	a tiros	*a balazos*	*a tiros*	*a balazos*
España	247	44	3,38	0,60
México	64	104	2,42	3,94

Tabla 1. Ocurrencias en CORPES[81]

2. FIJACIÓN INTERNA

Solo algunas realizaciones de la construcción [*a* + $S_{plural/acción\ violenta}$] se recogen como locuciones en repertorios o manuales de fraseología del español, como Corpas Pastor 1996 (*a patadas*), Ruiz Gurillo 1998 (*a gritos*), Martínez López 1999 (*a golpes, a gritos, a tortas, a patadas*), Seco *et al.* 2004 (*a escobazos, a puntapiés, a zapatazos*, etc.), Penadés Martínez 2005 (*a escobazos, a gritos, a patadas, a puntapiés*) o García-Page Sánchez 2007 (*a bofetadas, a empellones, a golpes, a puñetazos*).

La mayoría de ellas conoce empleos metaforizados (ej.: 'con desconsideración': *Me tratas a patadas*), pero no así *a tiros* ni *a balazos*, que presentan menor idiomaticidad porque mantienen el significado del sustantivo. En lo que sigue, comprobaremos si este bajo nivel de idiomaticidad repercute en el grado de fijación.

2.1. Valor de la preposición *a*

Desde el punto de vista semántico, la fijación interna de la CUP se relaciona con el grado de desemantización –y de gramaticalización– de la preposición: cuanto mayor sea este, mayor será el grado de cohesión interna de la construcción.

Como sabemos, *a* es, junto con *de*, la más desemantizada de las preposiciones, pues conoce usos estrictamente gramaticales, de modo que su significado básico es abstracto. Para Company Company y Flores Dávila (2014) tal significado es el de *locatividad directiva télica* hacia una *meta*; consideran que este «significado abstracto y esquemático, al entrar en diferentes construcciones y contextos, puede reelaborarse y adquirir distintos matices de

[81] Como se verá en la tabla 2, eliminaremos algunos casos de *a tiros*, por estar repetidos o porque no se corresponden con la CUP.

sentido» (2014: 1317), entre los cuales se halla el modal. Por tanto, cuando codifica metas modales, como en *a pie*, *a cuadros* o *a golpes* (2014: 1318-19), el grado de abstracción de *a* es incluso mayor que el que tiene en su significado básico. Si ello es así, parece factible inferir que el grado de fijación interna de *a tiros* y *a balazos* será alto. A continuación comprobaremos que la aplicación de otros parámetros para la medición de la fijación interna apunta también en esta dirección.

2.2. Fijación del sustantivo

Puede afirmarse que el sustantivo está fijado en el número plural: solo hemos detectado dos ejemplos de *a balazo* en el CORPES, ambos del área caribeña (*Resultó muerta* a balazo *en su residencia*); por otra parte, *a tiro* es una locución recogida como tal en los diccionarios, así como *a tiro de* y *a tiro hecho*.

Cabría considerar, como sugiere García Page (2007: 121), que la combinación [*a* + $S_{singular}$ + *limpio*] es una variante con el sustantivo en singular, en la cual el adjetivo «resalta el que la acción se lleve a cabo con contundencia y sin miramientos» (NGLE 2009: 2386). Sin embargo, hay restricciones en la selección de los sustantivos y, según los contextos, la construcción aporta un valor peculiar. P.e., las secuencias siguientes, extraídas del CORPES, carecen de correlato plural y además contienen un rasgo de significado específico ('únicamente con el objeto denotado por el sustantivo'): *caminar a pie limpio / a pata limpia, fabricar algo a mano limpia, tocar una canción a guitara limpia, trabajar a lomo limpio, abrir picadas a machete limpio*, etc. Parece, por tanto, que estamos ante una CUP diferente de la que analizamos, si bien esta cuestión merecería un análisis pormenorizado.

2.3. Grado de cohesión entre preposición y sustantivo

Tras aplicar el test de inserción a las dos realizaciones a través de búsquedas por proximidad en el CORPES y en el CREA (en los dos corpus, con distancia no superior a 2), es posible concluir que en ambos casos la tendencia a la expansión es baja.

2.3.1. La consulta de los dos corpus demuestra que la inserción por determinación se limita, prácticamente, a algunos ejemplos del español de Argentina[82], corroborando así una observación de la NGLE, donde, al describir las locuciones adverbiales que corresponden al esquema «*a* + sustantivo en plural», se comenta que «las variantes de estas construcciones con artículo determinado (*a los golpes, a los gritos, a los empujones, a los saltos, a las patadas*) son características del español rioplatense» (2009: 2385).

Los datos de los dos corpus demuestran que la presencia de la variante con artículo solo es relevante en relación con *a tiros* y en Argentina, donde alcanza aproximadamente el 30% de las ocurrencias. Aunque el porcentaje merecería un análisis más detallado, en todo caso creemos poder afirmar que, salvando esta excepción, la inserción de un determinante es insólita en la construcción, lo cual es un indicio de fuerte fijación.

2.3.2. También es muy inusual la inserción por modificación. Tras rastrear en los tres corpus, únicamente hemos encontrado un ejemplo de modificación por adjetivo:

1. [Los narcos] Arreglan todo *a puros balazos*, nada más.

[82] No hemos localizado ningún ejemplo de *a los balazos* en textos no argentinos, y solo hemos encontrado uno de *a los tiros* en otros tres países.

Respecto a este ejemplo, conviene recordar que con el significado 'mero, solo, no acompañado de otra cosa' (DRAE: *s.v. puro*), el adjetivo *puro* se asimila en posición prenominal a los determinativos (cfr. *Es pura agua* / *Es agua pura*), de modo que no estamos, en realidad, ante un modificador, sino ante un elemento intensificador, especialmente productivo en América, que se combina tanto con sustantivos plurales (*a puros culatazos, plomazos*) como con singulares (*a puro golpe, balinazo*), a veces con un alto grado de idiomaticidad (*a puro pulmón, a puro pulso, a puro huevo*).

Hemos encontrado ejemplos, bastante escasos, en los que el sustantivo está seguido de una FPREP, que en la mayoría de los casos sirve para identificar el arma empleada: *a tiros de pistola* / *de fusil* / *de escopeta* / *de revólver, a balazos de fusil,* aunque también hay alguno en el que se especifica la parte del cuerpo que recibe el impacto: *Me los mataron a tiros en la cabeza, Me matarían a tiros por el trasero*. En todos estos ejemplos el adjunto recibe una interpretación clasificativa o relacional, es decir, establece una restricción en la denotación del sustantivo, a lo cual no es ajeno el hecho de que este carezca de determinante (véase NGLE 2009: 861). No hemos detectado, sin embargo, ejemplos de modificación cualitativa del tipo de *a bestiales golpes de silla*, *a empujones violentos, a dentelladas secas y calientes*.

2.3.3. Algo más abundantes son los casos de expansión por coordinación: algunos ejemplos hallados en los corpus son *a tiros y botellazos, a tiros y pedradas, a tiros y puñaladas, a tiros y puñetazos, a hachazos y tiros, a balazos y granadazos, a balazos y puñaladas* y *a puñaladas, balazos y mordiscos*. La coordinación contribuye a a aumentar el énfasis, la intensificación que caracteriza pragmáticamente estas combinaciones.

2.3.4. Para ofrecer datos cuantitativos precisos acerca de la (escasa) significación que tienen todas estas modificaciones, hemos contabilizado los tres tipos de expansión en los subcorpus del CORPES correspondientes al español de España y México. El resultado puede verse en la tabla 2, donde se pone en evidencia lo inusual de la expansión. Concluimos, entonces, que la fijación interna de *a tiros* y *a balazos* es elevada.

	Subcorpus	Nº total de ocurrencias	Expansión por…		
			determinac.	modificac.	coordinac.
a tiros	España	241	-	2	4
	México	63	-	1	1
a balazos	España	44	-	-	1
	México	104	-	-	1

Tabla 2. Casos de expansión en relación con el número total de ocurrencias

3. FIJACIÓN EXTERNA

En este apartado vamos a indagar cuáles son las preferencias combinatorias de *a tiros* y *a balazos*, para comprobar si es posible establecer algún perfil combinatorio prototípico, tanto desde el punto de vista sintáctico como desde el punto de vista léxico.

3.1. Valores sintácticos

En general, las dos unidades objeto de análisis suelen adjuntarse a verbos (*matar a tiros, asesinar a balazos*), aunque también hemos detectado contextos en los que funcionan como modificadores de participios (*abatido a tiros, acribillado a balazos*) y de sustantivos eventivos (*muerte a tiros, enfrentamiento a balazos*). Si bien el tamaño de las muestras no permite extraer conclusiones firmes, sí puede observarse que la modificación (nominal y participial) es bastante más que anecdótica; concretamente, *a tiros* y *a balazos* se sitúan funcionalmente a nivel frástico en el 28% de los casos, como modificadores de sustantivo o de participio.

Por otra parte, no siempre que son adyacentes de un verbo su función es la de circunstante: hay contextos en los que este aporta básicamente valor aspectual, sea ingresivo (*empiezan a tiros*), terminativo (*acaban a balazos*) o durativo (*andan a balazos*), de modo que el elemento que aporta la carga semántica relativa a la acción realizada es el sustantivo (*tiros* o *balazos*); en estos casos, el vínculo sintáctico entre la CUP y el verbo es muy estrecho (cfr. *Ellos terminaron a tiros ≠ Ellos terminaron*).

El vínculo también se observa con los verbos de combinación prototípica, como *matar* o *morir*; por ejemplo, en las siguientes secuencias comprobamos que se coordinan –y por tanto se asimilan sintácticamente– a un complemento de régimen (2a) y a un predicativo (2b):

2. a. Muchos morían de malaria, otros *a balazos*, otros de hambre y soledad.

b. Cuando hablan de morir ¿qué más da si es *a tiros* o atravesados por la espada?

3.2. Combinatoria sintagmática léxica

Dado que los sustantivos que se construyen con *a tiros* y *a balazos* son eventivos y comparten lexema con los verbos, en este apartado y el siguiente nos limitaremos a examinar la combinatoria verbal, incluyendo en la misma las formas de participio.

El significado modal de la construcción, unido al de los sustantivos, limita mucho la combinabilidad de ambas realizaciones: como cabría esperar, los verbos que más frecuentemente se construyen con ellas remiten al significado general de 'quitar la vida': *abatir, acribillar, asesinar, dar muerte, ejecutar, matar*, *rematar*, *ultimar*, etc., a los que cabe añadir *morir*, que en estos contextos equivale a 'ser asesinado'.

Asimismo, hay bastantes verbos que se usan metafóricamente con este mismo sentido, pero aportando una carga expresiva hiperbólica; por lo general, son verbos que muestran muy plásticamente el destrozo físico: *coser, deshacer, desmenuzar, despedazar, destrozar, enhebrar, freír, mechar, reventar, siluetear*, etc.

También se contabilizan verbos que significan, más generalmente, agresión: *agredir, asaltar, atacar, derribar, herir, lesionar*, etc. y enfrentamiento: *agarrarse, ajustar (las) cuentas, batirse, defender(se), enfrentarse, enzarzarse, liarse, repeler*, etc.

En ciertos casos la acción violenta se asocia con un desplazamiento (tanto del que la lleva a cabo como del que la sufre), de modo que también encontramos verbos que denotan el movimiento con el que esta se inicia: *echar, entrar, irrumpir, llegar, recibir, salir*, etc.

Algunos de los verbos tienen sentido ingresivo, como *enzarzarse, irrumpir* o *liarse*, a los que cabría sumar *emprenderla* y *empezar*; otros, lo tienen terminativo: *acabar, detener, morir, resolver*. En relación con esta cuestión cabe destacar que, en general, los elementos modificados por *a tiros* y *a balazos* suelen referirse a actividades temporalmente limitadas y casi siempre breves, es decir, suelen ser aspectualmente puntuales (3a), raramente durativos (3b):

> 3. a. Se dice […] que *dio muerte a balazos* a un par de malandrines que le detuvieron el paso justo afuera del banco en el momento de ir a depositar.
>
> b. Mandad los apaches, estos cabrones nos *están friendo a tiros*.

El valor aspectual se refuerza con la selección de tiempos verbales, que por lo general son perfectivos: destaca el uso del pretérito perfecto, simple o compuesto, mientras que el presente, aunque no es inusual, se limita a titulares de prensa o a contextos dialógicos, en los que suele adquirir valor prospectivo empleado con sentido de amenaza (*Para o te mecho a tiros*).

Probablemente esta preferencia no sea más que una consecuencia del significado léxico de los sustantivos, ambos designación de una acción fuerte, inesperada y momentánea (*vid.* Monge 1972: 243).

Por otra parte, hemos detectado una presencia muy significativa de la perífrasis pasiva, en contextos en los que, por tanto, la acción se focaliza en el paciente. Aunque podría pensarse que esta presencia es consecuencia de la combinatoria verbal, ya que verbos como *asesinar*, *atacar*, *ejecutar*, etc., se construyen con frecuencia en voz pasiva, lo cierto es que en combinación con *a tiros* o *a balazos* la frecuencia es sensiblemente mayor. Solo como ejemplo: si bien la consulta de la BDS revela que los verbos *asesinar* y *acribillar* presentan unos porcentajes elevados de esquemas pasivos (35% y 25%, respectivamente), observamos que estos se encuentran muy lejos de los que arrojan sus combinaciones con *a tiros* y *a balazos* en la parte española del CORPES, que se sitúan, de media, en torno al 75%.

A la luz de todos estos datos podemos concluir que, en general, *a tiros* y *a balazos* son adyacentes de formas verbales que designan 'ataque, agresión', en estructuras en las que se destaca el carácter brusco, repentino, momentáneo de la acción y que se focalizan en el paciente con una frecuencia muy significativa.

3.3. Preferencia combinatoria léxico-verbal

En general, no se detectan grandes diferencias en la combinatoria de las dos realizaciones ni en la de los dos subcorpus (salvando algunas formas dialectales, como *agarrarse* o *ultimar*). Aun teniendo en cuenta el tamaño relativamente pequeño de las muestras, las coincidencias son altas: entre 5 y 7 casos en los 10 verbos más frecuentes, tanto si contrastamos los dos subcorpus como si contrastamos las dos realizaciones. Además, en unos y otras destacan los verbos *asesinar, matar, morir y acribillar*, si bien este último presenta un porcentaje especialmente alto (cercano al 50%) en el caso de la realización *a balazos* de la parte española.

Acribillar a balazos, de hecho, es claramente una coaparición preferida –o colocación– en el español de España: si bien no se trata de una combinación exclusiva, la consulta de la

parte española del CORPES nos permite constatar que otras combinaciones con *acribillar*, aunque posibles, son esporádicas (*a alfilerazos, a picotazos, a perdigonazos*).

Asimismo, entre los verbos que se usan metafóricamente en una descripción hiperbólica del destrozo físico, destacan *coser* y *freír*: el primero de ellos lo hemos encontrado en los dos subcorpus y con las dos realizaciones, aunque debe señalarse que, con este sentido de 'producir heridas', *coser* se combina preferentemente con sustantivos que denotan lesiones realizadas con armas cortantes y/o punzantes (*coser a navajazos, cuchilladas, machetazos, bayonetazos, puntillazos,* etc.), en particular con el sustantivo *puñaladas*. El segundo, sin embargo, solo lo hemos detectado en la parte española del corpus y combinado con *a tiros*; además, apenas hemos encontrado un ejemplo de *freír a balazos* en Chile y otro de *freír a descargas* en España, así que parece que *freír a tiros* puede considerarse, como *acribillar a balazos*, una coaparición preferida en el español de España.

Por otra parte, observamos que en el subcorpus español la combinación *liarse a tiros* está entre las más frecuentes, aunque también son habituales otras muchas coocurrencias, como *liarse a puñetazos*, *a mamporros*, *a bofetadas, a golpes, a patadas*, *a hostias*, etc. Además, *liarse a* se emplea también con infinitivos, en general para denotar actividades violentas (*liarse a dar / pegar tiros, pegar a alguien, dar golpes*, etc.) e, incluso más frecuentemente, actos de habla (*liarse a hablar, discutir, recitar, chatear, contar chistes, hacer preguntas*, etc.). Esta preferencia en la combinación verbal es una de las razones por las que «*liarse a* + infinitivo» no suele considerarse una perífrasis verbal, sino más bien un esquema fraseológico semiproductivo (*vid.* NGLE 2009: 2125). Sea como fuere, parece sensato concluir que estamos más bien ante un uso específico del verbo *liar*, en conjugación pronominal y con rección de la preposición *a,* y no ante la combinación *liarse* con la CUP [*a* + $S_{plural/acción\ violenta}$].

Un caso similar al de *liar* es el del verbo *emprender*, al que se adjunta un clítico de valor no referencial sin referente, con el cual se construye el esquema [*emprenderla a* + S_{pl} *con/contra* + S/Pron]: *La emprendió a tiros contra ellos*, *La emprendieron a golpes con todos, La emprendieron a naranjazos con la fachada*…

3.4. Esquematicidad en el cotexto

Finalmente, comprobaremos si existen elementos que se combinen con cierta frecuencia con alguna de las dos realizaciones.

En primer lugar, hemos detectado algún ejemplo en el que la construcción está introducida por un elemento focalizador con valor aproximativo, aunque sin relevancia cuantitativa:

4. a. Los echó de la casa *casi a tiros*.

b. En otras ocasiones […] hubo que echarlas [las liebres] *poco menos que a tiros*.

En segundo lugar, localizamos casos de la expresión *ni a tiros*, que la NGLE incluye en el extenso grupo de locuciones negativas que se forman con la conjunción *ni* (NGLE 2009: 3714) y que se recoge en los diccionarios como locución adverbial coloquial (DRAE: 'Ni aun con la mayor violencia, de ningún modo, en absoluto'). No cabe duda de que en este caso el grado de idiomaticidad es elevado; de hecho, apenas hemos encontrado contextos

en los que se exprese violencia. Sin embargo, aunque es indudable la productividad de *a tiros*, también se pueden encontrar ejemplos con *a balazos* y con otros sustantivos (*Homero no baila ni a balazos; No la despertaríamos ni a cañonazos; No se doblan ni a martillazos*), así que la fórmula merecería un análisis pormenorizado.

4. CONCLUSIONES. DESCRIPCIÓN DE LA CONSTRUCCIÓN

Tras todo lo visto, podemos concluir que las dos realizaciones examinadas de la CUP [*a* + $S_{plural/acción\ violenta}$] apenas presentan diferencias entre sí: ambas tienen un alto grado de fijación interna, si bien esta no es absoluta; su combinatoria léxica es muy similar, aunque detectamos algunas preferencias en el español de España (*acribillar a balazos, freír a tiros*); sintácticamente, no se limitan a funcionar como circunstantes, sino que con cierta frecuencia presentan un estrecho vínculo sintáctico con el verbo.

Por otra parte, las características que ambas exhiben apuntan a un valor pragmático apenas aludido en los párrafos precedentes: junto al significado denotativo de la CUP, claramente modal (algo así como 'realizando la acción denotada por el S'), es posible identificar un valor pragmático de intensificación, evidenciado por el empleo del plural en el sustantivo; la selección de verbos de significado hiperbólico (*aniquilar, destrozar, reventar*), a veces metafórico (*desmigar, desmenuzar, enhebrar*); la inserción de algún elemento focalizador (*a puros balazos*); etc.

Asimismo, *a tiros* y *a balazos* son formas más bien propias del registro coloquial, donde se han generado otras construcciones, como *ni a tiros*.

Finalmente, el análisis en contexto de estas dos realizaciones sugiere que en el español de España *a tiros*, de uso mucho más frecuente, tiene un mayor grado de fijación, mientras que *a balazos* se usa en contextos de mayor expresividad; sin embargo, por limitaciones obvias no hemos podido desarrollar esta idea. Tampoco hemos podido abordar otros muchos aspectos, que demuestran que nos hallamos en un terreno apenas desbrozado y a nuestro juicio de gran interés, como las combinaciones que nos hemos limitado a mostrar pero que deberían analizarse con mayor detenimiento: *a balazo limpio*, *a los tiros, a puros balazos, ni a tiros*… Por otra parte, y ya desde una perspectiva más amplia, cabría analizar construcciones con el mismo esquema formal pero con otros significados (*a trozos, a trompicones, a carretadas*…), así como la construcción [*de* + NUM + S_{pl}], de significado próximo al que nos ha ocupado: *Te mataré* de cinco balazos *como a un perro, en plena calle*.

Bibliografía

MALDONADO GONZÁLEZ, C., (dir.) 1996. *Clave. Diccionario de uso del español actual*. Madrid: Ediciones SM.

COMPANY COMPANY, C. & FLORES DÁVILA, R., 2014. La preposición *a*. In C. Company Company, ed. 2014. *Sintaxis histórica de la lengua española*, Tercera parte, *Preposiciones, adverbios y conjunciones. Relaciones interoracionales*, Volumen 2. México: Fondo de Cultura Económica. Cap.11.

CORPAS PASTOR, G., 1996. *Manual de fraseología española*. Madrid: Gredos.

DRAE = REAL ACADEMIA ESPAÑOLA, 2014[23]. *Diccionario de la lengua española*. Madrid: Espasa.

GARCÍA-PAGE SÁNCHEZ, M., 2007. Esquemas sintácticos de formación de locuciones adverbiales. *Moenia*, 13, pp. 121-144.

GARCIA-PAGE SANCHEZ, M., 2008. *Introducción a la fraseología española: estudio de las locuciones*. Barcelona: Arthropos.

GOLDBERG, A.E., 1995. *Constructions. A Construction Grammar Approach to Argument Structure.* Chicago: University of Chicago Press.

MONGE, F., 1972. Sufijos españoles para la designación de 'golpe'. In: *Homenaje a Francisco Yndurain*. Zaragoza: Universidad de Zaragoza. pp. 229-247.

NGLE = REAL ACADEMIA ESPAÑOLA Y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA (2009). *Nueva gramática de la lengua española.* Madrid: Espasa.

PENADÉS MARTÍNEZ, I., 2005. *Diccionario de locuciones adverbiales para la enseñanza del español*. Madrid: Arco Libros

PENADÉS MARTÍNEZ, I., 2012. *Gramática y semántica de las locuciones*. Alcalá de Henares: Universidad de Alcalá.

RUIZ GURILLO, L., 2001. *Las locuciones en español actual*. Madrid: Arco/Libros.

SECO, M. ANDRÉS, O. & RAMOS, G., 2004. *Diccionario fraseológico documentado del español actual. Locuciones y modismos españoles*. Madrid: Aguilar.

STEYER, K., 2013. *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. (Studien zur Deutschen Sprache 60). Tübingen: Narr.

TAYLOR, J. R., 2014. Cognitive linguistics (for Routledge Handbook of Linguistics), draft. Available at: https://www.academia.edu/7179973/Taylor_2014_Cognitive_linguistics_for_Routledge_Handbook_of_Linguistics .

Corpus consultados

BDS = *Base de datos sintácticos del español actual.* USC. http://www.bds.usc.es/

CORPES = REAL ACADEMIA ESPAÑOLA: Banco de datos (CORPES XXI) [en línea]. *Corpus del Español del Siglo XXI (CORPES).* http://www.rae.es

CREA = REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual.* http://www.rae.es

DAVIES = DAVIES, MARK. (2002-) *Corpus del Español: 100 million words, 1200s-1900s*. http://www.corpusdelespanol.org

AN EXPLORATION OF THE PHRASEOLOGY OF A LARGE CORPUS OF ACADEMIC ENGLISH: A NEW MORE TEACHABLE TAXONOMY OF MULTIWORD EXPRESSIONS

John Anthony McKenny

Abstract

This paper grew out of a corpus project which set out to compile the British University in Dubai (BUiD) corpus of written academic English. This corpus is the only sizeable corpus of tertiary English writing from the Middle East. Masters dissertations submitted by BUiD students in the Education, Business and Engineering Faculties since 2004 were available in machine-readable form. These were collected as PDFs and converted into text only (ASCII) format. Arabic is the L1 of the large majority of the students; all students on submission agreed to their work being used for research. TEI metadata (headers, section divisions) were introduced to make Abstracts, Introductions, Literature Reviews, Methods, Results, Discussion, Conclusion and References sections retrievable as more genetically focused sub-genres of the macro- genre, *Dissertation* (Paré et al., 2015). It is intended to carry out genre analysis on sub-corpora containing all the examples of the major sections of the dissertations (e.g. introductions, conclusions) possibly as optional student-led project work.

Makeup of the Buidcorp corpus:

91	M.Ed.	2.3 million academic words (TESOL and LEADERSHIP)
206 M.Sc. DISSERTATIONS: 5.3 million words (BUSINESS)		
140	M.Sc.	2.4 million words (ENGINEERING and IT)
Approximate total word count 10 million words		

1. INTRODUCCIÓN

The most useful application of our corpus seemed to be the production of a resource for writers in the same position and context as our corpus authors. Writers of English for Academic Purposes need to develop phraseological competence, formerly referred to as collocational competence (Howarth (1998). According to Bolinger (1975:7) there is a "greater degree of unfreedom in every syntactic combination that is not random" than

allowed for by generative linguists. There seems to be a level above the lexical, where the choice of one word constrains the words which might co-occur with it. These constraints seem to be arbitrary and conventionalised. One of the earliest and most influential accounts of this 'unfreedom' is found in the work of Pawley and Syder (1983) in their book chapter, 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency'. They call into question the Chomskyan notion of creativity in the generation of grammatical sentences. For Pawley and Syder, only a small subset of the innumerable or indefinitely large number of sentences which would be generated by a Chomskyan grammar are nativelike in form - in the sense of being acceptable to native informants as ordinary, natural forms of expression, in contrast to expressions that are grammatical but are judged to be 'unidiomatic', 'odd' or 'foreignisms.' How can EAP students be helped to come to terms with this phraseological dimension of language?

The problem can be viewed from a different perspective. According to Fox (1998):

> It is certainly true that there are times when you do want your audience to be impressed by your use of language...But most of the time you don't want that. You just want to get your message across. And you do that by being unremarkable in your language, by being conventional and predictable, and - dare I say it - boring! (Fox 1998:28)

As Fox points out here, the greatest challenge faced by novice writers of academic prose is the production of idiomatic English with a phraseology which expresses the writer's meaning well without drawing attention to itself through unnaturalness or strangeness. This involves deploying sufficient delexicalised words to achieve the right amount of redundancy. Many applied linguists have addressed this problem and two notable early attempts to come to the aid of the EAP writer was a proposal by Rosamund Moon in an unpublished talk (1996) and a talk by Sardinha (1996) where the two applied linguists proposed independently the use of software which could check the phraseology of the writing of students as they write and advising students when a sequence of words was unusual or in some way strange. The students' words would be run through a large corpus and the software could suggest alternative word sequences. This would come in the form of a yet-to be-invented add-in tool for use within a word processor. Thus a huge amount of learning now incumbent on a student of a language would be rendered unnecessary and the time currently devoted to it could be used to learn other areas of the language. While we wait for AI research to give us such a tool, writers can use corpora to test their phraseology. Google is one way writers can test phraseology that they are in doubt about. Frequency of occurrence of expressions on the Internet can be a useful but not infallible guide to idiomaticity. One of the strengths of the Buidcorp corpus is its large size which generates sufficient and manageable amounts of examples of locutions which enables writers to determine which turns of phrase are typical and expected in their field of study, thus addressing the problem referred to by Fox (1998) above.

The Longman Dictionary of Spoken and Written English (LGSWE) marked a milestone in the quest for ways of providing some kind of taxonomy of this protean area of linguistic studies. Biber et. Al. (1999) in their monumental LGWSE started by coining the term *lexical bundle.*

This term was very close in meaning to the established term *Ngram* except that Biber and his co- authors stipulated that to qualify as a lexical bundle the sequence of words would have to occur a certain minimum number of pre-ordained times per million words. Thus lexical bundle is determined purely on statistical grounds. Several systems have been used to remove those lexical bundles which have little psycholinguistic salience and also no pedagogical relevance (e.g. *and of the).* Vlach-Simpsonn and Ellis (2010) introduced quantitative and qualitative methods to do such filtering. Martinez and Schmitt (2012)

introduced 'teachability' as a further selection criterion. Critics see this as introducing too much subjectivity to the selection process, but it is difficult to see how subjectivity can be avoided. As in all corpus linguistics investigation, the data obtained must be appraised in some way by the intuition of the linguist or 'knowing subject'.

An earlier example of a selection system was that of De Cock, Granger, Leech and McEnery (1998:75):

1) the elimination of most combinations of closed-class items that are only phrase or clause fragments; e.g. *in the, and it* and the subjective elimination of those word combinations which are fragmentary in nature; e.g. *are a lot of, don't know if you;*
2) The next phase of the filtering process is an assessment of whether the remaining candidate formulae have the potential to serve any pragmatic or discourse functions;
3) The final stage examines each occurrence of a potential formula in its co-text.

McKenny (2011) attempted to create a taxonomy of the lexical bundles contained in a learner corpus using the three language macrofunctions of Halliday and Hasan (1989) viz. the ideational, the interpersonal and the textual. Recently there has become available a much more apposite taxonomy in the work of Hyland (2008) and Salazar (2010). The three main categories in this model fit much better with the main functions of academic prose. The textual function is retained. The interpersonal function becomes the procedural function, referring to hedging, boosting, all the other ways that the writer approximates to the reader in expressing epistemic stance, certainty, tentativeness. The ideational function is replaced by the Research function which covers all bundles which are used to refer to the research being done in the dissertation or which appears in the literature review to support the dissertation's research agenda. With this new taxonomy, it is hoped that the prefabricated language of the corpus can be better categorized and will more effectively support the creation of lists of idiomatic expressions with more psychological reality and which are more teachable in academic literacy classes. Salazar (2010) provides such a taxonomy of bundles for a corpus of health sciences.

This refining procedure could be applied in the creation of lists of lexical bundles extracted from the several sub-corpora of the Buidcorp corpus. These could be arranged according to genre and function as has been very skilfully done by John Morley (2015) for his very useful and generous website available to all. Morley assembles a list of useful phrases which he recommends writers uses as models with which to construct their texts, making lexical substitutions and grammatical adjustments where necessary. Morley's comment that his site is being used more by native writers is most interesting.

2. DATA DRIVEN LEARNING

Tim Johns (1994) in his influential work on Data Driven Learning described his students as becoming „language detectives' as they manipulated and investigated the language data discovering the rules for themselves. This kind of learning is different from traditional approaches to language teaching in that it stresses the incorporation of real data into the language classroom. One of the traditional approaches to language teaching is the 'Present-Practice- Produce' (PPP) method by which students practice a previously

presented grammatical or lexical construction and are then able to produce a sentence or text which exhibits the newly learned feature, paying special attention to the notion of 'fluency'. This approach is mainly 'product based' in that it presents specific aspects of the language to the students. Data-driven learning, on the other hand, uses corpora and concordance software to teach language. It is 'process based' in nature which means that it encourages students to discover certain aspects of language themselves and experiment with the language. It assumes that grammar is a relatively flexible system and not a set of static rules.

Advantages of the product based approach are obvious. It gives students a sense of direction and structure while learning many aspects of a language. However, a product based approach shows only fragments of language or grammatical rules, which are based to a certain extent on the intuition of the textbook author. Recent corpus research shows that these rules are often inaccurate. In the 'process based' approach, students are encouraged to actively think about texts and to recognize language patterns.

Carter and McCarthy (1995) suggest supplementing the PPP approach with an III approach (Illustration, Interaction, Induction). Illustration takes place via an examination of real data, while interaction with the text draws students' attention to certain aspects of the data, such as interpersonal grammar. The induction stage signifies the students' ability to draw conclusions about the interpersonal functions of lexico-grammatical constructions.

More research is necessary to determine how students learn from corpus data. For now it is clear that concordancing in the classroom is an extremely powerful hypothesis testing device which allows controlled speculation, makes hidden structures visible, enhances at the same time imagination and checks it by inductivity, thus making higher degrees of objectivity possible. Other corpus data, for example word and Ngram frequency lists can make students aware of what are the usual and preferred ways of saying things and what are the less usual ways.

We can see two main English for Academic Purposes (henceforth EAP) scenarios: the EAP practitioner relays or mediates corpus findings to writers who need to join a discourse community or who need to transfer/translate their L1 skills and knowledge to handle L2 discourse conventions. The second kind of EAP practice considered in this paper is the initiation of soon-to-be autonomous learners in the use of corpus methodology to become researchers of the language actually occurring in their discourse community. Working from the level of Word upwards, a student can quickly take stock of a dissertation or collection of dissertations.

Although word frequency should not be overemphasized, it provides a useful way into a text or a genre. Wordsmith Tools is particularly useful for looking into phraseology. For example all the 3-5 word clusters (lexical bundles) in a corpus can be generated to reveal the phraseological patters found in inside the corpus.

3. NOTE ON THE TWO MAIN TOOLS USED

In this paper two different tools are described that *language detectives* can use in exploring corpora: the already mentioned Wmatrix3 (Rayson, 2009) and Wordsmith Tools 6 (Scott, 2012) Wmatrix3 automates a good deal of the corpus analysis that needs to be done. The user uploads corpora to the University Centre for Computer Corpus Research Language (UCREL) and works mainly online. Wmatrix3 tags for Part of Speech (POS) and for semantic domain (USAS). It compares a corpus with a control corpus and produces

frequency lists of words, parts of speech, and lists of multiword expressions. At all stages, concordances and lists can be produced. Wmatrix3 presents almost instantly to the investigator a great deal of information about the language of the corpus at many levels, giving an overview of the whole corpus. Such lavish and all-encompassing data might support the enquiries of a student who prefers a top-down approach.

Wordsmith Tools lends itself to working with a bottom-up approach. The corpus linguist can start with Wordlists (lists providing words in descending order of frequency) and move on to Keywords, concordances, clusters (Ngrams), and concgrams. It has many more functions than can be described here. Sometimes Wordsmith compliments Wmatrix3. If a corpus contains many files, these files have to be joined using Wordsmith's file utilities before they can be uploaded as one file to Wmatrix3. Overall, Wmatrix3 suits a researcher who prefers a top-down approach to investigating language while Wordsmith Tools suits a more bottom-up approach. Students can be initiated into the workings of both tools in a relatively short time by a teacher in one-to-one or small group sessions. There are various tutorials within the system for users to deepen their knowledge of the tools' potentialities. A lifelong licence for each costs approximately 70 € to buy.

4. EXAMINING THE EDUCATION SUB-CORPUS

4.1. Keywords: the aboutness of texts

One approach to our corpus is to find the keywords of the texts in our corpus. Keywords are words which are significantly more frequent in one corpus than another. The concept of Keyness is now the preferred term. Keywords can be generated either by Wordsmith Tools or more simply, using Wmatrix3. The concept of Keyness is now the preferred term. The idea is to identify words specific to a particular kind of language use. The formula calculates the proportional use of a word in the specialized corpus divided by the proportional use of the word in a general corpus. The identification of Keywords can indicate what a corpus (or text) is about (i.e. its aboutness), (Cheng 2012).

Keywords have been championed and explained in many publications by Mike Scott, designer of Wordsmith Tools, and his associates (Scott 2012). Tables 1 lists the 50 top keywords in Education sub-corpus. If this Keyword table is examined, certain characteristics of the Education corpus become apparent.

	Word/phrase	Freq Edcorp	%	Freq Control corpus	%	+/-	Log Likelihood
1	students	441	1.52	305	0.03	+	2096.54
2	technology	339	1.17	94	0.01	+	1925.73
3	teachers	267	0.92	71	0.01	+	1525.91
4	writing	257	0.89	115	0.01	+	1346.96
5	laptop	199	0.69	22	0	+	1251.14
6	initiative	164	0.57	43	0	+	939.14
7	computers	130	0.45	11	0	+	833.46
8	laptops	119	0.41	6	0	+	785.24
9	learning	159	0.55	106	0.01	+	762.83
10	United_Arab_Emirates	101	0.35	1	0	+	695.84
11	student	142	0.49	94	0.01	+	682.46

12	science	147	0.51	157	0.02	+	617.55
13	schools	140	0.48	132	0.01	+	611.28
14	to-	82	0.28	0	0	+	574.01
15	one-	82	0.28	1	0	+	563.25
16	classes	94	0.32	26	0	+	534.17
17	study	161	0.56	321	0.03	+	532.65
18	one-to-one	75	0.26	2	0	+	506.59
19	100010	72	0.25	0	0	+	504.01
20	id	84	0.29	32	0.01	+	453.33
21	teaching	86	0.3	66	0.01	+	397.99
22	classroom	58	0.2	18	0.01	+	323.91
23	education	92	0.32	189	0.02	+	300.24
24	curriculum	50	0.17	18	0.01	+	272.52
25	research	108	0.37	360	0.04	+	272.46
26	skills	71	0.25	105	0.01	+	266.07
27	EFL	37	0.13	0	0	+	259.01
28	computer	59	0.2	57	0.01	+	255.73
29	high	92	0.32	291	0.03	+	239.55
30	using	100	0.35	373	0.04	+	234.92
31	teacher	53	0.18	48	0.01	+	234.18
32	use	111	0.38	567	0.06	+	207.31
33	emirati	29	0.1	0	0	+	203.01
34	educational	50	0.17	67	0.01	+	194.4
35	in	836	2.89	15816	1.7	+	190.63
36	marks	27	0.09	0	0	+	189
37	classrooms	31	0.11	5	0	+	188.3
38	IAT	25	0.09	0	0	+	175
39	high_school	26	0.09	1	0	+	173.51
40	implementation	40	0.14	39	0	+	172.89
41	achievement	40	0.14	39	0	+	172.89
42	researcher	34	0.12	19	0	+	170
43	Internet	24	0.08	0	0	+	168
44	program	34	0.12	23	0	+	162.53
45	UAE	23	0.08	0	0	+	161
46	grade	32	0.11	18	0	+	159.77
47	write	45	0.16	80	0.01	+	156.56
48	integrate	26	0.09	5	0	+	154.92
49	professional	43	0.15	84	0.01	+	143.58
50	2009	26	0.09	9	0	+	142.65

Table 1. First 50 keywords from education sub-corpus of BUIDCORP

Certain fairly predictable concepts emerge with high Log Likelihood, meaning their frequency of occurrence is much higher than could be predicted judging from the reference corpus (most notably *student(s), teacher(s), classroom(s), technology, computer(s) laptop(s), writing).* The prominence of these last words points to the major revolution which has occurred in education in recent years in which the uses of technology in education has moved centre stage. It might be surmised that the occurrence of two verbs among the keywords, *write* and *integrate* demonstrate two major foci of the Education dissertations, the teaching of writing and the integration of skills. Such speculation can be corroborated by follow-up concordancing.

Using Wordsmith Tools Index, the 3-5 word lexical bundles were extracted from the Education sub-corpus:

	BUNDLE	FREQ.	%	TEXTS
1	In the uae	1,704	0.07	72
2	in order to	1,510	0.06	89
4	gifted and talented	861	0.04	11
5	as well as	828	0.03	86
7	the use of	724	0.03	82
8	of the students	689	0.03	72
9	with special needs	642	0.03	33
10	of the study	626	0.03	78
11	united arab emirates	614	0.03	70
12	the importance of	607	0.03	82
13	in this study	580	0.02	77
14	on the other	543	0.02	75
15	in the classroom	534	0.02	69
16	the other hand	507	0.02	70
17	on the other hand	505	0.02	70
18	the impact of	501	0.02	61
19	of this study	492	0.02	77
20	the united arab	492	0.02	65
21	in terms of	491	0.02	80
22	ministry of education	489	0.02	64
23	the role of	486	0.02	77
24	the number of	484	0.02	80
26	the united arab emirates	483	0.02	65
27	the gifted and	468	0.02	9
28	the gifted and talented	454	0.02	8
29	in the school	448	0.02	51
30	do you think	444	0.02	79
31	a lot of	444	0.02	41
33	the implementation of	428	0.02	65
34	in the united	418	0.02	85
35	of the teachers	417	0.02	43
36	of the research	411	0.02	71
37	the ministry of	409	0.02	50
38	students with sen	407	0.02	71
39	due to the	407	0.02	63
40	the end of	406	0.02	11
41	that there is	405	0.02	84
42	there is no	397	0.02	75
43	in addition to	393	0.02	82
44	of the school	393	0.02	77
46	the fact that	392	0.02	60
48	be able to	386	0.02	76
49	university in dubai	385	0.02	32
50	teaching and learning	384	0.02	79
53	the needs of	380	0.02	80
54	the purpose of	371	0.02	60
55	of students with	371	0.02	65

Table 2 Most frequent 50 of the 3-5 word bundles from the corpus with original frequencies before filtering.

A cursory glance at this list of lexical bundles reveals many interesting features of the writing in the Buidcorp. The overuse of *a lot of* (444 occurrences in 41 dissertations) shows that students need to be advised that this expression is overly informal and could be replaced with more formal expressions such as *a great deal*. The expression *in terms of* used 491 times in 80 dissertations was probably overused. Students could be invited to examine a partial concordance of this term and decide whether its use was always necessary. Often it is used to sound formal but can be replaced by a much simpler preposition. The EAP teacher can derive a great deal of benefit from lists of lexical bundles as can the autonomous learner or language detective.

The most frequent words in Wordsmith Tools and the keywords in Wmatrix3 can be used by language detectives as starting points for further investigation using the concordancer in each tool. Students can begin to investigate the collocations of the words and the phraseology surrounding them and the functions that words and phrases serve in the texts they are found in.

4.2. An alternative approach to phraseology: the use of Wmatrix3 to extract multiword expressions (MWEs)

Lexical bundles do not exhaust all the possible types of fixed expressions found in natural language. Many of the tropes occur very infrequently (possibly once or never per 10,000,000 words) and so they would never appear in searches based on frequency. This is where the search for multiword expressions using Wmatrix3 can be highly revealing. MWEs are not lexical bundles or formulaic sequences although there is some overlap. Wmatrix3 searches out MWEs using a two-pronged approach. First it has a large lexicon of sayings, idioms, clichés, commonplaces and other fixed expressions for which it searches in the corpus. This is further supplemented by a set of matrices and formulae.

Thus Wmatrix3 provides an automated way of extracting MWEs. Such a tool could suit a certain kind of student and raise their awareness of this feature of language. Other student might prefer readymade models as those provided by Morley's website (2014) and others might prefer to engage autonomously with the data using say. Wordsmith Tools (to extract bundles or clusters as Scott 2012 calls them).

4.3. The control corpus

Lexical Within Wmatrix3 a number of corpora are provided to choose from or the corpus analyst can create one. The BAWE corpus was obtained from the University of Oxford Text Archive (ota.ox.ac.uk). The Masters dissertations in Arts and Humanities and in the Social Sciences which obtained Merit or Distinction were extracted. This yielded a control corpus of Expert User writing of just fewer than one million words. This BAWE corpus is smaller than my various sub-corpora. However, the measure of significance used in Wmatrix3 (i.e. log likelihood) already makes adjustments for differing corpus size (Rayson, 2003; Rayson, 2009; Dunning, 1993).

	Occurrences in BAWE		Occurrences in Buidcorp			
United_Arab_Emirates	1	0	201	0.04	-	435.55
private_schools	1	0	178	0.04	-	384.6
ministry_of_education	1	0	108	0.02	-	229.8
public_schools	2	0	89	0.02	-	180.45
high_school	6	0	97	0.02	-	174.91
human_rights	236	0.03	3	0	+	162.51
higher_education	15	0	106	0.02	-	157.18
United_States	210	0.02	1	0	+	156.87
a_lot	72	0.01	157	0.04	-	121.66
academic_year	4	0	57	0.01	-	100.52
Hong_Kong	148	0.02	3	0	+	95.16
special_needs	2	0	43	0.01	-	80.93
east_Asia	99	0.01	0	0	+	78.89
international_relations	126	0.01	3	0	+	78.58
english_speaking	2	0	40	0.01	-	74.54
North_Korea	90	0.01	0	0	+	71.72
i_think	2	0	38	0.01	-	70.29
rather_than	411	0.04	81	0.02	+	67.67
developing_countries	115	0.01	4	0	+	65.53
education_system	8	0	47	0.01	-	65.36
mother_tongue	8	0	46	0.01	-	63.45
exchange_rate	75	0.01	0	0	+	59.76
going_to	46	0.01	86	0.02	-	57.39
primary_school	4	0	36	0.01	-	57.3
in_fact	249	0.03	39	0.01	+	56.8

Table 3 Wmatrix3 counts MWEs as one word using underscores

The would-be language detective can pick up on interesting levels of Log Likelihood (which its inventor calls a measure of surprisingness and follow through with concordances of the most interesting MWEs. four entries from Table 3 will serve to illustrate the power of the concept of MWE as a tool for increasing language awareness. There is a startling contrast between the BAWE writers and the Buidcorp writers in relation to their interest in human rights (236 mentions in BAWE and 0 in Buidcorp. This could be said to be at the level of content analysis. *I think* is overused by the Buidcorp writers. This epistemic stance marker, usually assertive, sometimes a hedge, is more frequent in spoken English. The 'good' writers of BAWE almost completely abstain from it in their dissertations. On the contrary, these high achieving writers use *in fact* much more often as a stance marker (249 vs. 39).

I brought to the Malaga conference some initial findings and research ideas. I received from the participants searching questions and insightful suggestions as to how my research might continue. I thank the conference organizers for such a great opportunity to learn.

References

BERBER SARDINHA, A. 1996. Writing assessment and corpus linguistics. Paper presented at the Applications of Corpus Linguistics Seminar, Aston University, 19/04/1996.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. and FINEGAN, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman

BOLINGER, D. 1975. *Aspects of language*. Second edition. New York: Harcourt Brace Jovanovich.

CHENG, W. 2012. *Exploring Corpus Linguistics: Language in Action*. Milton Park, Abingdon: Routledge.

DE COCK, S., GRANGER, S., LEECH, G. and MCENERY, T. 1998. An automated approach to the phrasicon of EFL learners. In S. Granger (ed.) *Learner English on Computer*. London: Longman. 67-79.

FOX, G. 1998. Using data in the classroom. In B.Tomlinson (ed.) *Materials Development in Language Teaching*. Cambridge: Cambridge University Press. pp.25-43.

HOWARTH, P. 1998. Phraseology and second language proficiency. *Applied Linguistics* 19(1): pp. 24-44.

HYLAND, K. 2005. *Metadiscourse*. London: Continuum.

MAHLBERG, M. 2007. Corpus stylistics: bridging the gap between linguistics and literary studies. 191-208. In HOEY, M. MAHLBERG, M., STUBBS, M. and TEUBERT, W. 2007 *Text, Discourse and Corpora. Theory and Analysis*. London: Continuum.

MARTINEZ, R. and SCHMITT, N. 2012. A Phrasal Expressions List. *Applied Linguistics:* 33 (3): pp. 299-320.

MCKENNY, J. 2011. *A corpus study of the phraseology of written argumentative English*. Frankfurt: Lambert Academic Publishing.

MORLEY, J. 2014.Phrasebank. < http://www.phrasebank.manchester.ac.uk> [Accessed 25 May 2015]

PAWLEY, A. and SYDER, F. H. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. C. Richards and R. W. Schmidt (eds.) *Language and Communication*. London: Longman. Pp.191-226.

PARE, A., STARKE-MEYERRING, D. and MCALPINE, L. 9 The Dissertation as Multi-Genre: Many Readers, Many Readings. *PERSPECTIVES ON WRITING* (2007): p.179.

RAYSON, P. 2003. *Matrix: a statistical method and software tool for linguistic analysis through* corpus *comparison*. Unpublished Ph.D. thesis. Lancaster University.

RAYSON, P 2009. Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix [Accessed 20 May 2015] SALAZAR, D. 2011. Lexical bundles in scientific English: A corpus-based study of native and non-native writing. Unpublished PhD thesis, University of Barcelona.

SCOTT, M. (2012). *Wordsmith Tools, Version 6*. Oxford: Oxford University Press. SIMPSON-VLACH R. and ELLIS N. 2010. An Academic Formulas List: New Methods in Phraseology Research. Applied Linguistics. 31 (4): pp. 487-512.].

VERBALE KOLLOKATIONEN. *JEDER KENNT SEINEN PLATZ. JEDER WEIß, WO SEIN PLATZ IST* [83]

Nikoleta Olexová
Universität der hl. Kyrill und Methodius, Trnava, SR
nika.olex@gmail.com

Abstract

Wir kommunizieren nicht nur mit einzelnen Wörtern, sondern mit festen, bzw. typischen oder usuellen Wortkombinationen, die wir zur Bildung der sinnvollen Sätze verwenden. Im Mittelpunkt des vorliegenden Beitrags stehen Kollokationen, die zum Bereich der festen Wortverbindungen gehören und einen aktiven Bestandteil des Wortschatzes jeder Sprache bilden.

Der Grund für die Auswahl des Themas im Hinblick auf die aktuelle Situation der wissenschaftlichen Forschung der Kollokationen ist, die Ursachen und die Faktoren des kollokationellen Verhaltens der Wörter und präferierte Kombinatorik einzelner Kollokationskomponenten weiter zu erforschen. Es handelt sich um eine empirisch basierte korpusgestützte Untersuchung von ausgewählten kognitiven Verben *kennen* und *wissen* im Deutschen, mit dem Ziel, neue Erkenntnisse über ihre Semantik, Kollokabilität und Gebrauchsspezifik zu gewinnen und eventuelle Korrektur und Ergänzungen der lexikografischen Beschreibung vorzuschlagen. Es handelt sich um experimentale Forschung, die primär die Methoden und Mittel der Korpuslinguistik ausnutzt. Die Forschung beruht auf der Annahme, dass mit Hilfe von einer detaillierten Korpusanalyse des gewonnenen empirischen Materials ein detailliertes Bild über die Kollokabilität der untersuchten kognitiven Verben gegeben werden kann. Nach der Gewinnung der Informationen über Kollokationspotenzial der Verben können wir ihre Bedeutungsstruktur anhand von den lexikografischen Standardwerken in der Konfrontation mit gewonnenen Datenmengen aus Korpora überprüfen. Wir brauchen umfangreiche empirische Basis, um zuverlässige Ergebnisse für die Erstellung der Kollokationsprofile zu gewinnen. Die Ermittlung der Kollokabilität der Lexeme erfolgt auf Grund der Datengewinnung aus Korpora und lexikographischen Quellen. Die methodologische Ausgangsbasis der Forschung und Ermittlung der Kollokationen im Rahmen der Lexikographie stellen Erklärungswörterbüchern (Duden, DWDS, Leo.de u.a.), Kollokationswörterbüchern von Quasthoff und von Häcki Buhofer et al. Feste Wortverbindungen des Deutschen, Wörterbuch deutscher Präpositionen von Müller und Valenzwörterbüchern Wörterbuch zur Valenz und Distribution deutscher Verben von Helbig und Schenkel, VALBU u.a. als primäre Informationsquelle über das Aktantenpotential und Rektion der untersuchten Verben. Bei den korpusbasierten Untersuchungen stützen wir uns auf die Korpora (DeReKo, DeTeTen13, DWDS) und Kookkurrenzdatenbank (CCDB).

Die experimentale Erforschung der Kollokabilität der verbalen Kollokationen stützt sich auf die frequenzbasierten und musterbasierten Methoden der Korpuslinguistik. Im Vordergrund stehen die mathematisch-statistischen Modellen, die die Kookkurenz und die Kollokabilität der Verben im Text ermitteln können. Als sehr effektiv zur Unterscheidung zwischen festen und freien Wortverbindungen haben sich die kombinierten Verfahren der "7K-Methode"[84] (vgl. Ďurčo 2014) und die Anwendung des

[83] Der Artikel entstand im Rahmen des Projekts VEGA 1/075/14 Verbal collocations in German and Slovak (2014-2016)

[84] (1) Kookkurrenztest ermittelt die lexikalische Kollokabilität der untersuchten Einheiten und geht von der statistischen Ermittlung anhand von Korpora aus. Es handelt sich um erste Ermittlung der Kollokabilität der Lexeme. Es wird untersucht, ob die Lexeme gemeinsam vorkommen. Ergebnisse des Kookkurrenztests können

Frequenzkriteriums bewährt. Am objektivsten dokumentiert die Typikalität der Kollokationen die Frequenzdistribution mit vordefinierten paradigmatischen und syntagmatischen Filtern. Die Analyse der Frequenzdistribution identifiziert Kollokationen und zeigt klar und deutlich die typische Umgebung zum untersuchten Basislexem. Außerdem werden die statistischen Maße (MI-score, t-score, logDice, log likelyhood) verwendet, die die Kollokatoren zur untersuchten Basis aus unterschiedlichen Gesichtspunkten ermitteln. Das Verhalten der Kollokationen kann man nicht nur mittels der statistischen Maße, sondern auch in strukturellen Kettenmodellen erforschen. Eine manuelle Sortierung und linguistische Analyse sind trotzdem immer notwendig. Empirische Basis unserer Untersuchung stellt die Ausarbeitung der detaillierten Kollokationsprofile der ausgewählten kognitiven Verben dar. Die Grundschablone für Identifizierung der verbalen Kollokationen enthält folgende kombinatorische Beziehungen des Verbs: Infinitiv des Verbs mit Substantiv in Genitiv, Dativ, Akkusativ, Attribut oder mit anderem Verb; Negation mit Verb; Verb in erster, zweiter oder dritter Person in Singular oder Plural; Imperativ; Partizip 1 und 2.

Die gewonnen Erkenntnisse aus der Kollokationsforschung stellen eine solide, empirisch erprobte materielle Basis für eine vertiefte kontrastive Beschreibung und sie bilden eine neue Basis für eine detailliertere Beschreibung der morphologischen und syntaktischen Eigenschaften der Lexik und ihrer Verwendung, was natürlich auch zur Verbesserung der Muttersprache- und Fremdsprachendidaktik beitragen kann.

1. SEMANTISCHE UND DISTRIBUTIONELLE ANALYSE DER VERBEN KENNEN UND WISSEN

Auf den ersten Blick sind die Bedeutungserklärungen der ausgewählten kognitiven Verben *kennen* und *wissen* ähnlich strukturiert, doch es gibt ein paar Unterschiede. Beim Verb *kennen* erfasst Duden-WB 6 Bedeutungen und DWDS-WB spricht von 3 Bedeutungen. Duden-WB teilt die Mehrdeutigkeit des Verbs *wissen* in 5 Bedeutungen ein und DWDS-WB führt 2 Bedeutungen an. Duden-WB erfasst breiter gefächerte polysemantische Struktur im Unterschied zu DWDS-WB, das reicher um Bedeutungsuntergruppen bei beiden Verben ist. Mit diesem

folgendermaßen aussehen: *Bescheid… wissen; den Grund… kennen.* (2) Kollokationstest dient zur Unterscheidung der zufälligen und frequentierten Kookkurrenzen von realen syntagmatischen Konstruktionen aus der Sicht ihrer realen syntagmatischen Integrität. a) zufällige, frequentierte Kookkurrenzen: *Hörensagen* (?) *kennen* (Frequenz 346 in DeTenTen13) vs. Kollokation: *etwas/jemanden vom Hörensagen kennen* (Frequenz 312 in DeTenTen13), *etwas/jemanden nur vom Hörensagen kennen* (Frequenz 193 in DeTenTen). Durch den (3) Kategorientest wird die Vollständigkeit bzw. Unvollständigkeit der paradigmatischen Formen der Komponenten der festen Wortverbindung überprüft. Es werden die morphologischen Wortparadigmen untersucht. Z.B. Kollokation: *jmdn. am Gang/ *Gang/ *Gänge/ *an den Gängen kennen* (gleichgeltend für die Verwendung in Plural) oder *jmds. (meine, seine…) Pappenheimer/ *Pappenheimer kennen*. Durch das Testen der Vertauschbarkeit der Komponenten, d.h. durch den (4) Kommutationstest stellt man im Rahmen der semantischen Paradigmen (Synonymie bzw. Antonymie) die Festigkeit und Freiheit der Wortverbindung fest. Z.B. a) *die Bedeutung kennen / *die Bedeutung wissen (um Bedeutung des Themas wissen)*, b) **Bescheid kennen / Bescheid wissen*. (5) Konstruktionstest überprüft die Möglichkeiten der syntaktischen Transformation der Wortverbindung; unzulässige Transformationen: *jmds. Handwerk kennen → *das gekannte Handwerk*, *den Weg wissen → *der gewusste Weg*. Beim (6) Kompositionstest wird die Weglassprobe der Komponenten berücksichtigt. Die semantische Zerlegung der Komponenten wird in Bezug auf die Gesamtbedeutung der Wortverbindung überprüft. Es werden zufällige syntaktische Nachbarkombinationen, sog. Kolligationen, feste oder freie Wortverbindungen entdeckt. Z.B. *Ich glaube, ich kann guten Gewissens sagen, dass ich jeden Meter dieses Films auswendig kenne.* Daraus folgen Kolligationen: *Films kennen*; freie Wortverbindung: *jeden Meter dieses Films kennen*; Kollokation: *etwas/jemanden auswendig kennen* (Frequenz 1167 im Korpus DeTenTen13). (7) Kontrastiver Test bestimmt die Idiosynkrasie der festen Wortverbindung vor dem Hintergrund des Vergleichs mit einer anderen Sprache. Dieser Test ist relevant für Konfrontationsbeschreibung der Lexeme und die Maße der Gemeinsamkeiten und Unterschiede der Kombinatorik der Konfrontationslexeme, die Eigenschaften und das Verhalten der festen Wortverbindungen werden analysiert. Deutlicher kann man Bedeutungsgliederung des Lexems für lexikographische Konfrontationsbeschreibung definieren. Äquivalente in der Ausgangssprache werden mit Äquivalenten in der Zielsprache verglichen. Z.B. Kollokation *den Rummel kennen* / *poznat' krik, hurhaj, zábavný park; / vediet' si poradit'; vediet', ako na to; vediet' si rady; vediet', ako ďalej.

Bedeutungsvergleich konnten wir feststellen, wie bemerkenswert die Gemeinsamkeiten und Unterschiede der Bedeutungsbeschreibung der Verben in beiden Lexika sind. Die Wörterbücher beschreiben einzelne Bedeutungen oft gleich oder mit anderen Wörten und mit Hilfe von semantischer Distribution kann man konkrete Kollokationen erfahren.

Wie bereits erwähnt wurde, bilden die Verben *kennen* und *wissen* partielle Synonyme, weil ihre polysemantische Struktur unterschiedlich ist. Was diese Verben zu Synonymen vereint ist 2. Bedeutung bei *wissen* und 6. a) Bedeutung bei *kennen* im DUDEN-WB i.S.v. ´sich einer Sache bewusst sein´. Canoo.net, VALBU, DUDEN-WB erfassen bei beiden Verben die Bedeutung ´Kenntnis von etwas haben´. Bei 1.c) Bedeutung im DWDS-WB bei *wissen* wird angeführt, dass es eine Möglichkeit *wissen* mit *kennen* zu erklären gibt. Austauschbarkeit des Ausdrucks ´über etw./ jmdn. Bescheid wissen´ mit *kennen* wird im DUDEN-WB unter 1. a) Bedeutung erwähnt. Was diese Verben zu Synonymen verteilt, ist bei *kennen* 2. Bedeutung im DWDS-WB und 1. c) Bedeutung im DUDEN-WB ´mit jmdm. bekannt sein´ und 2. Bedeutung im DUDEN-WB ´verstehen, beherrschen´. Bei *wissen* geht es um 2. Bedeutung im DUDEN-WB und 4. Bedeutung im DWDS-WB ´in der Lage sein, etwas zu tun wissen´.

2. 'KORPUSMUSTERANALYSE' (CPA)

Ausgehend von Hanks Grundidee *Wie man aus Wörtern Bedeutungen macht?* (vgl. Hanks: 2011) und vom seinen lexikonbasierten und korpusgesteuerten Ansatz, der erklärt, wie Wörter Kollokationsmuster bilden, die Bedeutung entstehen lassen, untersuchen wir die Verben *kennen* und *wissen*, auf ihre konkrete Verwendungsweise, und ihre Kollokationen, denen semantische Typen[85] entsprechen. Die Häufigkeit des Gebrauchs von Kollokationsmustern in verschiedenen Kontexten kann mit modernen Korpusanalysetools, wie z.B. der Sketch Engine (vgl. Kilgarriff et al., 2004) festgestellt werden. Bei der lexikalischen Korpusanalyse geht es zuerst darum, den normalen und typischen Sprachgebrauch (Norms) zu identifizieren. In einem zweiten Schritt kann der unübliche und vom Typischen abweichende Sprachgebrauch als linguistische Kreativität bezeichnet werden (Exploitation) (vgl. Hanks: 2013).

CPA konzentriert sich auf Verben. Das Verb steht im Vordergrund, weil es als strukturelles Zentrum des Satzes aufgefasst wird (vgl. Helbig, Schenkel: 1969). Es handelt sich um eine empirische Methode zur semantischen Analyse und anders gesagt, geht es um eine Methode der Identifizierung relevanter semantischer Typen und Kollokationen. Kollokationen können in Gruppen von semantischen Typen zusammengefasst werden. Jeder semantische Typ in der Ontologie wird mit einem Set von lexikalischen Elementen ausgestattet sein, basierend auf den im Korpus für jedes Muster gefundenen Beispielen. Die Bedeutung des Verbs wird von semantischen Typen seiner Argumente beeinflusst. Verschiedene semantische Werte von Argumenten aktivieren verschiedene Bedeutungen jedes Verbs.

Für Verben *kennen* und *wissen* schlagen wir folgende Klassifizierung ihrer semantischen Typen, die sich in eckigen Klammern befinden. Es geht um deutsche bearbeitete Version. Die Kursivschrift bedeutet zugegebene semantische Typen zur CPA.

kennen

[Einheit/Wesen]

[abstrakte Einheit]:[Konzept][Regel][Vorschlag]*[Vorwürfe][Meinung][Behauptung][Argument]*

[85] insgesamt 200 semantischen Typen von 700 englischen Verben

[Lüge][Vorstellung][Regeln][Grenzen],
[Quelle der Information]:[Dokument][Sprache][Medium]*[Situation],*
[Wert]:[Geldwert], [psychische Eigenschaft]:[Einstellung][Verhalten][Emotion], *[Ziel]*, [Mittel], [Zeitperiode]:[Zeitpunkt]*[Geschichte][Krieg][Datumsangaben]*,
[Energie]:[Wellenlänge]:[Ton][Signal][Licht]
[physisches Objekt]:[belebtes Wesen]:
[Mensch/Person]:*[Namen][Eigennamen][Arzt],[Tierwelt]:[Hund]*,[Pflanzenwelt]:*[Baum]*
[unbelebtes Wesen]:
[Artefakt]:[Gebäude]:*[Ministerium][Schule][Denkmal]*,[Essen][Getränke]
[Kleidung][Dokument]*[Buch]*[Waffe][Gerät]:[Computer],
[Fahrzeug]:[Straßenfahrzeug][Wasserfahrzeug][Flugzeug][Zug],
[Ort/Platz]:[Gebäude]*[Lokal][Restaurant][Weg][Land][Welt]*,
[Material/Stoff]:[Gas][Luft][Feuer][Feststoff]:[Glas][Metall][Holz],[Flüssigkeit]:[Wasser],*[Geräusch]:[Pfiff]*
[Eigenschaft]
[kognitive Eigenschaften]:[Fähigkeit][Charaktereigenschaft]*[Verhaltensweise][psychischer Zustand][jmds. Stärken/Schwächen][Mentalität][Schönheit][Gefühl][Schrecken]*,[Rolle]: [Status in der Gesellschaft],[äußere Merkmale]:[Gewicht]*[Maß- und Gewichtsgrößen/längen]*
[Eventualität]
[Ereignis][Vorgang][Handlung][Aktivität][Tätigkeit][Ergebnis][Sprechakt]*[Arbeit][Beruf][Handwerk][Schachspiel][Lösung][Sprache]*,[Prozess]:[Wetter],[Tatbestand]:[Beziehung][Krankheit]
[Teil]
[Teil der Musik]:[Ton][Akkord]*[Sinfonien]*,[Teil der Sprache/des Sprechakts]:[Name]
[Phrase][Wort]*[Antwort][Frage][Witz][Sprache]*,[Teil des physischen Objekts]:[Körper]
[Körperteil]*[Gesicht]*,[Teil des Gebäudes]:[Zimmer],[Teil der Pflanze],[Teil des Fahrzeugs]
[Gruppe]
[Menschengruppe]:[Institution]*[Volk]*,[Fahrzeuggruppe],[Tiergruppe],[Gruppe des physischen Objekts]

wissen
[Einheit/Wesen]:
[physisches Objekt]:
[belebtes Wesen]:[Person]
[unbelebtes Wesen]:[Ort/Platz]:[Gebäude][Weg][Gerät]:[Fahrzeug]
[abstrakte Einheit]:*[Name][Anschrift][Antwort][Beruf][Lösung][Nummer][Preis][Note][Anlass][Urteil][Strafe][Grund][Befehl][Losung]*[Regel]*[Ziel]*
[Eventualität]:[Ereignis][Vorgang][Handlung][Aktivität][Tätigkeit][Ergebnis]
[Bescheid][Hilfe][Rat][Dank], [Zeitperiode]:[Datumsangaben]
[Eigenschaft]:[Maß- und Gewichtsgrößen]

3. KOMMENTAR ZUM VORSCHLAG UND ZUSAMMENFASSUNG DER ERGEBNISSE DER ANALYSEN

Die vorliegende Studie setzte sich zum Ziel neue Erkenntnisse über Semantik, Kollokabilität und Gebrauchsspezifik der Verben *kennen* und *wissen* zu gewinnen und eventuelle Korrektur und Ergänzungen der lexikografischen Beschreibung vorzuschlagen. Neue Erkenntnisse aus der Kollokationsforschung stellen auch eine solide, empirisch erprobte materielle Basis für eine vertiefte Beschreibung der Gemeinsamkeiten und Unterschiede in der Kollokabilität der ausgewählten kognitiven Verben und auf die Bestimmung des Einflusses der Kollokabilität auf ihre Semantik. Es wurden bemerkenswerte Bedeutungsunterschiede der Verben festgestellt.

Dank der Kookkurrenzanalyse, der Korpusmusteranalyse, der Analyse der Präpositionen anhand vom Wörterbuch deutscher Präpositionen und Korpus deTenTen13 zeigen uns die Bedeutungsbeschreibungen der Verben, die ihre Polysemie bestätigen und im Vorschlag der Klassifizierung der Verben angeführt sind. Der Vorschlag umfasst die Kernbedeutung, ihren Gebrauchs- und Situationsaspekt und Kollokationen, die dank Analysen festgestellt wurden.

Die Analysen zeigen uns keine Selektionsrestriktionen der Akkusativobjekte bei *kennen*. Bei der abstrakten Einheit als Akkusativobjekt bei *kennen*, z.B. [Vorschlag] [Behauptung][Lüge] [Argument][Regel][Grenze] denken wir über 1. Bedeutung ´mit etwas bekannt sein´ nach, wobei Eventualität, z.B. [Ereignis][Vorgang][Handlung][Aktivität]

[Tätigkeit][Ergebnis][Arbeit][Beruf][Handwerk][Schachspiel][Lösung][Sprache] richtet unsere Aufmerksamkeit eher auf die 2. Bedeutung ´etwas verstehen, beherrschen´.

Beim Verb *kennen* sieht man oft die Verwendung in der Negationsform: *kein Maß/ Ziel kennen*= maßlos, ziellos handeln, *keine Furcht kennen*= furchtlos sein, *keine Grenzen kennen*= sehr groß sein (z.B. *die Jubel kannte keine Grenzen*= der Jubel war unbeschreiblich groß); übertreiben, *keine Nerven kennen*= nicht nervös werden; mutig/ ruhig bleiben, *kein Pardon kennen*= streng/ gnadenlos/ rücksichtslos sein, *keine Rücksicht/ Schonung kennen* = sie gehen rücksichtslos, schonungslos vor bzw. er nimmt nie Rücksicht, *keine Unterschiede, Schranken kennen*= sich über alle Unterschiede, Schranken hinwegsetzen. Negierende Wörter, die sich mit *wissen* verbinden, zeigt uns die Korpusanalyse folgende: *Ich weiß mir keinen Rat/ keine Hilfe*, d.h. ich bin ratlos/ hilflos.

Bei *wissen* muss man daran beachten, dass [belebtes Wesen][Mensch], wie z.B. *Er weiß einen Arzt (in der Nähe vom Bahnhof)/ den Dichter des Fausts.* näher spezifiziert bzw. räumlich lokalisiert sein muss. Durch Paraphrasierung *Ich kenne einen Arzt* und *Ich weiß einen Arzt* versuchen wir den semantischen Unterschied klarzumachen. Im ersten Fall kann man einen Arzt zur Gruppe der Bekannten zählen, der zweite Fall zeigt uns, dass man jmdn. räumlich lokalisieren kann, d.h. Ich weiß, wo es einen Arzt gibt. Am Beispiel des Dialogs *Wissen Sie einen Arzt? Ja, ich weiß einen in der Nähe des Bahnhofs. Und kennen Sie ihn?* versuchen wir weiter die semantische Kompatibilität der beiden Verben zu vergleichen. Die Verwendung des Verbs *wissen* im ersten Satz könnte damit begründet sein, dass es um oberflächliche Kenntnis geht. Weiter benutzt der Fragende das Verb *kennen*, weil er die Bestehung der inneren bzw. näheren Beziehung zwischen dem Befragte und Arzt voraussetzt und der Fragender kann ihm ihn empfehlen (vgl. Butulussi 1991).

[unbelebtes Wesen][Ort/Platz][Gebäude], z.B. *Er weiß das renommierteste Theater Hamburgs/ den höchsten Berg der deutschen Alpen* müssen gleichfalls lokalisiert sein. [Gebäude] in diesem Fall *Ich weiß das Haus, in dem er wohnt.* ist durch attributive Zusätze spezifiziert.

Wenn der Weg jemandem bekannt, nicht neu bzw. genauer bekannt ist, verwendet man *Er kennt den Weg.* Im Satz *Er weiß den Weg* kann man die Bedeutung des Satzes doppelt auffassen. Einerseits bezeichnet der Weg ein Gebiet, oder Gelände, die zum Begehen und Befahren dienen bzw. Gang, Fahrt. Anderseits fassen wir der Weg im Sinne der Richtung.

Man verwendet *Er weiß die Hauptstadt von Italien*, aber wir können nicht sagen **Er weiß Rom*, obwohl mit Rom dieselbe Entität bezeichnet wird wie mit der Hauptstadt von Italien. Entscheidend ist nun, dass die Kennzeichnung nicht direkt das Gewusste bezeichnet. *Er weiß die Hauptstadt von Italien* ist paraphrasiert mit ´Er weiß, was die Hauptstadt von Italien ist´, ich kann die Stadt lokalisieren. Mit dem Satz stellt man fest, dass er ein bestimmtes Wissen hat. Die nominalen Objekte bei wissen scheinen generell ein Wissen über etwas und nicht einen klaren Wissensinhalt selbst auszudrücken. Das kann man mit *Er kennt/ *weiß Rom* bestätigen und die Sätze zeigen, dass der Wissensinhalt bei kennen stehen kann (vgl. Engelen 2010).

Beim [Fahrzeug] sieht man possessive Erweiterung *Ich weiß den Zug von… nach…*

[abstrakte Einheit] bei *wissen* muss durch attributive Zusätze spezifiziert sein:**Er weiß die Neuigkeit. Ich weiß die Neuigkeit, die ihn so erschreckt hat.* Dieses gilt nicht für *Er weiß das Neueste/ das Beste.* Engelen (2010) unterscheidet in diesem Fall zwischen dem einfachen und komplexen Merkmal, wobei *Er weiß das Beste* als einfaches Merkmal aufgefasst wird. Engelen (2010) spricht dazu die Hypothese aus, dass je einfacher das mit dem betreffenden Wort benannte Merkmal ist,

desto mehr neigt man dazu, wissen, und nicht kennen zu verwenden: *Den Krimi schau ich mir nicht an. Ich weiß/ *kenne* (das wäre umfassender, komplexes Merkmal) *den Mörder schon.*

Die Verbindbarkeit *Name, Antwort, Beruf, Familienstand, Lösung, Nummer, Preis, Strafe, Grund, Befehl, Losung, Regel, Ziel… wissen* begründen wir damit, dass der betreffende Sachverhalt als Ergebnis eines Vorgangs oder einer Handlung erfasst wird. In den Sätzen *Er weiß die Lösung/ den Satz des Pythagoras…* bezieht sich die Bedeutung auf das Endergebnis einer Rechenaufgabe oder auf den Wortlaut. Die Substitution mit den Ausdrücken ´die Lösung bewusst, gewiss sein; die Lösung erfahren, sich erinnern´ ist gleichfalls zulässig. Nach dem Ersetzen mit *kennen* sind diese Reduktionen nicht gegeben. *Ich kenne die Lösung/ den Satz des Pythagoras.* bedeutet, dass mir sein Wortlaut oder sein Inhalt bekannt ist oder im anderen Kontext *die Lösung/ den Satz des Pythagoras* verstehen.

Semantischer Unterschied ist bemerkenswert zwischen *Er kennt den Namen*= der Name ist ihm bekannt/ nicht neu/ genauer bekannt und *Er weiß den* Namen= er weiß, wie der Name lautet; er weiß den Namen zu nennen; er erfährt den Namen; er erinnert sich an den Namen.

Wenn der betreffende Sachverhalt als Vorgang oder Handlung erfasst wird, sind *Benotung, Beurteilung, Veranlassung, Verurteilung, Bestrafung, Begründung, Anordnung, Vereinbarung, Regelung* mit *wissen* nicht zulässig. Im Fall der Erweiterung um einen Genitivus subjectivus bzw. objectivus, um ein von-, durch-Gefüge oder Possessivelement, z.B. *die Benotung des Kollegen, die Benotung dieser Arbeit durch den Kollegen, die Verbindung von nach… wissen* ist die Verbindbarkeit möglich.

Zur [Eventualität] gehören bei *wissen* [Ereignis][Vorgang][Handlung][Aktivität][Tätigkeit] [Ergebnis][Bescheid][Hilfe][Rat][Dank], die uns 2. Bedeutung ´ etwas zu tun können/ wissen zeigen. *Rat wissen*= gut raten können, *Hilfe wissen*= helfen können, *Bescheid wissen*= sich gut auskennen, sich durchhelfen. I.S.v. etwas zu tun können, z.B. *Ich wusste viel zu berichten*= Ich war in der Lage, viel zu berichten bildet das Verb wissen Infinitiv mit zu. Bei *kennen* ist das nicht möglich.

[Datumsangaben] bei *wissen* treten als Kennzeichnungen aus oder sind mit einem possessiven Element spezifiziert, z.B. *Ich weiß den Todestag Napoleons/ seinen Geburtstag. Er wusste schon den neuen Termin.*

[Eigenschaft] wie [Charaktereigenschaft]: *Charakter*, [Verhaltensweise]: *Verhaltensweise, finanzielles Gebaren,* [psychische Zustände]: *Vitalität, Schönheit, Hässlichkeit* sind beim Verb wissen unzulässig.

[Maß- und Gewichtsgrößen] sind mit einem possessiven Element spezifiziert: *Höchstgeschwindigkeit (dieses Wagens), seine Schuhgröße, seinen Kontostand, Cholesterinspiegel (seines Großvaters), sein Blutdruck, sein Alter* und können nicht um qualifizierende Attribute erweitert sind: **Ich weiß seinen enorm hohen Kontostand.*

Ziel der Analyse der Präpositionen, die sich mit Verben *kennen* und *wissen* verbinden, beruht darauf, alle relevanten Präpositionen ans Licht zu bringen. Ausgehend vom Müllers Wörterbuch deutscher Präpositionen, richten wir zuerst unsere Aufmerksamkeit auf Präpositionen: jmdn./etwas **als/ von** jmdn./etwas kennen, jmdn./etwas **an/aus** etwas kennen, etwas **aus/ bei** etwas wissen, jmd. weiß **um** etwas, etwas **über** jmdn./etwas wissen, **von** jmdm./etwas wissen, jmd. weiß etwas **zu** etwas.

Dank Korpus deTenTen13 wurden andere potentiellen Präpositionen und ihres Frequenzvorkommen ausgewertet, die der Korpusanalyse untergeworfen wurden und die Ergebnisse aus dieser Analyse stellen den Beitrag für den lexikografischen Vorschlag:

etwas/jmdn. *auf Frage, aufgrund Erfahrung, bis ins Innerste, dank der Arbeit, durch die Arbeit, im Land, innerhalb des Vereines, mit Sicherheit, mit/ nach Namen, seit Jahren, trotz Liebe, unter dem Begriff, vom Studium, während der Zeit, wegen Fehlens* kennen; etwas/jmdn. *anhand Daten, als Kind, auf Frage, aufgrund Daten, beim Malen, dank Erfahrung, durch Talent und Erfahrung, durch Medien, mit Sicherheit, ohne Zweifel, ohne Hilfe, seit/ nach Jahren, trotz Alters, während Schulzeit, wegen Krankheit* wissen.

Außerdem wurden in der Analyse lexikografische Quelle, wie DUDEN, DWDS, Leo, OWID, PONS, wortschatz.uni-leipzig.de, Feste Wortverbindungen des Deutschen: Kollokationen-Wörterbuch, Kollokationenwörterbuch von Quasthoff, redensarten-index.de, VALBU, e-Valbu, Wörterbuch zur Valenz und Distribution deutscher Verben berücksichtigt.

Die Probe der Präposition jmdn./etwas **an** etwas kennen i.S.v. ´jmdn./etwas an dem Genannten erkennen´:

• WB deutscher Präpositionen: *jmdn. an der Stimme* (5)*, an der Uniform* (-)*, am Gang* (16) *kennen; die Meise an den Feldern* (-) *kennen, ein Instrument am Klang* (1) *kennen*

Präposition **an** scheint uns problematisch und sie wird nicht in der Word Sketsch Liste angeführt. Die Umgebung des konkreten Verbs liefert nicht nur Word Sketsch Liste bzw. Konkordanzliste. Außerdem kann man die Umgebung des konkreten Verbs mittels CQL Formeln untersuchen. Die Spanne zwischen dem ausgesuchten Substantiv mit Präposition und dem Verb kann man mit Hilfe vom Eintrag []{0,4} einstellen. Man kann den Negativfilter anwenden, der die Konjunktionen, Symbole, Interpunktionszeichen verbietet. CQL Formel für Auftreten der konkreten Präposition **an** in der Umgebung des konkreten Verbs *kennen* ohne Anwendung des Negativfilters:

[lemma="kennen"][]{0,4}[tag="N."][lemma="an"][]{0,2}[tag="N."] oder

[tag="N."][lemma="an"][]{0,2}[tag="N."][]{0,4}[lemma="kennen"]

• Beispiele aus Korpora deTenTen13: *das Leben am Hof* (2) / *an Bord* (3) *kennen, Arbeit am Film* (4) *kennen, die Freude am Kochen* (1) *kennen, Bild am Computer* (2) *kennen; Gesprächsführung am Telefon* (2) *kennt, Die Ikone an der Spitze* (1) *kennen wir nicht.*

• VALBU, DWDS, DUDEN, Leo: Verwendung in den Ausdrücken, in denen die Art und Weise des Bekanntseins charakterisiert wird, wird kennen i.S.v. erkennen verwendet:

jmdn. am Gang, am Schritt, an der Stimme.

Im Sinne von CPA beschreiben wir mit Hilfe von semantischen Typen was alles kann man *an etwas kennen*:

• *jmdn./etwas*: [physisches Objekt]: [belebtes Wesen]:[Mensch],[Tierwelt]; [unbelebtes Wesen]:[Artefakt]; [abstrakte Einheit]; [Eigenschaft] ([Mensch][Tier][Instrument][Wort] [Leben][Bild][Arbeit][Stadt][Freude][Stress] *an etwas*: [abstrakte Einheit]; [Ort/Platz]; [Artefakt] ([Stimme][Gang][Kleidung][Feld][Hof][Bord][Film][Computer][Telefon][Klang] [Schritt]) *kennen*

Weitere präpositionalen Gruppen:

• *jmdn./etwas als jmdn./etwas*: [physisches Objekt]: [belebtes Wesen]: [Mensch], [abstrakte Einheit]: [Beruf], [Ort/Platz]: [Markt], [Kunstwerk], [Beispiel], [Teil] [charakterliche Eigenschaft] *kennen*

• *jmdn./etwas aus etwas*: [Quelle] (mündliche, schriftliche, audiovisuelle): [Erfahrung], [Erinnerung], [Anschauung], [Hand], [Effeff], [Institution], [Zeitperiode] *kennen*

• *jmdn./etwas von jmdm./etwas*: [physisches Objekt]: [belebtes Wesen]: [Mensch], [Quelle]: [Sehen] [Hörensagen], [Institution], [Zeitperiode] *kennen*

• *etwas aus etwas*: [Quelle] (mündliche, schriftliche, audiovisuelle): [Erfahrung], [Erinnerung], [Erleben], [Leben], [Forschung], [abstrakte Einheit]: [Sicht][Grund][Bereich], [Zeitperiode], [Institution], [Körperteile] *wissen*

• *etwas bei* [Gelegenheit][Thema][Problem][Handlung][Prozess][Vorgang] *wissen*

• *jmd. weiß um etwas*: [abstrakte Einheit][Eigenschaft][Gefühl][Fähigkeit][Zeitperiode] [Zustand][Ort/ Platz][Artefakt][Ereignis][Vorgang][Handlung]

• *über etwas/jmdn.*: [physisches Objekt]: [belebtes Wesen]: [Mensch], [abstrakte Einheit], [Zeitperiode], [Institution], [Prozess], [Fähigkeit] *wissen*

• *jmd. weiß von* [physisches Objekt]: [belebtes Wesen]: [Mensch], [unbelebtes Wesen], [abstrakte Einheit], [Zeitperiode], [Institution], [Eigenschaft], [Eventualität]

4. VORSCHLAG EINER KLASSIFIZIERUNG DES VERBS *KENNEN*

1. von etwas/ jemandem Kenntnis haben

Gebrauchsaspekt, Situationsaspekt: diese Kenntnis über jemanden/ etwas wird durch Erfahrung/ Wahrnehmung erworben; man weiß, was und wie etwas ist und kann das bezeichnen/ benennen/ beurteilen/ einschätzen; durch bestimmte Eigenschaften kennzeichnen bzw. schon einmal gesehen/ gehört/ gelesen/ erlebt haben; mit etwas/jmdm. (näher, genauer) bekannt sein

Kollokationen: *die Absicht*[86]*, die Adresse, jmds. Alter, die Angst, die Antwort, den Arzt, die Bedeutung, das Buch, die Gefahren, das Gefühl, das Geheimnis, das Geschäft, die Geschichte, das Gesetz, jmds. Gesicht, die Gewalt, jmds. Gewicht, die Gewohnheit, die Grenze, den Grund, die Heimat, jmds. Herkunft, den Kniff, jmds. Körper, das Lokal, die Lösung, die Meinung, den Mensch, jmds. Namen, den Ort/ Platz, das Restaurant, den Schmerz, jmds. Schönheit, jmds. Seele, jmds. (Schatten-)Seiten, die Stärken/ Schwächen, die Ursache, die Wahrheit, den Weg, die Welt, die Zukunft, den Zusammenhang…*

2. etwas verstehen, beherrschen

Kollokationen: *die Arbeit, jmds. Beruf/ Handwerk, das Schachspiel, die Sprache...*

Attribute:

jmdn./ etw. ansatzweise, in- und auswendig, flüchtig, gründlich, oberflächlich, persönlich...

Präpositionalgruppe:

jmdn./ etwas kennen als… z.B. zuverlässige Person/ schreckliches Erlebnis/ bescheiden…; jmdn./ etwas kennen wie jmds. Hosentasche/ Westentasche

jmdn./ etwas an der Stimme, am Gang/ Schritt; (die Antwort) auf Frage; etwas/ jmdn. aufgrund Erfahrung, aus Erfahrung/ Quelle, bis ins Innerste, dank der Arbeit, durch die Arbeit, im Land, innerhalb des Vereines, mit Sicherheit, mit/ nach Namen, seit Jahren, trotz Liebe, unter dem Begriff, vom Hörensagen, vom Studium, während der Zeit, wegen Fehlens kennen

Vorschlag einer Klassifizierung des Verbs *wissen*

1. etwas als (Er)kenntnis zur Verfügung haben

Gebrauchsaspekt, Situationsaspekt: erfahren, sich erinnern, sich bewusst/ gewusst sein, sich auskennen/ in etwas bewandert sein

Kollokationen: *jmds. Adresse, die Antwort, keinen Ausweg, jmds. Entscheidung, das Ergebnis, den Gedanken, das Geheimnis, ein Lokal, die Lösung, den Namen, das Passwort, den Rat, ein Restaurant, das Schlimmste, den Weg, das Wort*

[86] Unterstreichung bedeutet von uns zugegebene Kollokationen, die uns die Korpusanalyse zeigte und die in den untersuchten Lexika nicht vorgekommen sind.

2. etwas zu tun wissen (etwas tun können/ vermögen)

Kollokationen: *zu berichten/ helfen/ leben/ schaffen/ schätzen/ singen/ würdigen wissen*

mit jemandem nichts anzufangen wissen

jmdn. zu nehmen wissen

Attribute:

etwas auswendig, gewiss, hundertprozentig, intuitiv, schon lange... wissen

Präpositionalgruppe:

anhand Genannten Kenntnis haben: *anhand Daten*

als jmd. Kenntnis haben: *etwas als Kind wissen*

auf Genannten Kenntnis haben: *die Antwort auf Fragen*

aufgrund Genannten Kenntnis haben: *aufgrund Daten*

Kenntnis stammt aus Genannten: *etwas aus dem Kopf wissen, aus (zuverlässiger) Quelle*

über das Wissen bei Genannten verfügen: *beim Malen*

dank Genannten Kenntnis haben: *dank Erfahrung*

durch Genannten Kenntnis haben: *durch Talent und Erfahrung, durch Medien*

mit Genannten Kenntnis haben: *mit Sicherheit*

ohne Genannten Kenntnis haben: *ohne Zweifel, ohne Hilfe*

seit/ nach Genannten Kenntnis haben: *seit/ nach Jahren*

trotz Genannten Kenntnis haben: *trotz Alters*

jemand ist sich der Kenntnis bewusst, die um Genannte besteht: *um Angst, um die Bedeutung von etwas, um Geheimnis, um jmds. Nöte, um Problem, um Qualität, um Wert*

sich auskennen, informiert sein (über das Genannte Bescheid wissen): *über jmdn./ etwas Bescheid, über Vergangenheit*

Kenntnis von der Existenz des Genannten haben: *etwas (nur) vom Hörensagen, von jmds. Erkrankung, von Menschen, von diesem Plan, von dieser Sache, von dem Schmuck, von jmds. Schwierigkeiten*

während Genannten Kenntnis haben: *während Schulzeit*

wegen Genannten Kenntnis haben: *wegen Krankheit*

jmd. hat das Wissen, das zum Genannten gehört: *dazu muss man wissen, dass…*

References

(Books)

BUTULUSSI, E. 1991. *Studien zur Valenz kognitiver Verben im Deutschen und Neugriechischen.* Tübingen: Niemeyer.

ĎURČO, P. 2014. *K princípom kolokačnej lexikografie. (Extrakcia a spracovanie kolokácií s adjektívami).* In: Balleková, K. – Múcsková, G. (ed.): Prirodzený vývin jazyka a jazykové kontakty. Bratislava: Vydavateľstvo Veda (im Druck).

ĎURČO, P., BANÁŠOVÁ, M., HANZLÍČKOVÁ A. 2010., *Feste Wortverbindungen im Kontrast.* Trnava: Univerzita sv. Cyrila a Metoda.

HANKS, P. 2011. *Wie man aus Wörtern Bedeutungen macht: semantische Typen treffen Valenzen.* In: Engelberg, S., Holler, A. and Proost, K. IDS. Berlin/Boston: de Gruyter.

HANks, P. 2013. *Lexical analysis: norms and exploitations.* Cambridge, Mass: MIT Press.

HÄCKI BUHOFER A., Dräger M., Meier S., Roth T. 2014. *Feste Wortverbindungen des Deutschen*. Tübingen: Francke.

HELBIG, G., SCHENKEL, W. 1969. *Wörterbuch zur Valenz und Distribution deutscher Verben*. Leipzig: VEB Bibliographisches Institut.

EISENBERG, P. 1986. *Grundriss der deutschen Grammatik*. Stuttgart: Metzler.

ENGELEN, B. 2010. *Die syntaktische Selektionsbeschränkung „kein Eigenname" – Untersuchungen zum Akkusativobjekt bei dem Verb wissen.* In: Schwierige sprachliche Strukturen: Aufsätze zur deutschen Grammatik. Frankfurt am Main: Lang.

KÁŇA, T. 2014. *Sprachkorpora in Unterricht und Forschung DaF/DaZ*. Brno: Masarykova univerzita.

LATZEL, S. 1978. *Die Verben "wissen", "kennen" und "können": eine Bedeutungs- und Gebrauchsbeschreibung mit Übungen für das Fachdeutsch als Fremdsprache*. München: Goethe-Inst.

MÜLLER, W. 2012. *Das Wörterbuch deutscher Präpositionen: die Verwendung als Anschluss an Verben, Substantive, Adjektive und Adverbien*. Berlin [u.a.]: De Gruyter

QUASTHOff, U. 2010. *Wörterbuch der Kollokationen im Deutschen*. Berlin: de Gruyter.

SCHUMACHER, H., KUBCZAK, J., SCHMIDT, R., DE RUITER, V. 2004. *VALBU – Valenzwörterbuch deutscher Verben*. Tübingen: Gunter Narr Verlag.

WAHRIG, G. 2006. *Wahrig Deutsches Wörterbuch: mit einem Lexikon der Sprachlehre.* München: Wissen-Media-Verl.

FÜHLEN ODER EMPFINDEN? EIN VERGLEICH DER KOOKKURRENZPROFILE DER PARTIELLEN SYNONYME[87]

Gabriela Orsolya
Universität der hl. Kyrill und Method in Trnava
lengabi@yahoo.com

Abstrakt

Die Kollokationen bilden heute ein heißes Thema in der linguistischen Forschung. Sie sind "Halbfertigprodukte" (Hausmann, (1984: 398) in der Sprache, die in unserem Gehirn gespeichert und in der Sprachproduktion automatisch hervorrufbar sind. Zum Gegenstand des vorliegenden Beitrags wird eine kontrastive Analyse der Kollokabilität der Verben *„fühlen"* und *„empfinden"* gestellt. Nach der Analyse des kollokationellen Verhaltens der untersuchten Verben wird versucht, die Bedeutungsstrukturen der Verben vor dem Hintergrund ihrer aktuellen lexikografischen Bearbeitung zu (re)interpretieren.

1. ZUR THEORETISCHEN GRUNDLAGE

Das Phänomen „Kollokation" verfügt in der Sprachwissenschaft über großes Diskussionspotential. Es gibt mehrere Betrachtungsweisen, wie Kollokationen in der Linguistik definiert werden. Wie Cieszkovszki (2003: 39) sagt, es herrsche eine Art „definitorischer Verwirrung" in der Kollokationsforschung[88].

Kollokationen sind eine Art von festen Wortverbindungen, die ihren Platz aus der Sicht des Idiomatizitätsgrades an der Peripherie des phraseologischen Bestandes haben. Hausmann beschreibt Kollokationen als Verbindungen von Wörtern mit begrenzter Kombinierbarkeit, die auf „differenzierten semantischen Regeln und einer gewissen zusätzlichen Üblichkeit" (Hausmann, 1984: 398) basieren. Wir verstehen unter Kollokationen sinnvolle Wortverbindungen, deren Entstehung sowohl durch die Kollokabilität als auch Kompatibilität bedingt ist, wobei aber die Frequenz und andere

[87] Der Artikel entstand im Rahmen des Projektes VEGA 1/0751/14 Verbal collocations in German and Slovak (2014 - 2016).

[88] In Konecny 2010 findet man mehr zu der Problematik der Begriffsbestimmung von Kollokationen

statistische, linguistische und pragmatische Kriterien auch eine wesentliche Rolle spielen (vgl. Ďurčo et al., 2010: 15).

2. METHODOLOGISCHES VERFAHREN

Bei der Analyse gehen wir von der Basis-Kollokator-Beziehung[89] aus. Als Basis dient das Matrixmodell von Ďurčo (2007), das für die Erfassung der Kollokationen alle primär binären Kombinationen von autosemantischen Wortarten wie Substantiv, Verb, Adjektiv, Adverb bezogen auf ihre konkret usualisierte Wortformenkombinatorik umfasst.

Mit Hilfe dieser universellen kombinatorischen Matrix wird versucht, ein möglichst vollständiges und komplexes Bild über die Kookkurrenzpartner von untersuchten lexikalischen Einheiten zu erfassen.

Für das Verb werden folgende Kombinationen untersucht:

Infinitiv	Negation	Singular	Plural	Partizipien
SubNom+VerbInf	Neg + Verb	VerbSg1 +	VerbPl1 +	Part I
SubGen+VerbInf		VerbSg2 +	VerbPl2 +	Part II
SubDat+VerbInf		VerbSg3 +	VerbPl3 +	
SubAkk+VerbInf		VerbImp +	VerbImp +	
Attr+VerbInf				
Verb2+VerbInf				

Abb. 1.

Die Analyse der Kompatibilität der lexikalischen Einheiten wird mit Hilfe von korpuslinguistischen und lexikographischen Methodik der Kollokationsforschung durchgeführt.

Als primäre Wörterbücher wurden das DUDEN Universalwörterbuch, DWDS online Wörterbuch verwendet[90]. Die in diesen Nachschlagewerken dargestellten Artikel werden zu den untersuchten Verben auf ihre Komplexität hin überprüft. Als Anschluss darauf wird ein Vorschlag für ihre Aktualisierung und Erweiterung mit relevanten Kollokationen entworfen.

Das Kollokationsprofil des Basislexems wird mit Hilfe von Korpora aufgrund von Häufigkeitsvorkommen mit mathematisch-statistischen Modellen und Methoden erstellt. Als führendes Kriterium dient das Frequenzkriterium, wobei wir die Tatsache betonen, dass „die Frequenz und Festigkeit der Bindung einer Kollokation keine direkt zusammenhängenden Korrelationen sind, da nicht alle festen Wortverbindungen auch hochfrequent sind und vice versa." (Ďurčo 2010: 6)

[89] Mehr dazu in Hausmann 1984, Konecny 2010

[90]Weitere Wörterbücher, die benutzt worden sind: elexico, dict.cc, dict.leo.org, Lingea digitales Deutsch-Slowakisches Wörterbuch, Web-Seiten, wo die Wörterbücher erreichbar sind: elexiko: www.owid.de, dict.cc: http://www.dict.cc/, English-German and Multilingual Dictionary, dict.leo.org: https://dict.leo.org/, Lingea WB – digitale Version, Lingea Lexicon Platinum 5, 2009.

Als primäres Korpus wird das DeTenTen13 mit der Suchmaschine *Sketch Engine*[91] benutzt. Weitere Korpora sind das DWDS-Korpora[92], der DeReKo Korpus mit Cosmas II[93] und die Kookkurrenzdatenbank CCDB[94].

3. BEDEUTUNGSSTRUKTUREN

Anhand der aktuellen lexikografischen Bearbeitung beider Verben wurde ein Vergleich durchgeführt.

Die Abbildung 2 fasst die Bedeutungen des Verbs *fühlen* zusammen:

FÜHLEN	
Bedeutungserklärung DUDEN/DWDS	**Beispiele**[95]
1a. Mit dem Tastsinn, den Nerven wahrnehmen, körperlich spüren 1b. Etwas körperlich und seelisch wahrnehmen	*einen Schmerz, die Wärme der Sonne fühlen,* *Er fühlte sein Herz schlagen,* *Kribbeln fühlen, u.a.* *die Not bitter fühlen,* *Alle Knochen fühlen, u.a.*
2.Seelisch empfinden, etw., besonders einen seelischen Vorgang, innerlich wahrnehmen	*Etw. instinktiv fühlen, Mitleid/Hass/Angst /den Drang fühlen,* *sie fühlen als Franzose* *Mit jmdm . fühlen, u.a.*
3a. Tastend nach etw. suchen, 3b. etw. durch Tasten prüfen	*Er fühlt den Puls,* *nach etw. fühlen, u.a.*

Abb. 2

Abbildung 3 demonstriert die Bedeutungen des Verbs *empfinden*:

EMPFINDEN	
Bedeutungserklärung DUDEN/DWDS	**Beispiele**
1.Als einen über die Sinne vermittelten Reiz wahrnehmen, verspüren, etw. durch Sinnesorgane wahrnehmen	*Hunger, Wärme, Schmerz, Lärm empfinden, u.a.*
2.Eine bestimmte Gemütsbewegung erfahren, erleiden, von einer bestimmten Emotion erfüllt sein	*Freude, Stolz, Dankbarkeit, Ekel, Scham, Vergnügen, Abneigung,* *etw. dunkel, empfinden,* *mit jmdm. empfinden,* *etw. als Beleidigung empfinden,* *er empfindet nichts für sie., u.a.*
3.In bestimmter Weise spüren, auffassen, für etw. halten	*etw. als kränkend empfinden, u.a.*

Abb. 3

[91] erreichbar unter Webseite: http://www.sketchengine.co.uk/. DeTenTen13 Korpus enthält 19,918,263,493 Token und 16,534,176,369 Wörter und damit gibt es eine relevante materielle Basis zur Untersuchung.

[92]Erreichbar unter: http://www.dwds.de/ressourcen/korpora/

[93] erreichbar unter http://www.ids-mannheim.de/cosmas2/

[94] http://corpora.ids-mannheim.de/ccdb/

[95] Beispiele sind aus allen oben angeführten Wörterbücher genommen

Wir haben folgende Bemerkungen[96]:

Eine synonymische Beziehung zwischen den Verben besteht in den Bedeutungen:

- Körperliche Wahrnehmung: empfinden, fühlen: *Schmerz empfinden/fühlen*
- Empfindung (sinnliche Wahrnehmung, Wahrnehmen durch die Sinnesorgane, innerliche Wahrnehmung): empfinden, fühlen: *Mitleid empfinden/fühlen*

Unterschiede lassen sich auch aufweisen:

Fühlen:

- Nach etwas tastend suchen, durch Tasten prüfen: fühlen nach etw. (= berühren, abtasten)*: Den Puls fühlen, nach dem Portemonnaie fühlen*

Empfinden:

- In bestimmter Weise etw. spüren, auffassen: *etw. als kränkend empfinden*

4. KORPUSLINGUISTISCHE ANALYSE

Die Verben *fühlen* und *empfinden* drücken einen Wahrnehmungsprozess aus. Schreiber/Sommerfeldt/Starke (1987: 77) definieren die Wahrnehmung als „Wiederspiegelung von Gegenständen und Erscheinungen der objektiven Realität mit Hilfe von Sinnesorgane von Menschen und Tieren." Die beiden Verben zeigen dieselbe Argumentstruktur auf: sie verlangen ein Agens und mindestens ein Patient. Als weitere Argumente in der Umgebung der Verben können noch lokale, modale und temporale Adverbialbestimmungen auftreten.

4.1 Die Kombination SubNom + VerbInf.

Als Subjekte kommen am meistens Lebewesen vor: Menschen und Tiere. Die häufigsten gemeinsamen Kollokatoren sind z.B.: *Betroffene, Patient, Mensch, Eltern, Frau, Mann, Bewohner, Kind, Jugendliche, Hund, Katze, Fisch, Kanninchen,* [97] *u.a.*

„auch der Hund Schmerzen empfindet, wenn er grob angefasst wird, (…)
„Alle Menschen fühlen gute und schlechte Energien"

Weitere Subjekte sind: Kollektiva *(z.B. Europäer, Deutsche*) und Institution *(z.B. Redaktion, die Regierung).*

Körperteile können auch als Subjekt auftreten: z.B. Haut, Lippen, Körper, u.a.:

„meine Zunge fühlt eine Lücke"
„und unser Körper empfindet/fühlt diese Gefühle augenblicklich"

[96]Die reflexive Form des Verbs *fühlen* wurde während der Analyse außer Acht gelassen. Bei dem Vergleich gingen wir von den Wörterbücher DUDEN.de und DWDS.de aus. Beispiele und eine ausführliche Bedeutungsgliederung findet man unter die Web-Seiten: www.duden.de, www.dwds.de

[97] Die Beispiele sind aus TeTenTen13 Korpus entnommen

4.2 Die Kombination SubsAkk + VerbInf

4.2.1 Ähnlichkeiten in der Kollokabilität

Die korpuslinguistische Untersuchung brachte die größte Zahl von Kollokatoren im Akkusativ, da beide Verben transitiv sind.

Die Listen demonstrieren die meist frequentierten Kollokatoren aus dem Korpus DWDS und DeTenTen13[98]:

Fühlen aus DWDS/ Frequenz:

Puls (315), Schmerz (307), Verantwortung (142), Druck (137), Kraft (100), Bedürfnis (99), Blicke (86), Verpflichtung (84), Leere (77), Mitleid (76), Liebe (70), Schuld (52), Energie (46), Trauer (45), Hass (49), Angst (43) Erleichterung (43), Drang (41), Wärme (38), Freude (37),

Fühlen aus DeTenTen13 /Frequenz:

Puls (1549), Schmerz (3840), Zahn (1374), Kraft (1331), Liebe (1201), Gefühl (1197), Druck (1069), Energie (1000), Haut (911), Angst (842), Wärme (991), Bedürfnis (652), Leere (569), Drang (494), Verbundenheit (429), Wut (427), Kribbeln (370), Atem (365), Kälte (365), Lippe (309),

Empfinden aus DWDS/Frequenz:

Schmerz (411), Mitleid (410), Genugtuung (322), Sympathie (291), Gefühl (279), Freude (203), Liebe (176) Situation (160), Trauer (159), Hass (159), Arbeit (136), Leben (136), Respekt (131), Dankbarkeit (123), Unbehagen (115), Mitgefühl (101), Kritik (100), Schadenfreude (98), Zuneigung (97), Ekel (94),

Empfinden aus DeTenTen13/Frequenz:

Dankbarkeit (1053), Sympathie (1212), Schmerz (4329), Freude (3111), Gefühl (3159), Liebe (1720), Situation (1393), Angst (1208), Lust (1048/977*[99]), Glück (941/760*), Verhalten (925), Zuneigung (792), Stress (689/361*), Trauer (682), Respekt (667), Scham (636/), Geruch (636), Mitgefühl (616), Wut (613), Bedürfnis (609).

Die Objekte sind meist Abstrakta. Beide Verben haben viele gemeinsame Kollokatoren, z.B.: *Leere, Freude, Liebe, Hass, Mitleid, Schmerz, Trauer, Wärme, Kälte, Angst, Einsamkeit, Sympathie, Sehnsucht, Leidenschaft, u.a.*

In den meisten Fällen geht es um die Bedeutung „etwas körperlich oder seelisch wahrnehmen", z.B.:

Verpflichtung fühlen/empfinden = etwas innerlich wahrnehmen

> *„wenn wir diese Verpflichtung fühlen (…), nur dann ist ihr Tod nicht umsonst gewesen."*
> *„eine besondere Verpflichtung empfinden wir gegenüber jungen Menschen, "*

[98]CCDB und IDS-Korpora mit Cosmas II geben die gleichen Kookkurrenzpartner an, mit anderer Rangordnung.

[99] Bei einigen Substantiven kann der Kollokator mit als oder in Akk. stehen. DeTenTen13 berücksichtigt es in Word Sketch nicht, ob das Substativ mit „als", also in als+Substantiv+Verb-Konstruktion (als Lust empfinden/Lust empfinden) oder nur im reinen Akkusativ, also SubAkk+VerbInf-Verbindung steht. Mit concordance Filter (negativ) und CQL-Formel (z.B. [word="Lust"][lemma="fühlen"]) wird nach der Zahl der Substantive ohne „als" recherchiert. * - ist ein Zeichen für die Zahl der Kollokatoren ohne „als"

Es gibt Wortverbindungen, in denen die Verben nicht die gleiche Semantik aufweisen, wie:

Hitze fühlen/empfinden:

„Er erklärte mir, dass er starke Hitze gefühlt hat."= körperlich wahrnehmen, spüren

„Feuchte Luft und Hitze empfanden schon die Römer als wohltuend."= etwas bewerten, erfassen in einer bestimmten Weise

Oft ist der Fall, dass ein Substantiv in Akk. mit dem Verb *empfinden*, wie z.B. *ein Kribbeln empfinden*, auch in einer als-Konstruktion auftreten kann, aber mit unterschiedlicher Bedeutung: *etw. (als) Kribbeln empfinden*[100]*:*

„Die Impulse (…) werden von den Patienten (…) als Kribbeln empfunden." = Objekt: Impulse

„Mein Problem ist nur, dass ich entweder einen Schmerz oder ein Kribbeln empfinde."= Objekt: Kribbeln

Andere Beispiele sind: *etwas. als Bedrohung, als Zumutung, als Belastung, als Provokation, als Beleidigung, als Schande, als Fremdkörper, als Ohrfeige, u.a. empfinden*

Diese Konstruktionen werden unter der Kombination „Verb + Substantiv in Akkusativ" aufzeigt. Es geht hier aber um eine modale adverbiale Bestimmung[101].

Als Objekt können mit *empfinden* auch Lebewesen auftreten:

„Ich empfinde/fühle euch Schweizer als sehr standhaft gegenüber der ganzen Gehirnwäsche, "*

Es handelt sich um die Bedeutung: jemanden finden als jmdn/halten für etwas.

4.2.2 Restriktionen in der Kollokabilität

Mit „Sketch Differenz„ im DeTenTen13 und Wort-Profil Vergleich im DWDS wurden die Restriktionen in der Kollokabilität beider Verben analysiert.[102] Die Ergebnisse wurden mit gezielten CQL-Formeln[103] im Korpus DeTenTen13 nachgeprüft, um diese Ergebnisse zu bestätigen oder zu widerlegen. Die Ergebnisse sind in Abb. 4 und Abb. 5 präsentiert.

Substantive nur mit *empfinden* kompatibel:

[100] Das Beispiel wurde aus DeTenTen13 entnommen

[101] In diesen Fällen sind CQL-Formel und das negative Filter von großer Bedeutung. Man kann die nicht passenden Data ausfiltern.

[102] Beide Korpora geben unterschiedliche Ergebnisse an. Die Ergebnisse über die Verträglichkeit der Verben mit Substantiven im Akkusativ hängen sehr eng mit der Größe der Texte, der Textsorten, u.a. Faktoren in Korpora zusammen. Im Falle des Korpus DWDS handelt es sich um ca. 100 Millionen Textwörter in 79.830 Dokumenten. DeTenTen13 umfasst 16,534,176,369 Wörter in 49,993,176 Dokumenten. Die Ergebnisse aus den Korpora bedeuten aber nicht, dass die aufgelisteten Substantive nicht als Kookkurrenzpartner mit den untersuchten Verben auftreten können. Ein Grund ist dafür z.B., dass das DWDS-Korpus in einem Wortprofil-Vergleich auch Kollokatoren als nicht kompatibel anführt, die weniger signifikant sind. Hier sind die CQL-Formel von großer Bedeutung, mit denen Hilfe man nach konkreten Formen recherchieren kann.

[103] Benutzte CQL-Formeln: [tag="N.*Acc.*"] [lemma=”empfinden"], [tag="N.*Acc.*"][]{0,3} [lemma=”fühlen"], [tag="N.*Acc.*"][]{0,3}[lemma="fühlen"]

Ergebnisse im DWDS mit Wortprofil Vergleich: nicht kompatible Subst. mit *fühlen*	**Ergebnisse im DeTT13 mit Sketch Differenz: nicht kompatible Subst. mit *fühlen***	**Überprüfung in DTT13 mit CQL-Formeln: Subst. nur mit *empfinden* möglich**	**Überprüfung in DeTenTen13 mit CQL-Formeln: kompatible Subst. auch mit *fühlen***
Schadenfreude, Mitgefühl, Respekt, Vergnügen, Verhalten, Zustand, Äußerung, Auftritt, Hochachtung, Wehmut, Atmosphäre, Vorgehen, Achtung, Lage, Solidarität, Kritik, Urteil, Diskussion, Tatsache, Glücksgefühl, Triumph, Arbeit, Skrupel, Verlust, Scheu, Niederlage, Umgang, Angebot, Verhältnis, Tatsache, Ergebnis, Bewunderung, Entscheidung.	*Schadenfreude, Zumutung*	*Verhalten, Äußerung, Auftritt, Hochachtung für jmdn., Atmosphäre, Vorgehen, Solidarität, Kritik, Urteil, Diskussion, Triumpf, Arbeit, Verlust, Angebot, Verhältnis, Aussage, Ergebnis, Entscheidung, Zumutung.*	*Schadenfreude, Mitgefühl, Respekt, Vergnügen, Zustand, Wehmut, Achtung, Lage, Tatsache, Tatsache, Bewunderung* ***1x-2x in DTT13 als kompatibel vorkommende Substantive:*** *Glücksgefühl* *Skrupel* *Scheu* *Umgang empfinden*

Abb.4

Substantive nur mit *fühlen* kompatibel:

Ergebnisse im DWDS mit Wortprofil Vergleich: nicht kompatible Subst. mit *empfinden*	**Ergebnisse im DeTT13 mit Sketch Differenz: nicht kompatible Subst. mit *empfinden***	**Überprüfung in DTT13 mit CQL-Formeln: Subst. nur mit *fühlen* möglich**	**Überprüfung in DeTenTen13 mit CQL-Formeln: kompatible Subst. auch mit *empfinden***
Blut, Stich, Atem, Energie, Blick, Hand, Wind, Kraft, Drang, Herz, Puls.	*Puls, Herzschlag*	*Blut, Hand, Herz, Puls, Herzschlag*	*Stich, Energie, Blick, Wind, Kraft, Drang* ***2x in DTT13 als kompatibel vorkommende Substantive:*** *Atem empfinden*

Abb. 5

Es gibt auch Substantive, deren Verträglichkeit mit den untersuchten Verben mit Hilfe von Sketch Differenz nicht veranschaulicht wurden Das Überprüfen der Kompatibilität dieser Substantive mit den Verben *fühlen* und *empfinden* wurde manuell, gemäß dem Frequenzvorkommen im DeTenTen13 Korpus durchgeführt.

Weitere Substantive, die nur mit dem Verb *fühlen* vorkommen:

- Knubbel fühlen
- Muskelkater fühlen
- Puls fühlen
- Herzschlag fühlen
- Adrenalin fühlen
- Beule fühlen
- Schweiß fühlen
- Verwandtschaft fühlen
- Körperteile, z.B. Stirn fühlen

Substantive, die nur mit dem Verb *empfinden* vorkommen, die nicht in Sketch Differenz angeführt wurden:

- Schuldgefühl empfinden
- Geräusch empfinden

Die Wortkombination „Geräusch fühlen" kommt im DTT13 einmal in Passiv-Konstruktion vor: *„Grad 5: sehr lautes Geräusch, mit der Hand kann das Geräusch gefühlt werden"*

- (als) Lärm empfinden
- Situation, Arbeitsklima, Vorgehen, Äußerung, Verhalten wie empfinden
- Freundlichkeit empfinden

Die Grenzen in der Kollokabilität beider Verben sind sehr vage. Es ist schwer, die Unterschiede im Gebrauch der beiden Verben mit Substantiven in Akkusativ zu begreifen. Die Verben sind in der Bedeutung „etwas physisch und innerlich wahrnehmen" Synonyme. Wir halten die Charakteristika „instinktiv, emotional ↔ bewusst, rational" im Falle dieser Bedeutung für Schlüsselbegriffe, wobei die Eigenschaft „emotional" für das Verb *fühlen*, die Eigenschaft „rational" für das Verb *empfinden* charakteristisch ist:

Jemand/etwas fühlt *etwas:*

- etwas irgendwo physisch wahrnehmen (psychischer Vorgang, Zustand, Substanz, wie Schweiß, Atem, Schmerz, u.a.) = **erleben, erleiden**
- etwas gefühlmäßig wahrnehmen (Stimmungen, Gefühle) = **etwas ahnen, intuitiv erfassen**, **nicht bewusst, instinktiv**

→ EMOTIONAL

Jemand/etwas empfindet *etwas:*

- mit den Sinnen wahrnehmen, etw. physisch wahrnehmen = **erleben, aufnehmen, begreifen, sehen, hören**
- im Gemüt bewegen, innerlich wahrnehmen = **sich etwas (Gen.) bewusst werden, etwas begreifen**
- etwas wahrnehmen als etwas = **etwas halten für etwas, meinen**

→ RATIONAL

4.2.3 Mit präpositionalem Kasus im Akkusativ

Das Verb empfinden kann mit der Präposition *für* stehen:

- Jemand empfindet jemanden/etwas *für (Adj., Attr.)*

„ich habe dich für reifer empfunden… „

- Jemand empfindet *für jemanden etwas/nichts/mehr, u.a.* = Liebe fühlen, verliebt sein

„Aber ich empfinde so viel für diesen anderen Mann,"

- *„Für jemanden* irgendwie empfinden = hat ein Gefühl

„(…) Die Aussage Jesu, es gebe Gott und er empfinde für uns wie ein Papa ist beinahe zu bekannt, (…)"

4.3 Mit präpositionalem Kasus im Dativ

Die typische Umgebung der Verben *fühlen* und *empfinden* wird oft auch mit präpositionalem Dativ als lokale, temporale adverbiale Bestimmung erweitert. Aus den Korpora ergeben sich folgende Ergebnisse:

1. Etwas irgendwo fühlen/empfinden – körperlich oder innerlich irgendwo wahrnehmen: Lokale adverbiale Bestimmung
   - Etwas fühlen/empfinden an etwas (Dat): *Am Hals, etwas am (eigenen) Leib fühlen, u.a.*
   - Etwas fühlen/empfinden auf. etwas (Dat): *auf der Haut, auf dem Rücken u.a.*
   - Etwas fühlen/empfinden in etwas (Dat): *im Rücken, im Gesicht, im Bauch, in Knochen, im Nacken, in der Seele, im Innersten*
   - Etwas aus tiefster Seele fühlen/empfinden

2. *Etwas fühlen/empfinden bei etwas (Dat*) – etwas körperlich oder innerlich irgendwann wahrnehmen: Temporale adverbiale Bestimmung: *beim Laufen, beim Schwimmen, beim Fahren, beim Lesen, u.a.*

Die Verben *fühlen* und *empfinden* sind weiter mit der Präposition *mit* kompatibel:

1. mit jmdm. fühlen/empfinden = Mitgefühl mit jmdm. haben: *Mitleid fühlen mit jemandem,*

„Ich hatte Pech. So empfanden/fühlten mit mir andere Patienten."

Das Verb fühlen kann weiter mit folgenden Präpositionen stehen*:*

1. nach etw. fühlen:
   a. nach etw. tastend suchen: *ich fühle nach meinem Schlüssel*
   b. Durch Berühren der genannten Stelle dessen Befindlichkeit zu ergründen: *er fühlt nach seinem Herzen*
2. fühlen als: jmds. Gefühl entspricht ganz dem des Genannten: *wir fühlen als Deutscher*

Das Verb *empfinden* kommt mit folgenden Präpositionen vor:

1. empfinden etwas als: jemand nimmt jemanden/etwas als das Genannte wahr, fasst es so auf: z.B. *etwas als Zumutung empfinden*

„Warum fällt es so schwer, anderen Grenzen zu setzen, auch wenn wir ihre Anliegen als Zumutung empfinden?“

Die Beispiele weisen darauf hin, wie vage der Unterschied zwischen beiden Verben ist, besonders im Falle, wenn sie sowohl eine körperliche als auch eine innerliche Wahrnehmung ausdrücken können. Die Semantik des jeweiligen Verbs ist von seiner ganzen Umgebung und Argumenten sehr stark beeinflusst.

4.4 Die Kombination Attr+VerbInf

Die Verben haben viele gemeinsame Attribute, wie z.B.: *deutlich, genau, hautnah, intensiv, instinktiv, richtig, schmerzhaft, schmerzlich, förmlich, intuitiv, schnell, sofort, überall, zugleich, körperlich, persönlich, wenig, tief, kaum, stark, dunkel, seelisch, füreinander, anders, genauso, wohl, national/sozial/freiheitlich u.a.*

Beispiele:
„Wir können unsere Verantwortlichkeit so tief fühlen, (…)“
„(…) das Gefühl der Minderwertigkeit, das die kolonisierte Bevölkerung so tief empfindet.“

Als Unterschied zwischen beiden Verben sind die als-Konstruktionen zu erwähnen. Das Verb *fühlen* bildet keine Verbindungen mit als-Konstruktionen, wie es beim Verb *empfinden* oft der Fall ist.

Als + Adjektiv: etwas *als unterecht, als angenehm, als unangenehm, als bedrohlich, als unerträglich, als peinlich, als befriedigend, als schmerzlich, als negativ/positiv, als wohltuend, u.a. empfinden.*

Die als- Konstruktion lässt sich in eine Adjektiv + Verb-Verbindung um transformieren, ohne die Bedeutung des Satzes zu verändern:

„Ich glaube Dir, dass Du das Alleinsein als sehr schmerzlich empfindest.“ = Ich glaube dir, dass Du das Alleinsein sehr schmerzlich empfindest.

In einigen Fällen steht das Verb mit einem Adjektiv nur ohne „als“: *etw. tief/genauso/wohl/anders/u.a. empfinden.*

Die untersuchten Attribute fungieren als modale Adverbialbestimmungen, die die Art und Weise, eventuell die Intensität, bezeichnen, in denen der Vorgang des Wahrnehmens realisiert wurde.

5. VORSCHLAG DER KLASSIFIZIERUNG DER BEDEUTUNGSSTRUKTUREN

In dem vorliegenden Beitrag haben wir die Antwort auf die Frage gesucht, wo die Grenzen in der Kombinatorik von beiden Verben liegen und welche Folgen auf die polysemantische Struktur die Kollokationsprofile beider Verben haben?

Die Analyse beweist, dass beide Verben eine viel breitere Semantik aufweisen, als es ihre denotativen Kernbedeutungen andeuten und als sie in den Wörterbüchern präsentiert werden.

Untersucht wurden nicht nur die Substantive im Akkusativ[104], sondern auch alle Argumente, die die Umgebung von Verben erweitern. Diese werden in den Wörterbüchern nur selten behandelt.

Das Verb *fühlen:*

Nach einem Vergleich von Bedeutungsgliederungen in den Wörterbüchern kommt man zum Ergebnis, dass die Gliederungen nicht vollkommend sind. In beiden untersuchten Nachschlagewerken wird auch die reflexive Form des Verbs *fühlen* in seine Bedeutungsstruktur eingebaut. Wir schlagen vor, die zwei Bedeutungen getrennt in den Wörterbüchern zu behandeln, da es um zwei selbstständigen Verben mit unterschiedlicher Bedeutungsstruktur geht.

Der Umfang der Beispiele mit dem transitiven *fühlen* ist klein. In den meisten Fällen gehören die Beispiele zu der reflexiven Form des Verbs *sich fühlen*.

Wir schlagen folgende einheitliche Gliederung vor:

1. Mit dem Tastsinn, mit den Nerven wahrnehmen, körperlich etwas wahrnehmen (Körperteil, Vorgang), irgendwas, irgendwo, irgendwie, irgendwann:
   - *Was: die Wärme fühlen, die Spritze in der Haut fühlen,*
   - *Wo: an etwas (Dat.)/in etwas (Dat.)/ auf etwas (Dat.) fühlen (eigenen Leib / auf (eigenen) Haut fühlen, auf dem Rücken, im Hals, u.a.*
   - *Wie: stark/persönlich/instinktiv/intuitiv, u.a. fühlen*
   - *Wann: beim Lesen, beim Laufen, beim Schwimmen, u.a.*
2. Etwas (Gefühle, Stimmungen, Zustand) seelisch, psychisch wahrnehmen, ahnen, merken irgendwas, irgendwo, irgendwie:
   - *Was: Mitleid / Liebe /Verantwortung / Leere fühlen, ich fühle, ich bin auf der richtigen Spur, u.a.*
   - *Wo: im Herzen, am eigenen Leib, u.a.*
   - *Wie: sozial/national //instinktiv/intuitiv, u.a. fühlen*
   - *Mit jmdm. fühlen* (Mitgefühlt haben)
3. Etwas (Gegenstand, Körperteil) durch Tasten prüfen/tastend prüfen:
   - *Nach etwas fühlen: nach Portemonnaie, nach der Tasche, u.a.*
   - *Etwas fühlen: den Puls, den Herzschlag, u.a*
4. Fühlen als, jmds. Gefühl entspricht dem des Genannten: *wir fühlen als Europäer, Deutsche, u.a.*

Das Verb *empfinden*

Auf Grund der Korpusanalyse schlagen wir vor, das Wortartikel mit weiteren relevanten und signifikanten Kollokationen zu erweitern:

1. etwas (durch Sinnesorgane) körperlich wahrnehmen, irgendwo, irgendwie, irgendwann*:*
   - *Was: die Wärme/Hitze empfinden, u.a.*
   - *Wo: in etwas (Dat.)/auf etwas (Dat.)/an etwas (Dat.), u.a. empfinden*
   - *Wie: schmerzhaft/ körperlich/ tief, u.a. empfinden,*
   - *Wann: sofort/zugleich/beim Laufen, u.a.*
2. etwas seelisch wahrnehmen, im Gemüt bewegen, irgendwie, irgendwo, irgendwann:

[104] Die Eigenschaft „transitiv" ist ein relevanter Faktor bei beiden Verben. Aus diesem Grund die meist frequentierten Kookkurrenzpartner waren Substantive in Akkusativ. Sie zeigen sich als obligatorische Partner der untersuchten Verben.

- *Was: Sympathie / Trauer / Schadenfreude / empfinden* (= ein Gefühl, Stimmung), *eine Situation / das Leben / die Arbeit, u.a. empfinden* (= Ereignis, Prozess*),*
- Jemand empfindet für jemanden etwas/nichts/mehr, u.a. (Liebe fühlen, verliebt sein*): ich empfinde so viel für den Mann, u.a*
- *Wie: Etwas schmerzlich/ /tief /dunkel/wohl, u.a. empfinden,*
- *Wo: überall, im Herzen, im Innersten u.a.*
- *Wann: zugleich/sofort/ beim Lesen, u.a.*
- Mit jemanden empfinden (Mitleid haben*): ich empfinde mit dir, die Ärzte empfinden mit dem Patienten*

3. Etwas empfinden als etwas (= etwas wahrnehmen als etwas):
   - *als Bedrohung / als Fremdkörper / als Herausforderung empfinden.*
4. Etwas empfinden als irgendwie (Adj, Atr)
   - *etwas als (un)gerecht / als (un)angenehm / als peinlich empfinden*
   - Etwas mit jmdm. (zusammen) empfinden (wahrnehmen, ergreifen*): Das Kind empfindet den Alltag mit dem behinderten Geschwister anstrengend*
5. Jemanden/etwas empfinden für (jmdn. finden, halten für): *ich empfand dich für reifer, u.a.*
6. Für jemanden irgendwie empfinden (ein Gefühl haben*): Gott empfindet für uns wie ein Papa.*

Schlussfolgerung

Die Verben der inneren Wahrnehmung bilden eine relativ kleine Gruppe der Wahrnehmungsverben, die in der Linguistik noch nicht genügend erforscht wurden. In diesem Beitrag stellten wir uns das Ziel, die Kombinationsmöglichkeiten der Wahrnehmungsverben *empfinden* und *fühlen* zu untersuchen. Als wichtigstes syntaktisches Charakteristikum dieser Verben hat sich ihre Transitivität angezeigt, was sich auch in einer großen Zahl von Kookkurrenzpartnern in Akkusativ wiederspiegelte. Um ein komplexes Bild über die weiteren Bindemöglichkeiten beider Verben zu bekommen, hielten wir es für wichtig, weitere binäre Strukturen und Kombinationsmöglichkeiten – wie Attribute - trotz ihrer niedrigeren Frequenz zu untersuchen.

Der Beitrag könnte als Anreiz für eine weitere Forschung sein, da er viele Fragen öffnet, die noch auf eine Untersuchung warten.

Literatur

BISCHOF, B.-B., 2007. *Französische Kollokationen diachron. Eine korpusbasierte Analyse*. Dissertation. Institut für Linguistik/Romanistik der Universität Stuttgart

CIESZKOWSKI, M., 2003. Kollokationen im Schnittpunkt zwischen Wort, Satz und Text. In: Cieszkowski, M./Szczepaniak, M. (Hrsg.): *Texte im Wandel der Zeit. Beiträge zur modernen Textwissenschaft. 1*. Aufl. Frankfurt am Main: Peter Lang. S. 39-54.

Cyril Belica: Kookkurrenzdatenbank: www.corpora.ids-mannheim.de/ccdb/
Das digitale WB der deutschen Sprache des 20. Jhs.: www.dwds.de
Der Leipziger Wortschatz- Lexikon: www. wortschatz.uni-leipzig.de
DeTenTen13: /www.sketchengine.co.uk

ĎURČO, P., 2008. Zum Konzept eines zweisprachigen Kollokationswörterbuchs. Prinzipien der Erstellung am Beispiel Deutsch – Slowakisch. In: Franz Joseph Hausmann (Hrsg.): Collocations in European lexicography and dictionary research. *Lexicographica*, Tübingen : Max Niemeyer Verlag 2008, s. 69-89.

ĎURČO, P., 2007. *Zásady spracovania slovníka kolokácií slovenského jazyka*. [online] Erreichbar: http://vronk.net/wicol/images/Zasady.pdf. [10. März 2015].

ĎURČO, P./BANÁŠOVÁ, M./HANZLÍČLOVÁ, A., 2010. *Feste Wortverbindungen im Kontrast.* Trnava*:* Univerzita sv. Cyrila a Metoda.

Elexiko: www. owid.de

HANKS, P., 2011. Wie man aus Wörtern Bedeutungen macht: Semantische Typen treffen Valenz. In: Engelberg,S./Holler, A./Proost, K. (Hrsg.) 2011. Berlin/Boston: de Gruyter. *Sprachliches Wissen zwischen Lexikon und Grammatik*, S. 483-505

HAUSMANN, F. J., 1984. Wortschatzlernen ist Kollokationslernen. In: (Hrsg.) Freudenstein,W./Jäger, J., 1984. *Praxis des neusprachlichen Unterrichts* 31. S. 395- 406

KONECNY, Ch., 2010. *Versuch einer semantisch-begrifflichen Annäherung und Klassifizierung anhand italienischer Beispiele*, Frankfurt: Lang Peter.

KONECNY, Ch., 2010. *Versuch einer semantisch-begrifflichen Annäherung und Klassifizierung anhand italienischer Beispiele*382382382382382晒晔晆喃暘曜輮朩栶槓橘檜欺殨毖昆汔汢沘沔沚沛沠沢
 铠鉞粎涎淲淲切壽貜奀<ⲁ 흜382j382哛Ĉj382ȣ 함갃哛ĈƧ˪ 鈁ȣ˪ 鈁

CONTRASTIVE ANALYSIS OF VERB-NOUN COLLOCATIONS OF 'UTTERANCE' IN FRENCH AND KOREAN

Sunock Shin
Laboratory LDI
Paris 13 University
sunockshin@gmail.com

Pierre-André Buvet
Laboratory LDI
Paris 13 University
Pierreandre.buvet@gmail.com

Abstract

This study aims to propose a method of investigating the collocation in line with contrastive approach whose main target is consisted of 'speech act' related collocations occurred in French and Korean. So, we could tell that our research have double interests in the collocation, especially collocations of 'utterance'.

Firstly we investigate the collocation through 'lexical selective combination' by which one can detect it with 'restriction of lexicon selection' and 'semantic transparency'. The support verbs point of view is proved particularly effective to our designated pattern "verb-noun".

Secondly we divided three types of contrastive correspondences between French and Korean collocations: [1] collocation *vs* collocation; [2] collocation *vs* one lexical item; [3] collocations *vs* idiom.

1. INTRODUCTION

This study aims to present a method to select and construct a collocation list of 'utterance' related expressions in French corresponding to Korean equivalent, within the framework of contrastive approach.

Our initial question to start was this: how could the translational asymmetries between French and Korean to be treated? To answer this question, we propose a systematic approach to the collocation. What is the collocation then? It is a linguistic expressions to manifest the conventional aspect of a country where the language in question is in use. For example, let's look at the expression 'play a musical instrument'; they would usually say it in French like this: *jouer (du piano + du violon + de la trompette)* – in these examples, we can see the one verb *jouer* singlehandedly covers the entire musical instruments mentioned here. However, it is not the case in Korean because each musical instrument needs a different verb to be used:

(1a) piano-leul chin-da (lit. piano-ACC hit-DECL) 'to play the piano'

(1b) baiollin-eul kyeo-da (lit. violin-ACC flick-DECL) 'to play the violin'

(1c) teuleompes-eul bul-da (lit. Trumpet-ACC blow-DECL) 'to play the trumpet'

Of course, we do not deny that there are also considerable amounts of cases where several French verbs are needed while they just correspond to one Korean verb. This questioning could be helpful to detect where an error could be occurred by, for example, a word-to-word translation effort without considering the characteristics inherent between the source and target language.

Our study will go following this step: firstly, it defines the predicative nouns (Npred) and support verbs (Vsup) and how the support verbs constructions (SVC) are being made within the framework of 'Three Primary Functions (MEJRI 2009)'. Secondly, it explicates the collocation by the 'fixation' point of view. Thirdly and lastly, it compares the typological patterns of "verb-noun[105]" between French and Korean and how their asymmetry problems could be solved.

2. SUPPORT VERBS CONSTRUCTIONS

2.1. Predicative nouns and support verbs

First of all, we introduce 'Three Primary Function' a framework employed in this study to explicate the Npred and Vsup. This theoretical framework takes the lexicon as the main object following Zellig Harris's proposals. It divides the functions of linguistic units as the predicate (function of predicate), the argument (function of argument), the actualisator (function of actualisator). The Primary Functions permit us to categorize the linguistic units according to the syntactical-semantic plan and to explain their roles in the course of speech act. The first two functions realize the content of a speech: the function of predicate puts the light into the structural element; the function of argument conditions which elements are to be subordinated to it. The function of actualisator states the elements of the instantiation of predicate-argument structure in the utterance (BUVET 2013).

For example, the sentence *Ce garçon aimait son père* has the verb root *aim-* analyzed as a predicate, the nouns *garçon* and *père* as arguments, and the determiners *ce* and *son* and the verb ending *-ait* as actualisator.

We now take into consideration two simple sentences including *prendre*. Let's think about whether or not these two sentences could share the same predicate-argument structures.

[105] "Verb-noun" in French correspond to "noun-verb" in Korean, because French is a SVO/prepositional language while Korean is a SOV/postpositional language.

(2) Paul a pris une décision.

(3) Paul a pris le bus.

The sentence (2) has the predicate *décision* taking the human argument *Paul*. The example (3) has the predicate *prendre* carrying double arguments *Paul* (Nhumain) and *bus* (Ntransport). We can see that (4) takes *décision*, not *prendre* a kind of syntactical transformation.

(4) La décision de Paul

Which status can *prendre* have in (2)? Being unable to select the argument, it functions only as a role of 'support' to support 'actualization' of the grammatical categories like tenses, persons and numbers by filling the syntactical predicative position which permits a construction where the predicative noun becomes primordial: hence we call this kind of verb as 'Vsup'.

We now come to see that the Npred always brings together the Vsup when it constitutes a simple sentence thanks to the series of its characteristics we have investigated above. Therefore this simple sentence can be termed as 'SVC' due to the combination between its own arguments and Vsup. During this combination, the predicative noun selects the semantic arguments and the latter is syntactically being realized in a complement position. That realization seems to make less evident the surface difference, especially syntactically, does exist between them because a SVC takes a verb its predicate, not a noun as in case of generic verb construction.

2.2. Types of support verbs: "Vsup-Npred"

We would now investigate an example "Vsup-Npred", one of many types of Vsup. We have chosen this Vsup for effectiveness without being too exhaustive for a brief presentation. We propose three types of Vsup. They are general, aspectual, transferring support verbs. The first type is divided into *faire, avoir* and stylistic variants.

A. Generic Support Verbs

Faire/ha-da is one of several examples in 'Generic Vsup', as we can see when a considerable number of Npred select *faire/ha-da*, one of the most frequently selected generic Vsup. However, other generic Vsup could also be observed to be combination, a phenomenon we might interpret as a result of stylistic variation of *faire*. Other than *faire*, the verbs like *donner, prendre etc.* fall under this type which is able to bring together a fairely large range of Npred, for exemple:

(5) (*faire+donner) une gifle 'to slap a person's face'

(6) (*faire+commettre) un crime ‘to commit a crime’

Now let’s look at the stylistic variations in Korean. They fairly are similar to French correspondences:

(7) daehwa-leul (ha-da + gaj-da + nanu-da)

lit. conversation-ACC (do+DECL + have-DECL+ swap -DECL)

‘to talk with’

A Npred *daehwa* is not only combined with *ha-da* but with *gaj-da* or *nanu-da*.

B. Aspectual Support Verbs

‘Aspectual Vsup’ are one of the type of Vsup which has the aspectual information: iterative, intensive, inchoative, progressive, terminative and telic aspects, for instance.

(8) Inchoative

a. entamer un discours ‘to begin a speech’

b. yeonsel-eul sijagha-da (lit. speech-ACC begin-DECL)

(9) Intensive

a. assener des injures ‘to hurl the insults at’

b. yogseol-eul peobus-da (lit. insult-ACC hurl-DECL)

C. Transferring Support Verbs

Transferring Support Verbs generally are ‘exchanging concept’. Then what is ‘exchange’? According to Maurice Gross in his 1978 study, it is a formal description of an intuitive exchange into a semantic predicate, namely, transfer. This predicate manifests the transferring act between two peoples, a giver and a receiver. Also it represents a more general concept like the transfer of a place’s object into another.

(10a) (offrir, recevoir) de l'aide 'to (give, receive) a help'

(10b) doum-eul (ju-da, bad-da) (lit. help-ACC (give-DECL, receive-DECL))

In the next chapter we will discuss about how the collocation comes to have a correlation with SVC we have studied above.

3. COLLOCATIONS AND SUPPORT VERBS CONSTRUCTIONS

We have studied a SVC because there is a correlation with the collocation. Here is the example: the sentence *Max a donné une gifle à Luc* is not interchangeable with *faire une gifle* or *offrir une gifle* due to the fact that a Npred *gifle* has selected a Vsup are being realized lexically by an internally selective correlation which is difficult to be predicated externally. This is also analogous to characteristics of the collocation, because the 'lexical collocation' is constructed due to selective correlation between two lexemes on a grammatically 'base construction'. Then we found those characteristics being built around the fact the base has a semantic autonomy while the collocation, which is not autonomic, is able to do it only in carrying with base.

Then what characteristics of collocation can allow this construction? We will now find it out.

3.1. What is the collocation?

In this study which adapted the 'fixation' point of view, to have a clearer understanding of lexical collocations calls of for a distinction between 'free expressions' and 'conventional expressions' in the first place. Free expressions and conventional expressions are usually distinguished from each other whether their degrees of being composed of explicative co-occurrences are high (the former, also termed as 'free combination') or low (the latter, also called as 'non-free combination'). These elements manifest themselves the arguments in which the conventional expression coexists with the collocation and then the collocation is defined as a middle ground between free expression and conventional expression. By virtue of it, the collocation proceeds from the phraseology as it concerns "a subcategory from variable degrees of fixation possessing units of lexicon which are being constructed by specific contexts and, because of this, manifest a discourse type" (NEUVEU 2011). In addition to this theory, MEL'CUK (2003) proposed a 'phraseme' or a 'phraseological unit' – a noun phrase that cannot be constructed regularly and freely from a given informational content. There are three distinctive types of phrasemes: (i) complete phrasemes; (ii) semi-phrasemes; (iii) quasi-phrasemes. Collocations correspond to semi-phrasemes; for example, an expression BA, formed with lexeme A and B, is a collocation because, in uttering this expression, a speaker freely selects A according to A's own meaning on one hand and, on the other hand she selects B to express a meaning C in accordance with A.

The collocation is made of two elements: one is called 'base' as semantically transparent core element; the other 'collocate' whose transparency is not quite guaranteed as the former. For example, the sentence *donner un avertissement* has a semantically transparent base

avertissement and a collocate *donner* which is not semantically transparent due to the irrelevance to its original meaning.

3.2. Criteria of collocations

3.2.1. Restriction of lexicon selection

When we say 'restriction of the lexicon selection', it would summon us to think about firstly 'who selects what' during the selective combination of lexicons we had investigated above and secondly about which conditions have to be imposed in it. But one thing has to be borne in mind: this restriction is not semantic and syntactic but rather lexical. For example, while a transitive verb *donner* itself takes a concrete being as object, however in the collocation context, it is not the transitive verb who chooses the argument by a lexical correlation *donner un ordre* the restriction makes *ordre* select *donner* at the expense of *offrir* and *fournir* while donner is not able to select *ordre*.

We now propose to distinguish a collocate from a non-collocate (IM 2002) a collocation test which is consisted of synonymic substitution and antonymic substitution. Here are examples in Korean:

(11a) gyeoljeong-eul naeli-da, myeonglyeong-eul naeli-da

lit. decision-ACC , order-ACC

'to make a decision, to give an order'

(11b) *gyeoljeong-eul hagang-siki-da

lit. decision-ACC descend-make-DECL

(11c) *gyeoljeong-eul olli-da

lit. decision-ACC lift-DECL

(12a) jim-eul naelida, gisbal-eul naeli-da

lit. luggage-ACC let down-DECL, flag-ACC let down-DECL

'to send a luggage down, to lower a flag'

(12b) jim-eul hagang-siki-da

lit. luggage-ACC descend-make-DECL

'to make descend a luggage'

(12c) jim-eul olli-da

lit. luggage-ACC lift-DECL

'to lift a luggage'

While (11a) might inspire the idea that *gyeoljeong-eul naeli-da* be lexical due to its substitution with *myeonglyeong-eul naeli-da*. Also, as we can see in (11b, c), *naeli-da* in the expression (11a) cannot be substituted neither with *hagang-siki-da*, its synonym, nor with *olli-da*, its antonym. We now could say the expression is substituted neither nor antonymically, and that *gyeoljeong-eul naeli-da* be defined as a collocation while (12a) *gyeoljeong-eul naeli-da* as a free combination.

3.2.2. Semantic Transparency of the base noun

Unlike the usual method of determining the semantic transparency according to the transfer degree of original meaning of a lexical and the degree of this transferred transparency within a new context, we have selected to determine the semantic transparency by dividing a collocation into a 'base' and a 'collocate' and then by categorizing them each as 'transparent meaning' and 'non-transparent meaning'. There are three type of it as below:

[1] Transparent Meaning + Transparent Meaning

a. manger (du pain, un gâteau, de la salade) 'to have (a bread, a cake, a salad)'

b. (bab-eul, gwaja-leul, yag-eul) meog-da

lit. (rice-ACC, biscuit-ACC, medical pill-ACC) eat-DECL

'to eat (a rice, a biscuit, a medical pill)'

[2] Transparent Meaning + Non-Transparent Meaning

c. manger (du regard, de baisers) 'to stare at, to give a kiss'

d. (yog-eul, nail-eul) meog-da

lit. (insult-ACC, age-ACC) eat-DECL

'to receive an insult, to have an age'

[3] Non-Transparent Meaning

e. manger (la grenouille, son chapeau) 'to waste the fortune, to eat one's hat'

f. (kongbab-eul, miyeoggug-eul) meog-da

lit. (rice bean bowl-ACC, seaweed soup-ACC) eat-DECL

'to be in jail', to be failed'

Among these three types, [2] is the collocation, [3] the idiom. But, [1] being said as a free combination, as well as we witnessed the fact that support verb constructions might be considered as a collocation and that SVC might be considered as a collocation and that *faire* is the verb whose frequency makes it as the most representatives Generic Vsup, we ask ourselves this question: how can the Vsup groups below which carry *faire* be termed when we adopt the semantic transparency point of view?

These constructions will not permit us to interpret with [2] type. Or, with semantic transparency point of view, their collocate *faire* emptying predicative lexical semantics, their presence solely depend on their virtue of actualizing the predicative noun. This characteristic convinces us that they are the collocation.

Now, thanks to the restriction of lexicon selection and transparency, we could define the base of collocation as 'semantically transparent being' and the collocate as a combination which is selective about the semantic combination whose semantic transparency (or 'emptying') could be manifested or not.

4. CONTRASTIVE ANALYSIS

4.1. Typology of French-Korean collocation

We now come to set a list of French-Korean correspondence of collocation whose method of being defined and selected had been discussed above. We propose four collocation corresponding patterns between French and Korean as following:

Type A: collocation *vs* collocation

Type B: collocation *vs* one lexical item

Type C: collocation *vs* idiom

Now let us investigate the way how these types actually work by observing their respective examples.

Type A is "Vsup-Npred" that French collocation corresponds typologically to Korean equivalent. However, it is admitted that while there are numerous case to which this formula could be applied, individual verb itself employed in each case might be remarkably diverse.

Our first example is *faire* type, a sort of representative case in 'Generic Vsup'. *Faire* usually corresponds to *ha-da*, due to their similarity of respective semantic characteristics.

(13) Type A : Generic Vsup – FAIRE

a. faire une assertion *vs* jujang-eul ha-da (lit. assertion-ACC do-DECL)

b. faire un aveu *vs* gobaeg-eul ha-da (lit. confession-ACC do-DECL)

Avoir, a French Vsup to which Korean *gaji-da* (to have) is supposed to correspond, is the not the case because *gaji-da* has a difficulty to be combined with a speech act noun. So, we instead witness *avoir* has a combination with a verb *ha-da*.

(14) Type A : Generic Vsup – AVOIR

a. avoir une dispute *vs* maldatum-eul ha-da (lit. quarrel-ACC do-DECL)

b. avoir une altercation *vs* eonjaeng-eul ha-da (lit. quarrel-ACC do-DECL)

Donner, adresser, exprimer, pousser, etc. are a 'stylistics variation' type employed usually with Npred of utterance. In this case of sentence construction, they select frequently these Vsup to mean a communicative act concerning fundamentally the urtterance Npred: to emit a sound or speech (*pousser, exprimer, etc*).

(15) Type A : Stylistics Variants

a. exprimer ses condoléances *vs* jouil-eul pyoha-da (lit. condolence-ACC express-DECL)

b. pousser un cri *vs* goham-eul jileu-da (lit. yell-ACC let out-DECL)

We also mention that the reciprocal nature of speech act make possible a combination of the argument not only with the stylistic variants type like *donner* but also with transformative Vsup.

Below are the correspondences of a collocation of one language to a lexicon of anther due to the absence of Npred or Vsup in the collocation of the former in the lexicon of the latter.

(16) Type B: collocation *vs* one lexical item

a. faire une réprimande *vs* kkujij-da (lit. scold-DECL)

b. dire des malices *vs* nolli-da (lit. mock-DECL)

This is the case where a collocation in French corresponds to an idiom in Korean. We have to say it's not very common in general.

(17) Type C: collocation *vs* idiom

a. dire (des énormités + des idiotes) *vs* gaesoli-leul chi-da (lit. sound of dog-ACC shout-DECL) 'to talk a load of bollocks'

The collocation in French in (17a) corresponds to an expression in Korean *gaesoli-leul chi-da*, a combination between two nouns *gae* (dog) and *soli* (sound) which mean 'a derogatory expression for a complete utter nonsense'. This expression could be interpreted as a total meaning transfer and therefore a semantic non transparency so that it could be seen as an idiom.

In this chapter we have investigated how the French "Vsup-Npred" collocation formula would behave correspondingly towards Korean. And now, we have selected Type A where the differences of correspondence pattern between the two were manifested more clearly than the other types.

4.2. Difference between the French-Korean correspodences

Let us consider now how French-Korean "Vsup-Npred" formula correspondences differ from each other.

The first difference was observed when they correspond to each on the same formula while on the other hand their Vsup don't share the semantic correspondences:

(18a) avoir une entrevue *vs* hoedam-eul ha-da (lit. talks-ACC do-DECL)

(18b) adresser des felicitations *vs* chugha-leul ha-da (lit. congratulation-ACC do-DECL)

(18c) prononcer une sermon *vs* seolgyo-leul ha-da (lit. sermon-ACC do-DECL)

These examples demonstrate us that the French Vsup *avoir* (to have) *addresser* (to address), *prononcer* (to pronounce) correspond to *ha-da*, but not to *gaji-da*, *bonae-da*, *baleumha-da*, their semantic equivalents.

Secondly, there are patterns found in some French generic Vsup corresponding only to Korean aspectual Vsup.

(19) faire du tapage *vs* somun-i (*ha-da + dol-da + peoji-da) (lit. rumor-NOM (*do-DECL + turn-DECL + spread-DECL) 'rumors spread'

French *faire du tapage* cannot correspond to *ha-da*. Instead, it corresponds to *dol-da* and *peoji-da*. Those two Vsup are inchoative.

Finally, we have witness the Korean collocations containing honorific Vsup, the case we cannot find in French. Korean *deuli-da*, *olli-da* which mean 'give' are one of its typical examples. The honorific Vsup is observed being marked as verbal lexical form when it's realized as grammatical function word like a penultimate verb ending *-si-* or verb ending or lexical forms like *jinji* (meal-HON), *daeg* (house-HON) *jabsu-si-da* (eat-HON), *jumu-si-da* (sleep-HON), *etc* (HONG 1999:146). It could correspond to *ha-da* or *ju-da* in an irregular basis.

(20) faire un exposé vs bogo-leul (ha-da + *ju-da + deuli-da + olli-da) (lit. report-ACC (do-DECL + *give-DECL + give-HON-DECL + give-HON-DECL)

'to report on'

When one combine a speech act, its lexicon is always employed with the other which are conditioned to be connected with the former. This pattern constantly is influenced by a series of selections or peculiarities whose presence manifests as equally as in a dual language, say, translation, so that we need to keep these phenomena in check to understand and use them appropriately.

5. CONCLUSIONS

This study is an investigation about how the collocation correspondence between French and Korean could be made through the lens made of four kinds of typology we had established with the speech predicate and how these two languages could behave differently toward each other – in other words, an asymmetry problem.

This asymmetricality creates one of the main difficulties against which a foreign language learner could meet and it also creates a (re-)production of incorrect or awkward sentences occurred in which, for example, a machine translation might multiply by proposing the erroneous lexical elements for a sentence to be translated. We consider we can avoid this problem by manifesting the collocation information between in the nouns to be translated and their substitution words to be proposed. We would like to have more

future researches about co-occurrence, restriction of lexicon selection, semantic transparency in 'utterance' noun, which will certainly contribute both to refine the electronic dictionary system and the second language acquisition method.

References

(Books)

GIRY-SCHNIDER, J., 1987. *Les prédicats nominaux en français : les phrases simples à verbes supports.* Genève : Droz.

GROSS, M., 1996. *Les expressions figées en français.* Paris : Ophrys.

NEUVEU, F., 2011. *Dictionnaire des sciences du langge.* Paris : Armand Colin.

(Book chapters)

MEJRI, S., 2008. Construction à verbes supports, collocations et locutions verbales. In: P. Mogorron AND S. Mejeri, dir. *Las construcciones verbo-nominales libres y fijas. Aproximacion contrastiva y traductologica.* Université d'Alicante & Université Paris 13. pp.191-202.

MEL'CUK, I., 1996. Lexical functions: a tool for the description of lexical relation in a lexicon in a lexicon. In L. Wanner, ed. *Lexical functions in lexicography and natural language processing.* Amsterdam: John Benjamins. pp.37-102.

MEL'CUK, I., 1998. Collocations and Lexical Functions. In A. P. Cowie, ed. *Phraseology: Theory, Analysis and Applications.* Oxford: Oxford University Press. pp.23-53.

MEL'CUK, I., 2003. Collocations dans le dictionnaire. In Th. Szende (réd.), *Les écarts culturels dans les Dictionnaires bilingues.* Paris : Honoré Champion, pp.19-64.

(Journal articles)

BUVET, P.-A., 2013. Collocation, restriction de selection et predication. *Cahiers de lexicologie*, 102, pp.169-184.

GROSS, M., 1979. Observations on semantic theories. *Theoretical Linguistics*, 5(1), pp.3-17.

HONG, J.S., 1999. Approche lexicale des verbes supports. *Humanities collection of writing*, 41, pp.135-173.

HONG, J.S., 2003. Contribution à l'étude contrastive des adjectifs en français et en coréen. *Humanities collection of writing*, 48, pp.27-54.

IM, H.P., 2002. On the nature of collocation in Korean and its syntactic and semantic properties. *Korean Linguistics*, 39, pp.279-311.

LEE, S.H. AND PARK, S.Y., 2007. How to Design a Multilingual Database of Korean Collocations. *Eoneohag* 48, pp.46-64.

LEE, S.H., IM, H.P. AND HONG, J.S., 2009. Intérêt des classes sémantiques dans la construction d'une base de données en vue de l'étude constrative plurilingue des collocations : la cas de l'étude contrastive coréen-français. *Société Coréenne d'Enseignement de Langue et Littérature Françaises*, 31, pp.197-220.

MEJRI, S., 2009. Le mot: problématique théorique. *Le Français Moderne*, 77(1), pp.68-82.

PARK, M.G., 2005. Analyse contrastive des constructions avec verbes supports en français et en coréen. *Société Coréenne d'Enseignement de Langue et Littérature Françaises*, 20, pp.189-224.

VIVÈS, R., 1984. L'Aspect dans les constructions nominales prédicatives: avoir, prendre, verbe support et extension aspectulle. *Linguisticae Investigationes* 3(1), pp.161-185.

THE PARALLEL POLISH-BULGARIAN-RUSSIAN CORPUS: PROBLEMS AND SOLUTIONS

Wojciech Sosnowski
The Institute of Slavic Studies of
the Polish Academy of Sciences

Abstract

The parallel Polish-Bulgarian-Russian corpus we are currently developing as part of CLARIN-PL framework will become an essential tool for translators producing both traditional and digital translations. The electronic tools developed within the project facilitate fast search for and retrieval of multilingual equivalents of lexemes, phrases and sentences. Selected sentences and texts have been semantically annotated for the quantification of nomen, time and aspect. Our definition of equivalent stems from the contemporary contrastive linguistics theory. The guiding principle in the construction of the corpus was to proceed from meaning to form; the principle was first introduced in Koseska-Toszewa (2006).

During our work on the Polish-Bulgarian-Russian corpus, we have come across a number of issues, which we regard as characteristic of multilingual corpora: (1) the selection and procurement of texts, (2) the development of computer tools used for the construction of the corpus, (3) multilingual equivalence, and (4) semantic annotation.

Multilingual corpora have proved to be exceptionally helpful in language teaching, traditional and digital lexicography, as well as traditional and digital translations. The usefulness of multilingual corpora in each of these areas will be demonstrated through example corpus queries.

1. INTRODUCTION

The parallel corpora we are currently developing as part of CLARIN-PL framework (a Polish-Bulgarian-Russian corpus and a Polish-Lithuanian corpus) will become essential tools for translators producing both traditional and digital translations. In linguistics, parallel corpora will enable provide large amounts of data for the study of language and its evolution. Parallel corpora are also useful in language teaching, sociology, cultural studies as well as other fields related to linguistics and information technology. In the 2000s, many countries developed their national corpora, e.g. Poland (the so-called "one-million" National Corpus of Polish), Bulgaria (Bulgarian National Corpus, http://search.dcl.bas.bg/) and Russia (Russian National Corpus, http://www.ruscorpora.ru). Although the above corpora have proved to be valuable tools for linguists studying these languages in isolation, they were of little use to scholars working in contrastive linguistics, lexicography, translation studies and language teaching.

2. PARALLEL CORPORA IN CLARIN-PL

The Department of Corpus Linguistics and Semantics of the Polish Academy of Sciences has been developing a parallel Polish-Bulgarian-Russian corpus, which is to be incorporated into the CLARIN framework[106]. Our corpus will become the first multilingual corpus of Slavonic languages. It has been our priority to develop a multilingual corpus, because monolingual and even bilingual corpora are inadequate tools for comparative linguistics.

The European Union aims to make its ubiquitous digital market truly multilingual. The ubiquitous digital market strategy must address all issues that relate to multilingualism in order to ensure that EU offers equal opportunities for speakers of each of EU's official languages. Nevertheless, the language barrier still remains the main barrier to a truly integrated European economy and society. In order to overcome this barrier, we have been working on a number of corpora as part of CLARIN-PL: a Polish-Bulgarian-Russian corpus and a Polish-Lithuanian corpus. These corpora will bridge the gap in Slavonic and Balto-Slavonic digital linguistic resources and will help provide accurate translations of digital and conventional texts.

As soon as we began our work on the parallel corpora, a number of problems emerged that were specific to multilingual corpora. The remainder of this section will give an overview of the issues we encountered and the solutions that we chose to address them.

2.1. Selecting the languages

We have chosen Polish, Bulgarian and Russian because they are representative of the West, South and East Slavonic group respectively. The languages exhibit different structures: synthetic (Polish and Russian) and analytic (Bulgarian). They also employ different writing systems: the Latin script (Polish) and the Cyrillic script (Russian and Bulgarian).

2.2. Selecting the texts

The first version of the corpus will contain 6 million words — 2 million words for each language. We plan to add another 2 million in the second stage of the project. The aim of the planned expansion (the second stage) is to make our corpus a medium-sized corpus, which will enable researchers to conduct novel types of studies with the use of the corpus. The selection of texts for the corpus was based on the following criteria:

1) Texts must come from different styles, genres and registers (general language, languages for special purposes)

2) Texts must come from different sources (the original text in one of the corpus's languages or a translation from a different language into all three languages of the corpus)

3) Texts must come from different historical periods

4) Every text must exhibit a high standard of language correctness (critically acclaimed translations, canonical literary texts)

---

106 Common Language Resources and Technology Infrastructure is a project granted the status of ERIC (European Research Infrastructure Consortium) by the European Commission in February, 2012. CLARIN was founded by eight countries: Austria, Bulgaria, the Czech Republic, Denmark, Estonia, Germany, the Netherlands and Poland. CLARIN is part of the ESFRI (European Roadmap for Research Infrastructures, European Strategy Forum on Research Infrastructures). The project's primary aim is to combine language tools and resources for multiple European languages into one unified network, which will become an important research tool for scholars in arts, humanities and social sciences.

Eventually, we have included multiple text genres in the corpus: literary texts from the 19th, 20th and 21st century, instruction manuals and technical documentation, legal texts, as well as other types of documents. The table below presents some example texts included in the corpus:

Name	Word count
Additional Protocol to the European Convention on Mutual Assistance in Criminal Matters	3371
Antoine de Saint-Exupery, *The Little Prince*	35228
European Convention on Transfrontier Television	14621
Amendments to the European Convention on Transfrontier Television	14328
European Convention the Archaelogical Heritage	6843
Convention on the Protection of Children against Sexual Exploitation and Sexual Abuse	21749
Convention on the Recognition of Qualifications concerning Higher Education in the European Region	14937
Council of Europe Convention on Action against Trafficking in Human Beings	21563
Statute of the Council of Europe	8330
Paulo Coelho, *Eleven Minutes*	18946
Statute of the Council of Europe	34613
European Convention on Recognition and Enforcement of Decisions concerning Custody of Children and on Restoration of Custody of Children	9690
Universal Declaration of Human Rights	4442
European Cultural Convention	2859
Council of Europe Convention on preventing and combating violence against women and domestic violence	29668
European Convention for the Prevention of Torture and Inhuman or Degrading Treatment or Punishment	6481
European Convention on Mutual Assistance in Criminal Matters	8884
Additional Protocol to the European Convention on Extradition	2964
Joseph Conrad, *Lord Jim*	92307
Stefan Żeromski, *Ashes*	68132
Angel Wagenstein, *Far From Toledo*	184 421
Kyoto Protocol	21 257
Paulo Coelho, *The Alchemist*	17 636
Alexandre Dumas, *The Count of Monte Cristo*	417 620
Paulo Coelho, *The Witch of Portobello*	20 096

As we can see, beside literary texts the corpus also incorporates a large number of documents produced by international institutions, e.g. Council of Europe treaties and official EU documents.

2.3. Obtaining texts

The texts in the corpus come from three sources: (1) open source publications; (2) copyrighted documents for which we have obtained licenses[107] and (3) public domain texts (i.e. texts whose intellectual property rights have expired or have been forfeited). The search engine in the final version of the corpus will only display as much text as it is allowed by the right to quote. Every text will be annotated with metadata, which will also be displayed by the search engine. Yet another problem that we encountered while working on the corpus was that some texts have not yet been converted to an electronic format and therefore we had to digitise them manually. So as to obtain the most accurate version possible, after every phase of digitising a text was proofread and edited.

2.4. Developing the corpus with computer tools

The first step in developing the corpus was to choose a computer application that would enable us to align three languages in parallel. When we began the work on our corpus, it became clear that there was no application that would allow us to split large texts in three different languages in parallel. Eventually, we decided to use NOVA Text Aligner. NOVA Text Aligner is a tool designed to make manual text alignment as easy and simple as possible. There are automated paragraph/sentence alignment tools but there is one thing that they all have in common – they are not 100% accurate (and they can not be due to the nature of the task they are supposed to do). So this means that in the end you'll have to go trough the whole text yourself and check it and correct it (http://www.supernova-soft.com/wpsite/products/text-aligner/). First, we would align Polish and Bulgarian texts and afterwards we would supplement them with the third language. While aligning the texts, we found that the sentence-level equivalence was very difficult to achieve.

3. MULTILINGUAL EQUIVALENTS IN CONTRASTIVE STUDIES

The definition of equivalence that we follow in our research derives from the contemporary semantic theory and contrastive studies of natural languages developed in the multi-volume *Gramatyka konfrontatywna bułgarsko-polska* [further referred to as: GKBP] (Koseska-Toszewa and Gargov, 1990; Koseska-Toszewa, 2006; Koseska-Toszewa, Korytkowska and Roszko, 2007). GKBP is the first contrastive grammar in the world that makes use of an intermediate semantic interlanguage. Using a semantic interlanguage to compare multiple languages provides an innovative solution for contrastive studies and diverges from traditional principles of applied contrastive studies. Traditionally, the comparison between two (or more) languages relied heavily on the primary language of description. In consequence, it was always incomplete and could also be misleading, if not grossly inaccurate.

In theoretical contrastive studies, the analysis of language data proceeds from meaning to form. This stands in contrast to traditional contrastive grammars, which tend to depart from a form in one language and then proceed to a form in another language. The above procedure – outlined in GKBP – enabled us to treat the data from every language as equal.

[107] The development of a model licence agreement took approximately one year.

Equivalence or the lack of equivalence is a widely debated phenomenon in linguistics:

> Equivalence (or lack thereof) is a marginal phenomenon, if comparative studies take under consideration only one language. The notion of equivalence, on the other hand, plays a crucial role in contrastive lexicology. Accordingly, the notion of equivalence in lexicology concentrates on the language system, therefore it is relatively vague. On the basis of designation lies a polysemic understanding of the linguistic sign. Consequently, an element of the lexicon can have several values, i.e. meanings. When comparing an element from the language A with another element in the language B, generally the denotative relationship is the basis for such a comparison. Thus, there is an equivalence, which is usually called semantic equivalence with the provisos that, firstly, the number of sememes in language A is the same as in language B (and thus they have the same value), and, secondly, their denotation (paired sememes) is the same. (Jaskot, in press).

The table below presents a selection of equivalent sentences that we have encountered:

Agent okrętowy nie potrzebuje zdawać żadnych egzaminów, ale musi posiadać zdolność abstrakcyjnego myślenia i umieć wykazać je w praktyce	На морския кларк не са нужни никакви изпити, но той трябва да се отличава с абстрактно умение и да го проявява на практика.	Морскому клерку не нужно сдавать никаких экзаменов, но предполагается, что он должен отличаться сноровкой и проявляет ее на практике.

Joseph Conrad, *Lord Jim*

Tu postanowił spędzić noc. Wprowadził swoje owce przez rozpadającą się bramę i zagrodził wejście deskami tak, by w nocy nie mogły się wymknąć.	Реши да пренощува тук. Вкара овцете през разнебитената порта и я залости с няколко дъски така, че да не могат да избягат.	Он решил заночевать там, загнал через обветшавшую дверь своих овец и обломками досок закрыл выход, чтобы стадо не выбралось наружу.

Paulo Coelho, *The Alchemist*

Szum ów dobiegał z dołu, wznosił się ku stropom, ku twarzy prokuratora. A za plecami Piłata, za skrzydłem pałacu grały larum trąbki, słychać było stamtąd ociężały chrzęst setek nóg i pobrzękiwanie żelastwa. Procurator zrozumiał, że to już wymarsz piechoty rzymskiej, która spełniając jego rozkaz, udaje się na straszną dla buntowników i zbójców przedśmiertelną defiladę.	Този шум се надигаше отдолу към нозете и лицето на прокуратора, а зад гърба му, там, отвъд крилата на двореца, се чуваха тревожните сигнали на тръбите, тежкото скърцане на стотици стъпки, звън на желязо. Прокураторът разбра, че римската пехота вече тръгва, съгласно неговата заповед, за страшния за бунтарите и разбойниците предсмъртен парад.	Этот шум поднимался снизу к ногам и в лицо прокуратору. А за спиной у него, там, за крыльями дворца, слышались тревожные трубные сигналы, тяжкий хруст сотен ног, железное бряцание, – тут прокуратор понял, что римская пехота уже выходит, согласно его приказу, стремясь на страшный для бунтовщиков и разбойников предсмертный парад.

Mikhail Bulgakov, *The Master and Margarita*

4. SEMANTIC ANNOTATION

We are currently working on the semantic annotation of 2000 sentences in the parallel Polish-Bulgarian-Russian corpus and the Polish-Lithuanian corpus. The preliminary annotation (i.e. the 2000 sentences we are working on at the moment) must be performed manually. Once it is completed, it will serve as a basis of an automatic tagger. Koseska-Toszewa & Roszko (2015) developed an innovative semantic annotation scheme, which can be applied to entire sentences in multilingual parallel dictionaries. Instead of choosing a number of separate sentences and annotating them, we will annotate longer fragments of texts. The semantic annotation scheme outlined in Koseska-Toszewa & Roszko (2015) will

make contrastive studies of natural languages easier and, in consequence, facilitate more efficient manual and automatic translations. Below, I will present some examples of the semantic annotation scheme at work.

5. POSSIBLE IMPLEMENTATIONS OF THE CORPUS

Our corpus constitutes a comprehensive resource for scholars developing on multi-lingual dictionaries or conducting studies comparing multiple languages (e.g. Bulgarian-Polish and Russian-Polish contrastive grammar, Polish-Lithuanian contrastive grammar, Balto-Slavonic contrastive studies. The target group of our corpora are linguistics; stylisticians; translators (e.g. to investigate translation strategies employed in the works available in the corpus); scholars (e.g. for studies on terminologies and lexical equivalence); students of literary studies (for comparative research), cultural studies (to study the forms culturemes take in different languages), sociology (the texts we included in our corpus are a reflection of the social processes that took place in their respective periods), political studies, history, intercultural communication or anthropology.

- Searching for equivalents necessary for synchronic contrastive studies.

- Developing translation memories (TMs) based on contemporary lexis; these translation memories can be later incorporated into translators' own translation memories. Translation memories should be developed in a widely recognised format (e.g. TMX), which can be imported into the most popular CAT application suites. TMs significantly reduce the amount of time and labour translators and teachers have to spend on their tasks. More importantly, they also enable automated database search, which ensures high stylistic and terminological coherence of texts produced by translators and teachers.

- Quantitative studies (frequencies of word types and tokens as well as syntactic structures and contexts they appear in).

- Data necessary for the construction of grammatical models of languages.

- Research on intercomprehension[108]: it provides data for the construction of exercises that aim at the activation of the passive knowledge of cognate languages. In the area of Slavonic languages, exercises of this type are quite an innovation; they are of paramount importance, especially taking into consideration the fact that Slavonic languages form a significant part of the linguistic landscape in the EU and, what is more, they are quite closely related to each other.

- Investigating translation strategies: comparing the lexical and grammatical constructions in different languages used for the expression of similar semantic content; studying how different languages convey phraseological units, culturemes and non-equivalent lexis; stylistic and terminological coherence, etc.

- Teaching of first and second languages.

- Studies of text-level equivalence of culturemes.

- Quotation search.

- And many more.

[108] Cf. The European Intercomprehension Network REDINTER: http://www.redinter.eu/web/

5.1. Phraseology in multilingual corpora

The process of searching for the equivalents of phraseological units provides a good illustration of how multilingual corpora can be used in language teaching, dictionary development and translation.

Before we could investigate any phraseological units in the corpus, we need to develop a working definition of a phraseological unit. We decided to work with the definition developed by our colleagues from NASU (The National Academy of Sciences of Ukraine), who are currently working on a Polish-Ukrainian phraseological dictionary. They defined phraseological units as follows:

> Phraseological units are distinguished among other types of phrases by their complicated semantics, which is strongly oriented towards national linguistic worldview. Thus, the main problem in compiling of a bilingual phraseological dictionary is the selection of adequate translational equivalents with due account for differences in worldview represented in the respective language systems. This is why the task of a comprehensive translational phraseological dictionary is to convey the phraseological system of one language by the means of the other language". (Tymoshuk, Vilchynska, Shyrokov and Nadutenko, in press)

As we can see, the only way to provide a description of a phraseological unit is through its ontology, because every language expresses phraseological semantic content in a different manner. Most scholars studying the relations between lexemes and phraseological units argue that a semantic and functional correlation exists between them, which is reflected in the organisation of different levels of language systems. No consensus has yet been attained on how we should determine the position of phraseology among other levels of language systems. V. L. Arkhangel'skyi proposed a structural semantic classification. He defined lexemes and phrasemes as different units organised in a hierarchical relationship, however these units are units of the same level that constitute "building blocks" of sentences [3, pp. 182-188]. This apparent incongruence is a result of the great complexity of the semantics of a phraseological unit and of the priority it takes over a word (after M. M. Shanskyi). A phraseological unit takes the form of a free association of words on the phrase level, whereas on the text level it assumes the role of a word [10, p. 12].

It is equally difficult to clearly delineate the dividing line between a phraseme and a non-phraseme. As a consequence, the selection of phrasemes for contrastive studies is always problematic, because one always needs to decide which linguistic tradition to choose as the source of comparison with other languages. The above applies also to the selection of collocations, which we can also categorise as phrasemes:

> Separated into the so-called "rhombed" zone of the Dictionary are also the collocations – set phrases that allow slight desemantization of one component (eg. **вовчий апетит**), word equivalents (eg. **до безмежжя**) and terminological phrases (eg. **топографічна анатомія**). (Tymoshuk, Vilchynska, Shyrokov and Nadutenko, in press)

Scholars studying phraseology must be prepared to face numerous pitfalls. Idioms that appear strikingly similar may actually carry different, sometimes exactly opposite meanings: Pol. *lekarz z bożej łaski* (= a very bad doctor) / Rus. *милости божьей врач* (= a very good

doctor) [lit. doctor of God's grace], *idzie jak krew z nosa* (= very slowly), *кровь из носу* (= immediately) [lit. flows like blood out of a bleeding nose), *owinąć sobie wokół palca kogoś* (= have somebody under one's command), *обвести вокруг пальца* (= lie to someone in a particularly cunning way) [lit. wrap somebody around one's finger].

The parallel Bulgarian-Polish-Russian corpus allows users to search for phraseological units. We must take into consideration, however, that phrasemes can exist in:

1. Only one language

To ludzie bez ducha, bez dumnych snów, bez wzniosłych porywów. A człowiek bez tego to zwykły tchórz, to szmata.	Те нямат дух, те не знаят какво е горди мечти и горди въжделения, а всеки, който не познава нито едното, нито другото — боже мой! — та той е пълен със страхове и опасения!	У них нет мужества, нет гордости, они не умеют сильно желать. А без этого человек гроша ломаного не стоит.

Bronte, E *Wuthering heights*

Myślałam już nawet, że brak jej piątej klepki. Uciekła do swego pokoju wołając mnie do siebie, chociaż powinnam była ubierać dzieci.	Докато траяха тия неща, по държането й разбрах, че е доста глупавичка. Тя се втурна в стаята си и ме застави да отида при нея, макар че в това време трябваше да обличам децата.	Я приняла ее за полоумную, — так она себя вела, пока совершали обряд: она убежала к себе в комнату и велела мне пойти с нею, хотя мне нужно было переодевать детей.

Bronte, E *Wuthering heights*

2. In two languages

Matki i wychowawczynie - nie żadne lalkowate ślicznotki ze słodkimi ślepkami.	Никакви превзети дамички, никакво въртене на очи!	Только не сентиментальные дамы, не те, что строят глазки.

Wells, H. G. *The War of the Worlds*

W mgnieniu oka wdarłem się na wał i stanąłem na jego koronie. Przede mną leżała twierdza.	След още един миг се бях покатерил по земния насип и стоях на гребена му — вътрешността на редута лежеше в краката ми.	Еще через минуту я взобрался по насыпи и стоял на гребне вала – внутренняя площадка редута была внизу, подо мной.

Wells, H. G. *The War of the Worlds*

3. In three languages

Nigdy nie wyznałem swej miłości słowami, ale jeżeli oczy mają wymowę, to każda gąska musiałaby odgadnąć, że byłem zakochany po uszy.	„Не се признах в любов“2 гласно; и все пак, ако очите могат да говорят, дори един идиот би могъл да долови, че съм влюбен до уши.	Я «не позволял своей любви высказаться вслух»; однако, если взгляды могут говорить, и круглый дурак догадался бы, что я по уши влюблен.

Wells, H. G. *The War of the Worlds*

6. CONCLUSIONS

The data presented above shows how many new insights into phraseology multilingual corpora can provide, even though phraseological units usually exist only in one of the languages being compared. The usual situation is that translators only translate the words in phrasemes. The data also indicates that we need to study equivalents of phrasemes departing from their ontology, following the example of scholars from NASU (see 5.1). It is also worth noting that phrasemes in different languages evoke very different associations and mental images, e.g. Bul. *бързата кучка слепи си ражда* [lit. 'the hasty bitch gives birth to blind pups, Pol. *co nagle to po diable* [lit. 'rush is the devil's thing'], Eng. *haste makes waste.*

During our work on the corpus, we encountered a number of different issues. At the same time, it allowed us to find many new solutions and to introduce some innovations. We have learned that multilingual corpora need to be supplemented with more languages. Every language we add to a corpus enables researchers and practitioners to find new questions relevant to translators, interpreters and language teachers.

References:

DIMITROVA, L., KOSESKA-TOSZEWA, V., 2012. Bulgarian-Polish parallel digital corpus and quantification of time. *Cognitive Studies/Études cognitives*, 12, pp. 199–208.

GARABÍK, R., DIMITROVA, L., KOSESKA-TOSZEWA, V., 2011. Web-presentation of bilingual corpora (Slovak-Bulgarian and Bulgarian-Polish). *Cognitive Studies/Études cognitives*, 11, pp. 227–239.

GRAMATYKA KONFRONTATYWNA BUŁGARSKO-POLSKA BPCG [GKBP]., 1988–2007. (Vol. 1-12). Sofia, Warsaw.

JASKOT, M. P., 2014. Buscando las brechas de significado: las lagunas léxicas entre el español y el polaco In: Zuzanna Bułat Silva, Monika Głowicka and Justyna Wesoła [eds.] *Variación, contraste,*

circulación. Perspectivas lingüísticas en el hispanismo actual. Acta Universitatis Wratislaviensis. Wrocław, Wydawnictwo Uniwersytetu Wrocławskiego, pp. 127-136.

JASKOT, M. P., (in press). *Lexical non-equivalence in chosen European languages in the context of the policy towards multilingualism in Europe*

KISIEL, A. (in press). Korpusowe badania nad metatekstem. Problem homografii, *Prace Filologiczne*.

KISIEL, A., SATOŁA-STAŚKOWIAK, J., SOSNOWSKI, W., 2014. О работе над многоязычным словарём. *Прикладна лінгвістика та лінгвістичні технології (MEGALING-2013),* pp. 111–121.

KORPUS JĘZYKA BUŁGARSKIEGO IBE BAN, n.d. Available at: http://search.dcl.bas.bg/ [Accessed 6 November 2014].

KORPUS JĘZYKA ROSYJSKIEGO, n.d. Available at: <http://www.ruscorpora.ru/> [Accessed 6 November 2014].

KOSESKA-TOSZEWA, V., 1974. Z problematyki temporalno-aspektowej w języku bułgarskim (Relacja imperfectum - aoryst). *Studia z Filologii Polskiej i Słowiańskiej*, 14, pp. 213–226.

KOSESKA-TOSZEWA, V., 2006. *Gramatyka konfrontatywna bułgarsko-polska* (T. 7: *Semantyczna kategoria czasu*). Warsaw: SOW.

KOSESKA-TOSZEWA, V., GARGOV, G., 1990. *Bylgarsko-polska sypostavitelna gramatika* (T. 2: *Semanticznata kategorija opredelenost/neopredelenost*). Sofia: BAN.

KOSESKA-TOSZEWA, V., MAZURKIEWICZ, A., 1988. Net representation of sentences in natural languages. In: *Advances in Petri Nets. Lecture Notes in Computer Science*, 340*,* pp. 249–266. Berlin: Springer-Verlag.

KOSESKA-TOSZEWA, V., MAZURKIEWICZ, A., 2010. *Time flow and tenses*. Warsaw: SOW.

KOSESKA-TOSZEWA, V. ROSZKO, R., (2015. On Semantic Annotation in CLARIN-PL Parallel Corpora. *Cognitive Studies/Études cognitives*, 15

KOSESKA-TOSZEWA, V., KORYTKOWSKA, M., ROSZKO, R., 2007. *Polsko-bułgarska gramatyka konfrontatywna.* Warsaw: Wydawnictwo Akademickie Dialog.

KOSESKA-TOSZEWA, V., SATOŁA-STAŚKOWIAK, J., SOSNOWSKI, W., 2013. From the problems of dictionaries and multi-lingual corpora. *Cognitive Studies/Études cognitives*, 13, pp. 113–122.

KOSESKA-TOSZEWA, V., SATOŁA-STAŚKOWIAK, J., SOSNOWSKI, W., 2013. О работе над книжными и электронными словарями с польским, болгарским и русским языками. W *Прикладна лінгвістика та лінгвістичні технології (MEGALING-2012)*, pp. 124–135.

ROSZKO, D., 2015. Zagadnienia kwantyfikacyjne i modalne w litewskiej gwarze puńskiej (Na tle literackich języków polskiego i litewskiego). Warsaw: Instytut Slawistyki PAN

ROSZKO, D., 2013. Experimental Corpus of the Lithuanian Local Dialect of Punsk in Poland. Examples of the Lexical and Semantic Annotation. *Cognitive Studies/Études cognitives*, 13, pp. 79–95. DOI: 10.11649/cs.2013.005.

SATOŁA-STAŚKOWIAK, J. 2013, Contemporary Contrastive Studies of Polish, Bulgarian and Russian Neologisms versus Language Corpora, *Cognitive Studies/Études Cognitives,* 13, pp. 143-160.

SATOŁA-STAŚKOWIAK, J. 2014, *Edukacja przyszłych tłumaczy w oparciu o korpusy językowe.* In: Прікладна лінгвістика та лінгвістичні технології, MegaLing-2013: Зб. наук. пр. / НАН України, Укр. мовно-інформ. фонд, Київ, pp.. 211-223.

SATOŁA-STAŚKOWIAK, J., KOSESKA-TOSZEWA, V. 2014, *Współczesny słownik bułgarsko-polski*, Warsaw: Slawistyczny Ośrodek Wydawniczy.

SOSNOWSKI, W., 2013. Forms of address and their meaning in contrast in Polish and Russian languages. *Cognitive Studies/Études cognitives*, 13, pp. 225–235.

TYMOSHUK, R., VILCHYNSKA, K., SHYROKOV, V. and NADUTENKO, M., (in press). *Semantic interpretation of phraseological units in ukrainian-polish electronic phraseological dictionary*

АНТОНОВА, О., ДУБРОВСЬКА, І., and ЛУЧИК, А., 2011. *Українсько-польський словник еквівалентів слова.* (V. Koseska-Toszewa & A. Kisiel, Ed.). Київ: Український комітет славістів, Український мовно-інформаційний фонд НАН України, Національний Університет «Києво-могилянська Академія», Інститут Славістики Польської Академії Наук.

ЛУЧИК, А., АНТОНОВА, О., 2012. *Польсько-український словник еквівалентів слова.* (A. Kisiel & V. Koseska-Toszewa, Ed.). Київ: Український мовно-інформаційний фонд НАН України, Національний Університет «Києво-могилянська Академія», Інститут Славістики Польської Академії Наук.

ШАНСКИЙ Н.М., 1957. Лексика и фразеология современного русского языка: пособие для студентов-заочников факультетов русского языка и литературы педагогических институтов.

APRENDER FRASEOLOGÍA MEDIANTE CORPUS: UN CASO APLICADO A LA ENSEÑANZA DEL ALEMÁN

María Rosario
Bautista Zambrana
Universidad de Málaga
mrbautista@uma.es

Resumen

El presente trabajo da cuenta de una experiencia realizada en la asignatura Idioma Moderno II (Alemán), de nivel A1.2, perteneciente al título de Graduado en Estudios Ingleses. Partiendo de la premisa de la utilidad del corpus para la enseñanza-aprendizaje de lenguas extranjeras, hemos diseñado una serie de ejercicios basados en corpus para el aprendizaje de elementos gramaticales y léxicos de la lengua alemana; dentro del componente léxico nos hemos centrado en el aprendizaje de unidades fraseológicas, especialmente colocaciones y *lexical bundles*. Previamente al trabajo en el aula, compilamos un corpus acerca del tema elegido, *Orientierung in der Stadt* ("orientación en la ciudad"), con textos basados en la descripción del camino para llegar a un sitio. Siguiendo las recomendaciones de Leal Riol (2013), tratamos de que el corpus fuera un fiel reflejo del alemán en uso y que ofreciera muestras lingüísticas adecuadas. A continuación, seleccionamos las unidades fraseológicas con las que íbamos a trabajar, basándonos en criterios de frecuencia, facilidad, productividad y necesidad (Leal Riol, 2013). El siguiente paso fue utilizar el corpus en clase con los alumnos. Primero, se les impartió una breve introducción del concepto de corpus y de su utilidad para el aprendizaje de lenguas; a continuación, se les enseñaron nociones básicas de utilización del programa de análisis de corpus *AntConc*. Con esta base procedimos a trabajar en clase con el corpus preparado: en primer lugar, se realizó un ejercicio de extracción de las palabras más frecuentes, a fin de que los alumnos se familiarizasen con las palabras más comunes usadas en este ámbito; en segundo lugar, se realizaron ejercicios para descubrir, por medio del corpus, unidades fraseológicas útiles en relación con la descripción de un camino. Finalmente se les pidió que, con lo aprendido, redactaran dos breves diálogos donde explicaran el camino a un viandante. Ofreceremos las conclusiones a las que hemos llegado con esta actividad, aportando los resultados del corpus de textos de los alumnos, así como los de una encuesta que rellenaron los estudiantes acerca de la actividad.

1. INTRODUCCIÓN

El presente trabajo da cuenta de una experiencia realizada en la asignatura Idioma Moderno II (Alemán), de nivel A1.2 del Marco Común Europeo de Referencia para las Lenguas (MCER), y perteneciente al título de Graduado en Estudios Ingleses (Universidad de Málaga). Partiendo de la premisa de la utilidad del corpus para la enseñanza-aprendizaje

de lenguas extranjeras (tema abordado, por ejemplo, en Flowerdew, 1996 y López Sanjuán, 2008), hemos diseñado una serie de actividades basadas en corpus para el aprendizaje de elementos gramaticales y léxicos de la lengua alemana; dentro del componente léxico nos hemos centrado en el aprendizaje de unidades fraseológicas (UF), en concreto colocaciones y *lexical bundles*. El uso de corpus (especialmente, el análisis de concordancias) por parte de los alumnos ha sido igualmente fundamental en el desarrollo de la actividad; así lo propone, por ejemplo, Flowerdew (1996: 87), para quien se pueden realizar en clase ejercicios de concordancias con corpus pequeños, para la enseñanza y el aprendizaje de lenguas.

Entenderemos *colocación* en este trabajo tal como la define el diccionario *Feste Verbindungen des Deutschen. Kollokationenwörterbuch für den Alltag* (Häcki Buhofer, Dräger, Meier y Roth, 2014: XI), según el cual se trata de combinaciones fijas de palabras, que cumplen tres criterios:

- Se componen de al menos dos palabras; por ejemplo, *Wolken verziehen sich*.

- Los componentes de una colocación se encuentran unidos unos a otros de forma más fija de lo que es usual entre las palabras de una oración; por ejemplo, *Wolken anschauen* es una combinación libre de palabras, mientras que se observa un rasgo de fijación en la combinación *Wolken ziehen vorüber*.

- Las colocaciones son en mayor o menor medida literales y no idiomáticas. Es decir, el significado global de la combinación de palabras no es figurado o metafórico. Sin embargo, sí se recogen en el diccionario combinaciones en las que la entrada (base) constituye una palabra literal y su colocativo es una palabra figurada, como *die Sonne lacht* ("el sol luce", literalmente "el sol ríe").

Por razones de coherencia y simplicidad, hemos seleccionado para nuestro estudio únicamente colocaciones que se encuentren presentes en el citado diccionario.

Por otro lado, los *lexical bundles* ("grupos léxicos") son secuencias de palabras que muestran una fuerte tendencia a coocurrir en un determinado registro, por ejemplo *do you want me to*, *I don't know what* en el registro hablado, y *in the case of the*, *it should be noted that* en el registro escrito académico (Biber et ál., 1999: 989).[109] Fuster-Márquez (2014: 88-89) resume las principales características que presentan los *lexical bundles* tras analizar varios trabajos basados en Biber et ál. (1999):

1. They may be regarded as extended collocations.

2. They are empirically identifiable through corpus analysis since they show a statistical tendency to co-occur.

3. To guard against any idiosyncrasy of individual texts or writers, they must occur at least in five different texts per million words.

[109] Otra definición aportada por Biber et ál. (1999: 990) es la siguiente: "Lexical bundles are recurrent expressions, regardless of their idiomaticity, and regardless of their structural status. That is, lexical bundles are simply sequences of word forms that commonly go together in natural discourse." De aquí se desprende que los *lexical bundles* pueden corresponderse tanto con expresiones literales (mucho más frecuentes) como idiomáticas, así como con estructuras sintácticas incompletas (mucho más frecuentes) o completas.

4. They are not perceptually salient: they are single-choice word combinations even though addressers or addressees would not recognise them as such.

5. Most LBs are unidiomatic.

6. Most LBs are not structurally complete, crossing the boundaries of grammatical categories.

7. They are the most frequent sequences in a register.

8. Their length is variable, but they should contain a minimum of three words.

9. They are fundamental building blocks of discourse, having discernible functions in particular registers.

Como vemos, los *lexical bundles* se pueden considerar colocaciones extendidas, con una tendencia estadística a coocurrir, en general no idiomáticas e incompletas estructuralmente; se trata además de combinaciones de al menos tres palabras que constituyen las secuencias más frecuentes de un registro.

Nos basamos en la idea de que la adquisición de estas combinaciones de palabras (colocaciones, *lexical bundles*) juega un papel muy importante en el aprendizaje global de una lengua extranjera. Así, partimos del principio de que gran parte de la lengua que usamos es de naturaleza fraseológica:

> It is now accepted that much of the language we use is phraseological in nature; that it is acquired, stored and retrieved as pre-formulated constructions (Bolinger, 1976; Pawley and Syder, 1983). These insights began to be supported empirically as computer technology permitted the identification of recurrent phraseological patterns in very large corpora of spoken and written English using specialised software (e.g. Sinclair, 1991). (*Academic Phrasebank*, 2015)

Este principio puede entonces emplearse para la enseñanza-aprendizaje de lenguas extranjeras. Allen (2009: 105) realza la importancia de los *lexical bundles* para la producción de un discurso efectivo: "language learners need to assimilate appropriate use of lexical bundles in order to create effective and successful, register convergent discourse." Consideramos que esta idea puede extenderse igualmente al aprendizaje de colocaciones.

Con ello, nuestro objetivo ha sido doble: por un lado, presentar la metodología de corpus a nuestros alumnos; en segundo lugar, aprovechar dicha metodología para la adquisición de unidades fraseológicas.[110]

El presente artículo se divide como sigue: la sección 2 presenta la metodología usada para llevar a cabo nuestro estudio; la sección 3 muestra los resultados obtenidos; finalmente, el apartado 4 aporta la discusión sobre los resultados y las conclusiones del estudio.

[110] En el presente trabajo emplearemos *unidad fraseológica* (UF) como término genérico para referirnos tanto a colocaciones como a *lexical bundles* (véase por ejemplo Rica Peromingo, 2010).

2. METODOLOGÍA

2.1. Corpus

Tanto el trabajo previo de preparación como el realizado propiamente en clase con los alumnos se han basado en corpus. Así, en primer lugar, compilamos un corpus acerca del tema elegido, *Orientierung in der Stadt* ("orientación en la ciudad"), con textos basados en la descripción del camino para llegar a un sitio dentro de una ciudad. Siguiendo las recomendaciones de Leal Riol (2013), tratamos de que el corpus fuera un fiel reflejo del alemán en uso y que ofreciera muestras lingüísticas adecuadas; en ese sentido, hemos tratado de escoger textos sencillos (en la medida de lo posible ajustados al nivel A1) y que contuvieran unidades fraseológicas incluidas en el libro de texto empleado en clase, *DaF kompakt A1* (Braun et ál., 2010). Así, a la hora de buscar los textos, se han realizado búsquedas en *Google* con algunas de las unidades fraseológicas presentes en el libro.

Finalmente compilamos un corpus formado por 27 textos, con 640 types y 4376 tokens. Aunque se trata de un corpus muy pequeño, creemos que la cuidadosa selección de los textos puede contribuir a realizar un estudio adecuado acerca de este dominio concreto, para un nivel de competencia básico. En ese sentido, estamos de acuerdo con Pérez Basanta y Rodríguez Martín (2007: 149):

> […] a corpus is large enough if it sufficiently exemplifies what we want our students to notice. In other words, the authority of a corpus will not depend solely on size but on the selection of adequate texts.

2.2. Selección de unidades fraseológicas

El siguiente paso fue seleccionar las unidades fraseológicas que íbamos a tratar con nuestros alumnos, basándonos en criterios de frecuencia, facilidad, productividad y necesidad (Leal Riol, 2013); de esta forma, hemos tenido en cuenta lo siguiente:

- Criterio de frecuencia:[111] Hemos tenido en cuenta cuáles son las UF más frecuentes en nuestro corpus, tras analizar mediante la aplicación de análisis de corpus *AntConc* (Anthony, 2015) cuáles son los *clusters/n-grams* de cuatro palabras más comunes. Así, los tres primeros son *biegen Sie links ab*, *auf der linken Seite* y *biegen Sie rechts ab*.

- Criterio de facilidad: Con respecto a este criterio hemos considerado el nivel de competencia de los alumnos, aún básico, ya que estudian alemán a un nivel A1.2. Por ello, hemos escogido UF que puedan emplear con los conocimientos gramaticales a su disposición.

- Criterio de productividad: Con UF productivas Leal Riol (2013: 169) se refiere a "aquéllas que mediante la adición o sustitución de alguno de sus elementos dan lugar a nuevas UF de significado parecido o no, pero de fácil comprensión." En nuestro caso, hemos escogido, por ejemplo, *biegen Sie [nach] links/rechts ab*.

[111] Al tratarse de un corpus muy pequeño, hemos tenido que renunciar al principio, aplicable a los *lexical bundles*, de que estas secuencias aparezcan en al menos cinco textos por millón de palabras. Sin embargo, como se ha comentado, esperamos haber compensado este defecto con una selección muy cuidadosa de los textos dentro de un ámbito muy específico.

- Criterio de necesidad: Se han seleccionado UF que, como indica Leal Riol (2013: 169), respondan a funciones comunicativas, es decir, que permitan a los estudiantes desenvolverse en situaciones cotidianas de comunicación. En nuestro caso, hemos seleccionado UF importantes para preguntar y responder acerca de la orientación en la ciudad, con vistas además a que puedan ser integradas en producciones escritas y orales.

Por otro lado, tuvimos en cuenta que las UF aparecieran en el libro de texto usado en la asignatura, a fin de que los alumnos pudieran relacionar la actividad que íbamos a realizar con los temas impartidos en clase y con los contenidos que deben estudiar; de hecho, como ya se ha indicado, tomamos algunas UF del libro como punto de partida para la búsqueda de textos.

Mostramos algunas de las colocaciones seleccionadas para nuestro estudio; para su inclusión en esta categoría se ha tenido en cuenta que figuraran en el diccionario *Feste Wortverbindungen des Deutschen* (Häcki Buhofer et ál., 2014):

auf der linken/rechten Seite[112, 113]

(immer) geradeaus gehen[114]

zu Fuß gehen[115]

die Straße überqueren

den Bus/die S-Bahn/[medios de transporte] nehmen

die Straße/die Kreuzung/[tipos de vía] nehmen

De otro lado, hemos seleccionado algunos *lexical bundles*, teniendo en cuenta las combinaciones de palabras más frecuentes de nuestro corpus (sobre todo n-gramas de cuatro y cinco palabras). Hemos escogido expresiones que no cumplen los criterios específicos para ser colocaciones: a veces se trata de estructuras incompletas, o no aparecen en el diccionario de colocaciones que hemos tomado como base. Sin embargo, se trata de combinaciones muy frecuentes y relevantes para los alumnos dentro del dominio que queremos estudiar y dentro del registro hablado. Algunos ejemplos son:

Wie komme ich zu …?

---

[112] En la obra de Häcki Buhofer et ál. (2014: 735) figuran las colocaciones *die linke | rechte | gegenüberliegende | entgegengesetzte | andere Seite*, acompañadas del siguiente ejemplo: "Auf der anderen Seite der Bank hockte ein dicker, nasser Frosch." Por lo tanto creemos justificado incluir *auf der linken/rechten Seite*, en su forma de sintagma preposicional (muy frecuente en nuestro corpus), como UF objeto de estudio.

[113] Por su frecuencia, *auf der linken/rechten Seite* también puede considerarse un *lexical bundle*. De hecho, Biber et ál. (1999: 1007) dan cuenta en su trabajo de algunas colocaciones (*have a cup of tea*, *have a look*).

[114] En Häcki Buhofer et ál. (2014: 312) encontramos *geradeaus gehen*, mientras que *Das Stilwörterbuch* (Duden, 2001: 358) incluye bajo la entrada *geradeaus* las combinaciones *geradeaus gehen* e *immer geradeaus*.

[115] *Zu Fuß* puede ser considerada también una locución adverbial, según la concepción de muchos investigadores; por ejemplo, García-Page (2008: 124) clasifica la UF equivalente en español *a pie* bajo el grupo de las locuciones adverbiales.

Biegen Sie [nach] links/rechts ab.

Gehen Sie [nach] links/rechts in die/den/das [nombre de calle o vía].

Con estos datos preparamos una hoja de actividades[116] para ser usada en clase con nuestros alumnos, que describiremos en el siguiente apartado.

2.3. Actividad en clase

Previamente a la sesión de trabajo con corpus, ya había sido tratado brevemente en clase el tema de *Orientierung in der Stadt*. Luego, en una sesión de clase en el aula de informática, pasamos a realizar la actividad con corpus. En primer lugar, se explicó a los estudiantes el concepto de corpus y se resaltó la utilidad que pueden tener para el aprendizaje de lenguas extranjeras. A continuación, se les proporcionaron algunas nociones básicas de *AntConc 3.4.3*, que instalaron y comenzaron a usar en sus respectivos ordenadores. Finalmente pasamos a realizar la actividad con corpus. Después de abrir el corpus que habíamos preparado previamente, se les pidió que extrajeran las palabras más frecuentes del corpus, mediante la herramienta *Word List* de *AntConc*; a continuación, se les indicó que seleccionaran en la lista los tres verbos más frecuentes. Como segunda parte de la actividad, pasaron a realizar varios ejercicios para descubrir UF, por medio de la herramienta *Concordance* del programa. Se propusieron dos tipos de ejercicios: de rellenar huecos y de buscar oraciones de ejemplo dado un componente de la UF. Mostramos algunos de los ejercicios propuestos:

Wie _____ ich zum Bahnhof?
immer _____
_____ _____ _____ Seite

Busca una oración donde aparezca el verbo *überqueren*.

¿Qué dos tipos de sustantivos suelen acompañar a *nehmen*? Escribe dos ejemplos basados en el corpus.

Nehmen Sie …….
Nehmen Sie …….

Finalmente se pidió a los alumnos que, con lo aprendido y haciendo uso del corpus, realizaran una breve tarea de redacción: dos diálogos donde explicaran a un viandante cómo llegar a dos sitios, respectivamente, sobre la base de un mapa disponible en su libro de texto.

3. RESULTADOS

Un total de 22 alumnos entregaron la redacción, por lo que contamos con 44 textos: contienen 180 types y 1536 tokens. Mediante la herramienta *Clusters/N-Grams* de *AntConc*, obtuvimos que las cadenas de cuatro palabras más frecuentes fueron:

[116] En Flowerdew (1996) y Olivier et ál. (2007) encontramos algunos ejercicios de muestra que nos sirvieron de ayuda.

1. wie komme ich zum (frec. 16)
2. gehen Sie nach links (frec. 14)
3. gehen Sie hier geradeaus (frec. 13)
4. Entschuldigung, wie komme ich (frec. 11)
5. gehen Sie geradeaus bis (frec. 11)
6. komme ich zum Kornhaus (frec. 9)
7. biegen Sie links ab (frec. 8)
8. gehen Sie geradeaus und (frec. 8)
9. gehen Sie zuerst die (frec. 8)

Otros resultados interesantes los encontramos en puestos más bajos:

19. auf der linken Seite (frec. 5)
22. gehen Sie immer geradeaus (frec. 5)
23. gehen Sie nach rechts (frec. 5)
30. auf der rechten Seite (frec. 4)
31. biegen Sie rechts ab (frec. 4)

Por otro lado, si buscamos específicamente mediante la herramienta *Concordance* las UF que tratamos en la actividad, obtenemos lo siguiente:

auf der linken Seite: frec. 5
auf der rechten Seite: frec. 4
gehen Sie geradeaus: frec. 36
gehen Sie immer geradeaus: frec. 5
immer geradeaus: frec. 8
zu Fuß gehen/gehen Sie zu Fuß: frec. 0
zu Fuß: frec. 1
die Straße überqueren/Überqueren Sie die Straße: frec. 0
den Bus/die S-Bahn/[medios de transporte] nehmen: frec. 0
die Straße/die Kreuzung/[tipos de vía] nehmen: frec. 2

Wie komme ich zu …?: frec. 20
Biegen Sie [nach] links ab: frec. 10
Biegen Sie [nach] rechts ab: frec. 5
Gehen Sie [nach] links in die/den/das [nombre de calle o vía]: frec. 5
Gehen Sie [nach] rechts in die/den/das [nombre de calle o vía]: frec. 1

Como vemos, los alumnos emplearon con frecuencia algunas de las UF estudiadas, mientras que otras han tenido una ocurrencia muy baja o no se han utilizado en ningún caso.

En días posteriores a la actividad en clase, se pasó a los alumnos una encuesta en línea anónima, con el fin de averiguar si les había parecido útil el corpus para aprender una lengua extranjera, y para redactar acerca del dominio concreto estudiado. Contestaron 13 de los alumnos. Mostramos aquí algunas de las preguntas y los resultados:

¿Es la primera vez que utiliza un buscador de concordancias para para estudiar una lengua extranjera o preparar una redacción (por ejemplo, AntConc*)?*
Sí: 12 (92,31%)
No: 1 (7,69%)

¿Le ha sido de ayuda el corpus para adquirir vocabulario acerca del dominio estudiado?
Sí: 11 (84,62%)
No: 2 (15,38%)

¿Le ha sido de ayuda el corpus para realizar la redacción acerca del dominio estudiado?
Sí: 10 (76,92%)
No: 3 (23,08%)

¿Considera que volverá a utilizar un programa de concordancias como recurso para aprender un idioma?
Sí: 11 (84,62%)
No: 2 (15,38%)

En la siguiente sección comentaremos los resultados y ofreceremos algunas conclusiones.

4. DISCUSIÓN Y CONCLUSIONES

Los *clusters* extraídos del corpus de textos de los alumnos muestran que estos usaron siete de las nueve UF presentadas aquí, aunque *zu Fuß*, sin el verbo *gehen*, sólo apareció una vez. Algunas combinaciones han sido muy empleadas, lo que consideramos muy positivo: por ejemplo, *wie komme ich zum/zur*, *gehen Sie nach links* (aunque a menudo sin el complemento de dirección "in + sintagma nominal en acusativo"), o *gehen Sie hier geradeaus*. Por el contrario, UF como las que incluyen el verbo *nehmen* (junto con un medio de transporte o una vía), no han sido apenas usadas.

Aunque indagar en los motivos de estas diferencias excede los límites de este artículo, sí podemos aproximarnos a algunas causas: por un lado, el plano en el que los alumnos tenían que basarse era pequeño, así como las rutas que tenían que describir; por ello resulta comprensible que no tuvieran que usar las expresiones *zu Fuß gehen* o *den Bus nehmen*, por ejemplo. Por otro lado, como se ha manifestado en el artículo, el tema de *Orientierung in der Stadt* ya se había tratado en clase, incluyendo las estructuras gramaticales más típicas y necesarias en este ámbito, al tiempo que la actividad de corpus preparada contenía igualmente algunas tareas más bien centradas en la gramática; por ello es posible que muchos alumnos hayan preferido usar la combinación *über [die Straße/die Kreuzung/den Platz…] gehen* en lugar de la UF *die Straße überqueren*.

De otro lado, los resultados de la encuesta anónima en línea apuntan a que los alumnos, a pesar de desconocer en su mayoría el concepto de corpus y concordancias, han considerado útil la actividad, tanto en lo referente a la adquisición de vocabulario, como en lo concerniente a la producción escrita. La mayoría considera además que volverá a utilizar un programa de concordancias como recurso para aprender una lengua extranjera.

Con todo, consideramos que el uso de corpus en la clase de lengua extranjera es, de un lado, posible, y de otro, útil para la enseñanza y refuerzo de diversos aspectos lingüísticos,

entre ellos, las UF (en este caso, colocaciones y *lexical bundles*). Además, como se ha puesto de manifiesto, este tipo de actividades también se pueden llevar a cabo en niveles básicos (A1, A2).

Agradecimientos

El presente trabajo ha sido realizado en el seno de los proyectos 'INTELITERM: Sistema inteligente de gestión terminológica para traductores' (n. ref. FFI2012-38881, 2012-2015. MEC), 'TERMITUR: Diccionario inteligente TERMInológico para el sector TURístico (alemán-inglés-español)' (Ref. HUM2754, 2014-2017. Junta de Andalucía) y 'TRADICOR: Sistema de gestión de corpus para la innovación didáctica en traducción e interpretación' (PIE 13-054).

Bibliografía

ALLEN, D., 2009. Lexical Bundles in Learner Writing: An Analysis of Formulaic Language in the ALESS Learner Corpus. *Komaba Journal of English Education*, 1, pp. 105-127.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. Y FINEGAN, E., 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

BRAUN, B., DOUBEK, M., FRATER-VOGEL, A., TREBESIUS, U., VITALE, R. Y SANDER, I., 2010. *DaF kompakt A1*. Klett.

DUDEN, 2001. *Das Stilwörterbuch*. 8.ª ed. Mannheim: Bibliographisches Institut & F. A. Brockhaus.

FLOWERDEW, J., 1996. Concordancing in language learning. En: M. Pennington, ed. 1996. *The power of call*. Houston, TX: Athelstan. pp. 97-113.

FUSTER-MÁRQUEZ, M., 2014. Lexical bundles and phrase frames in the language of hotel websites. *English Text Construction*, 7(1), pp. 84-121.

GARCÍA-PAGE SÁNCHEZ, M., 2008. *Introducción a la fraseología española. Estudio de las locuciones*. Rubí (Barcelona): Anthropos Editorial.

HÄCKI BUHOFER, A., DRÄGER, M., MEIER, S. Y ROTH, T., 2014. *Feste Wortverbindugen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen: Francke Verlag.

LEAL RIOL, M. J., 2013. Estrategias para la enseñanza y aprendizaje de la fraseología en español como lengua extranjera. *Paremia*, 22, pp. 161-170.

LÓPEZ SANJUÁN, V., 2008. Integración de los corpus como herramienta de apoyo en la enseñanza de ESP. *Porta Linguarum*, 10, pp. 115-136.

OLIVIER, N., BREMS, L., DAVIDSE, K., SPEELMAN, D. Y CUYCKENS, H., 2007. Pattern-learning and pattern-description: an integrated approach to proficiency and research for students of English. En: E. Hidalgo, L. Quereda y J. Santana, eds. 2007. *Corpora in the Foreign Language Classroom. Selected Papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6)*. Ámsterdam, Nueva York: Rodopi. pp. 221-236.

PÉREZ BASANTA, C. Y RODRÍGUEZ MARTÍN, M. E., 2007. The application of data-driven learning to a small-scale corpus: using film transcripts for teaching conversational skills. En: E. Hidalgo, L. Quereda y J. Santana, eds. 2007. *Corpora in the Foreign Language Classroom. Selected Papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6)*. Ámsterdam, Nueva York: Rodopi. pp. 141-158.

RICA PEROMINGO, J. P., 2010. Colocaciones gramaticales en la producción escrita de estudiantes universitarios españoles. *Reduca (Filología). Serie Lengua Inglesa*, 2(1), pp. 1-25.

(Páginas web)

ACADEMIC PHRASEBANK, The University of Manchester, 2015. *About Academic Phrasebank*. [en línea] Disponible en: http://www.phrasebank.manchester.ac.uk/about-academic-phrasebank/ [Fecha de acceso 30 de junio de 2015].

ANTHONY, L., 2015. *AntConc Homepage*. [en línea] Disponible en: http://www.laurenceanthony.net/software/antconc/ [Fecha de acceso 30 marzo 2015].

¿DAR O ECHAR UN PIROPO? ME QUEDO LOCO, NUNCA ACIERTO. COLOCACIONES VERBALES EN ESPAÑOL Y PORTUGÉS

Javier Martín Salcedo
Universidad Federal do Ceará (UFC)
javims29@hotmail.com

Resumen

El proceso de enseñanza y aprendizaje de las Unidades Fraseológicas, y en el caso concreto de las colocaciones verbales, es bastante complejo, ya que sólo la propia consagración del uso determina la configuración de las mismas. En este sentido, este estudio pretende presentar algunos usos lingüísticos que difieren en ambas lenguas en cuanto al uso y a la elección del colocativo en dichas estructuras. Asimismo, cabe destacar que la investigación se desarrolló en la Universidad Federal de Ceará, en Brasil, con base a producciones de alumnos brasileños de español en dicha universidad y alumnos españoles de portugués de la E.O.I de Castellón a través de la creación de un grupo de Facebook, en el que se proporcionaron intercambios lingüísticos dirigidos y basados en cuestionarios. Para fundamentar nuestro artículo, tendremos en cuenta a diversos autores como Corpas Pastor, Molina Plaza, Penadés Martínez, entre otros etc… De este modo, la elaboración de un compendio de colocaciones verbales en contraste es más que necesario, seleccionando las unidades con más frecuencia de uso en la lengua coloquial e intentando resolver los principales problemas de interlengua de los estudiantes que desconocen el uso habitual de este aspecto fraseológico entre dos lenguas tan próximas, pero a la vez tan distantes, como son el español y el portugués.

1. CONSIDERACIONES INICIALES

En este texto trataremos, observados los procesos de interlengua en los alumnos, algunos aspectos relacionados al aprendizaje y uso de las colocaciones verbales. Pudimos observar, por una parte, que existían una amplia variedad de colocaciones verbales con los verbos *ficar* y *tirar* del portugués brasileño que los discentes de E\LE no entendían o no sabían producir y, por otra, se demostró asimismo que los alumnos de portugués como lengua extranjera no eran capaces de abordar la amplia gama de estructuras con el verbo echar y que en el español peninsular tienen un gran uso. Antes de continuar con la exposición, cabe resaltar que este artículo toma como muestra lingüística la variante del

portugués brasileño (PB), contrastándola con la variante del español europeo o peninsular (EE).

Iniciamos esta investigación con una breve, pero necesaria, revisión en torno a las definiciones y características de estas unidades fraseológicas. Seguiremos nuestro recorrido con unas reflexiones sobre el proceso de traducción y la búsqueda de equivalencias como coadyuvante al proceso de interlingua tan necesario dentro del proceso de aprendizaje de lenguas extranjeras. Una vez analizadas las diferentes perspectivas en torno a los procesos de transferencia e interferencia, culminaremos con una muestra extraída de un grupo intercultural de Facebook a través de la observación en producciones escritas de alumnos brasileños de español y alumnos españoles de portugués con el simple objetivo de reflexionar sobre la elaboración de actividades didácticas que tengan como base la fraseología y la lingüística contrastiva.

2. ¿QUÉ ENTENDEMOS POR COLOCACIÓN VERBAL? ES PARA VOLVERSE LOCO

Los alumnos brasileños de E\LE con poco dominio en la lengua extranjera creen erróneamente que el verbo *ficar* (PB) encuentra su equivalente en el verbo *quedarse,* cometiendo una gran cantidad de usos lingüísticos impropios como quedarse loco (ficar doido) en lugar de volverse loco; quedarse amigos (ficar amigos) en vez de hacerse amigos, etc.... Lo mismo sucede si proponemos los siguientes ejemplos: sacar la razón de alguien (tirar a razão dele) cuando debería ser quitarle la razón a alguien; sacarse la ropa (tirar a roupa) en vez de quitarse la ropa; estructuras que nos demuestran cómo estos estudiantes traducen el verbo tirar siempre por el verbo sacar en contextos donde no cabe el mismo, debido al desconocimiento de que las posibilidades combinatorias son reducidas, ya que existe cierto grado de fijación en el uso del colocativo, pues se tratan de colocaciones verbales. Ahora bien, ¿qué entendemos por colocación verbal?

Según Corpas Pastor (1996, p.61) Firth fue el primero autor que usa el término colocación en lingüística y añade que éste fue introducido al ámbito hispánico gracias a Manuel Seco en 1978. En cuanto a su definición, la autora afirma que éste se trata de una "coaparición frecuente y lineal de las palabras en el discurso" (CORPAS PASTOR, 1996, P. 56). Sin embargo, el problema es mucho más complejo de lo que parece y los diferentes autores no se ponen de acuerdo a la hora de clasificar, determinar o denominar las diferentes unidades fraseológicas (UFs) que existen. Del mismo modo, Penadés Martínez (1999, p.19), resalta lo que venimos manifestando, es decir, que no hay unanimidad en cuanto a la sistematización de las unidades fraseológicas. Tal hecho podemos observarlo en la clasificación de Sevilla Muñoz (1999) con respecto a las UFs, donde resulta cuanto menos curioso que en su trabajo no mencione el término colocación, ya que en ocasiones se confunde o se toma como sinónimo de locución.

No obstante, Corpas Pastor (1996, p.50-52), sí que recoge el concepto y clasifica las unidades fraseológicas (UFs) en colocaciones y locuciones, que llama de oraciones sin sentido completo y enunciados fraseológicos u oraciones con sentido completo en donde tendríamos las paremias y las fórmulas rutinarias. Asimismo, Penadés Martínez (1999, p. 12-14) incluye también las mismas, como expresiones o unidades pluriverbales, lexicalizadas o habituales dentro de la fraseología. En cuanto a la terminología colocación/locución y sus características, podríamos distinguir dos rasgos que nos permitan diferenciar las colocaciones de las locuciones y que a continuación detallamos. Si

bien es cierto que las colocaciones tienen rasgos propios de las UFs como fijación o como la inalterabilidad semántica de las palabras que las componen, podemos resaltar que en éstas hay una cierta autonomía de los componentes y, lo que es más importante, que existe la posibilidad de sustituir el colocativo (palabra que acompaña al colocado o al elemento principal de la colocación), que en el caso de las locuciones no se da. Sirvámonos de un ejemplo para entender la diferencia: es cierto que "volverse loco" encierra un sentido propio, sin embargo la estructura admite otras combinaciones como podrían ser "volverse tonto", "hacerse el loco" o simplemente "estar loco". Aunque cabe resaltar que cada estructura tiene un significado y un contexto de uso diferente.

De cualquier forma, las colocaciones han sido objeto de estudio en fraseología desde las más diversas perspectivas. Ahora cabe preguntarse, cómo definimos colocación y cómo distinguir lo que es colocación de lo que no es. Para Molina Plaza (2003, p. 429): son sintagmas completamente libres a los que el uso ha dado cierto grado de restricción combinatoria. Es interesante reflexionar y detenerse en esta definición, ya que parece que es el propio uso de la lengua el que hace que una determinada comunidad lingüística sea la que consagre y fije la colocación. Veamos un ejemplo para entenderlo mejor, en PB tenemos la colocación "tirar a roupa" que en español peninsular diríamos "quitarse la ropa". No obstante, existe también la variante "sacarse la ropa" como posible equivalente del verbo tirar en este contexto, sin embargo, a pesar de estar extendido este "sacarse la ropa" en todos los países hispanohablantes, es mucho menos empleado que la primera opción debido a la consagración del uso. Ahora bien, cabe plantearse qué debemos enseñar a nuestros alumnos de E/LE: quitarse la ropa o sacarse la ropa. Parece claro que es el uso mayoritario el que debe imponer la norma.

Para concluir este apartado veamos la definición de colocación que propone Corpas Pastor (1996, p.66):

> Unidad fraseológica formada por dos unidades léxicas en relación sintáctica, que no constituyen, por sí mismas, actos de habla ni enunciados, y que debido a su fijación en la norma, presentan restricciones de combinación establecidas por el uso, generalmente de base semántica.

Por tanto, cabe entender que las colocaciones son unidades fraseológicas de pleno derecho, ya que son polilexicales, es decir, mantienen una relación entre dos o más palabras y tienen una fijación o institucionalización debido al uso. De la misma forma, Corpas Pastor menciona acerca de las unidades fraseológicas que "entre sus rasgos figuran: polilexicalidad, alta frecuencia de aparición, institucionalización, estabilidad (fijación y especialización semántica), idiomaticidad y variación potenciales". (CORPAS PASTOR, 1998a, p.167). Sin lugar a dudas, estos rasgos son comunes también a las colocaciones, por lo tanto, éstas serían objeto de la fraseología.

3. BUSCANDO EQUIVALENCIAS PARA NO QUEDARSE TRISTE O QUEDARSE VIEJO EN LA INTERLENGUA

Es cierto que buscar equivalentes perfectos entre dos lenguas no siempre es fácil. Es más, en determinadas ocasiones, se vuelve una tarea difícil e incompleta. Tomaszczyk (1983, p. 48) nos propone una clasificación en torno a este asunto y que divide en tres aspectos: equivalencia completa, equivalencia parcial, y la no equivalencia, dando respuesta a la dificultad a veces de no encontrar equivalentes perfectos. De este modo, Molina Plaza

(2003, pp. 429-430): señala que la equivalencia parcial se da, sin poder hablar de completa, cuando se provocan incoherencias de tipo semántico, figurativo, connotativo.

Por otra parte, según Amparo Hurtado Albir (2007, p. 204) el concepto de equivalencia es causa de gran polémica debido a su carácter central e importancia en la definición de traducción. Sin lugar a dudas, y en la misma línea de pensamiento de esta autora es un concepto que involucra ambigüedad, hecho también advertido por Nord (1991) y que relatamos en palabras textuales: "El concepto de la equivalencia es uno de los conceptos más ambiguos en los estudios sobre traducción y, por consiguiente, se ha interpretado de muchas formas distintas". Además tenemos como hemos destacado anteriormente el concepto de no equivalencia y que sustentamos en la obra de Baker (1992) en la cual expone que cuando no existe una expresión idéntica o similar, podemos recurrir a la paráfrasis que consiste en la explicación del significado de la expresión.

Por desgracia, aún queda mucho camino por recorrer en relación a estudios, materiales didácticos, manuales, etc… con un enfoque que realmente considere la lingüística contrastiva y la fraseología entre el portugués brasileño y el español peninsular, y que responda tanto a las necesidades de los aprendices brasileños como también sirva de ayuda a los aprendices españoles de PB. Hoyos Andrade (1993, p.11) afirma que al enseñar castellano a alumnos brasileños hay que tener presente y distinguir cuidadosamente, lo que es igual a las dos lenguas, lo que parece igual y no lo es y lo que es efectivamente distinto. En realidad, lo que propone este autor es materiales basados en lingüística contrastiva portugués-español, teniendo mucho cuidado a la hora de saber distinguir lo que es portugués de lo que es español. Por otra parte, Takeuchi (1980) afirma que las lenguas portuguesa y española poseen un pequeño número de palabras diferentes y una gran variedad de vocablos comunes. Podríamos estar de acuerdo con este autor desde una perspectiva morfológica de la lengua, sin embargo, el mismo no podría hacer una afirmación de este tipo si hubiera tenido en cuenta el uso y distribución de los mismos, las connotaciones de los términos, diferentes contextos, la formación dentro de las unidades fraseológicas, los coloquialismos, etc… donde el portugués y el español posee un número importante de diferencias. Sírvase como ejemplo dos palabras entre muchas otras: "molestar": en portugués tiene el significado de abusar (EE) y en español sería incomodar, perturbar (PB).

Briones (2000), al igual que muchos otros autores, afirma que "los errores más frecuentes de traducción entre el portugués y el español o viceversa, ocurren debido al desconocimiento de los falsos amigos". Dicha afirmación se encaja perfectamente en el concepto de hablante ingenuo de Filmore (1979), ya que parece olvidar no sólo los errores de índole gramatical, sino que no tiene en cuenta de nuevo la importancia de la fraseología, el uso, registro que es, en realidad, donde radican los verdaderos problemas en la traducción y no tanto en el desconocimiento de los falsos amigos y sí en el desconocimiento de las colocaciones, locuciones, binomios, frases hechas, expresiones idiomáticas, etc… así como sus respectivos usos.

Para terminar, queremos llamar la atención del título: quedarse viejo y quedarse triste, dos ejemplos claros de interlengua de los aprendices brasileños de español, cuando debería ser hacerse viejo en el primer caso, y en muchos otros casos ponerse triste, y no porque lo digamos en este artículo, sino que es el uso de la propia colocación el que habla más alto.

4. LA OBSERVACIÓN Y RECOGIDA DE DATOS CONFIRMAN LOS USOS

El procedimiento metodológico para la recopilación de las estructuras y sus diferentes formas y usos se realizó a través de la elaboración de un cuestionario de 80 preguntas dirigidas en portugués y español para incentivar indirectamente que los intervinientes se vieran afectados en su proceso de transferencia negativa (JESSNER, 1996) de la lengua materna a la lengua meta. Para que se entienda mejor, se propone una de las preguntas que formó parte de dicho cuestionario: ¿para qué sirve un sacapuntas?. Todos los participantes brasileños sin excepción respondieron: para hacer la punta del lápiz, dado que es la estructura del portugués (para fazer a ponta do lápis) transferida literalmente al español. Del mismo modo, sucedió con los españoles cuando escribieron tirar a ponta ao lápis. Tal hecho nos demuestra que estamos ante colocaciones en contraste lingüístico y que el colocativo funciona de forma diferente en ambas lenguas debido a la fijación en el uso y que, por lo tanto, induce en errores de expresión por parte del estudiante extranjero. Esta observación se realizó durante tres meses a través de un grupo cerrado de Facebook llamado PortuÑhol, que estaba formado por alumnos españoles de portugués y brasileños de español. Se eligió este canal de comunicación porque entendemos que tanto la interacciones en *facebook* como en *whatsapp* son un claro ejemplo de corpus escrito de la interacción oral. Dicho esto, observamos que en estos intercambios comunicativos resultaron una gran cantidad de interferencias y malentendidos como era de esperar y de los que no podremos abarcar aquí una buena parte por cuestiones de espacio. De esta forma, sí que parece claro en relación a las colocaciones verbales, los españoles emplean mucho el verbo echar como colocativo que adquiere varios equivalentes en portugués y que supone, en consecuencia, una dificultad para el estudiante brasileño y estos, en su caso emplean a menudo los verbos ficar y tirar en estructuras que en español se expresan de otra manera. En este sentido, queremos ofrecer aquí una pequeña muestra sistematizada para entender que la cuestión es bastante compleja, así como hasta hoy poco investigada, con el fin de que la misma sirva como preámbulo a trabajos posteriores.

Por tanto, presentamos algunas estructuras con ejemplos extraídos de nuestro corpus:

ECHAR (SE) + SUSTANTIVO

- rato = Aquí **echando el rato** viendo la tele mientras él llega // *Estou aqui assistindo tevé, só para* ***passar o tempo*** *enquanto ele chega.*
- bronca = Mi padre me **echó una bronca** porque llegué tarde a casa // *Meu pai me* ***deu uma bronca*** *porque cheguei tarde em casa.*
- piropo = El graciosillo le **echó un piropo** a la vecina // *O gaiato* ***deu uma cantada*** *na vizinha.*
- novio = Juana **se echó un novio** en Carnaval // *Joana* ***arranjou um namorado*** *no Carnaval.*
- culpa = Él hizo el proyecto mal y me **echó la culpa** // *Ele fez o projeto errado e* ***jogou a culpa*** *em mim.*

FICAR + ADJETIVO

- apaixonado (a) = Ele **ficou apaixonado** por ela // *Él* ***se enamoró*** *de ella.*
- bom / boa = A situação **ficou boa** // *La situación* ***se puso bien*** *(mejoró).*
- doente = Ao saber da notícia, **ficou doente** // *Al saber la noticia,* ***se puso enfermo.***

- famoso (a) = Após a apresentação na televisão, eles **ficaram famosos** // *Tras su presentación en televisión* ***se hicieron famosos***.
- feliz = **Fico feliz** em te ver progredir // ***Me alegro*** *de ver como progresas.*
- miserável = Depois que enricou, **ficou miserável** // *Tras hacerse rico,* ***se volvió un miserable.***
- namorado / amigo (a) = Depois da festa, ele **ficou** muito **amigo** do rapaz // *Después de la fiesta,* ***se hizo*** *muy* ***amigo*** *del chico.*
- solteiro ou viúvo (a) = Com a morte da esposa, ele **ficou viúvo** // *Con la muerte de la esposa,* ***se quedó viudo***.
- tímido (a) = Depois de velho, você vai **ficar tímido**? // *De viejo, ¿ahora te vas a* ***volver tímido****?.*
- velho (a) = Ele **ficou velho** muito rápido // ***Se hizo viejo*** *muy rápido.*

TIRAR + SUSTANTIVO

- brincadeira = Ele foi **tirar brincadera** e se deu mal // *Él fue a* ***hacer una broma*** *y salió mal parado.*
- cochilo = Ele **tirou um cochilo** depois do almoço // ***Se echó una siesta*** *después de almorzar.*
- cópia = Vou **tirar uma cópia** desse CD // *Voy a* ***sacar una copia*** *del disco.*
- dúviva = O professor ***tira as dúvidas*** dos alunos // *El profesor* ***saca de dudas*** *a (aclara las dudas de) los alumnos.*
- férias = No próximo mês vou **tirar** minhas **férias** // *El próximo mes voy a* ***cogerme vacaciones***.
- mesa = Depois de comerem, o pai **tirou a mesa** // *Después de comer, el padre* ***quitó la mesa.***
- minuto / hora = Eu **tiro** dez **minutos** de casa para a universidade // ***Echó*** *diez* ***minuto****s de casa a la universidad.*
- nota = Eu **tirei uma** boa **nota** em matemática // ***Saqué una*** *buena* ***nota*** *en matemáticas.*
- poeira = Ela **tirou a poeira** dos móveis // *Ella* ***quitó el polvo*** *de los muebles.*
- ponto = O professor vai **tirar um ponto** de cada um // *El profesor nos va a* ***quitar un punto*** *a cada uno.*

5. CONCLUSIONES

Podemos constatar que el verbo *echar* (EE) encuentra equivalentes en estos casos en *dar, jogar, passar* y *arranjar* (PB); el verbo *ficar* (PB) en *volverse, quedarse, hacerse, ponerse,* etc… (EE); y el verbo *tirar* (PB) en *sacar, tirar, coger, hacer, echar, etc…* dependiendo de la estructura más o menos fija en la lengua meta. Por tanto, estamos ante una propuesta de colocaciones verbales en contraste, queriendo defender la validez y utilidad de los análisis contrastivos bajo una perspectiva didáctica del mismo contraste entre la lengua materna y la lengua meta. Estos estudios deben aplicarse y tomarse como referencia a la hora de la elaboración de materiales en clase de lenguas extranjeras, teniendo como objetivo primordial entender mejor las dificultades de los alumnos ante su propia interlengua. En este sentido, el estudio contrastivo es uno, entre otros tantos elementos, en el proceso de adquisición de idiomas.

Es un instrumento pedagógico que puede ser sumamente importante para determinar la aplicación didáctica de determinada metodología y actividades específicas en clase. Asimismo, no podemos olvidar la importancia de la fraseología y las estructuras sintácticas en estos estudios como punto de partida que oriente cualquier trabajo en lingüística contrastiva, ya que como hemos demostrado en este artículo una misma palabra en la lengua de partida puede comportarse de diversas formas, adquirir nuevos significados, e incluso nuevas estructuras gramaticales en la lengua meta debido al contexto y uso que los hablantes de una comunidad lingüística hacen de la misma.

Bibliografía

BAKER, M., 1992. Idioms and fixed expressions. In *other words, a coursebook on translation,* London: Routledge, pp. 63-79.

BRIONES, A. I., 2002. Dificultades de la traducción portugués-español vistas a través de la lingüística contrastiva, *Actas del IX Congreso Brasileño de Profesores de Español,* Brasilia: Consejería de Educación, Embajada de España. pp. 59-68.

CORPAS PASTOR, G., 1996. *Manual de fraseología española,* Madrid, Gredos.

CORPAS PASTOR, G., 1998. Criterios generales de clasificación del universo fraseológico de las lenguas con ejemplos tomados del español y del inglés, In M. Alvar Ezquerra y G. Corpas Pastor. 1998. *Diccionarios, frases, palabras.* Málaga. Servicio de publicaciones de la universidad pp. 157-187

FILLMORE, CHARLES J., 1979. Innocence: a Second Idealization for Linguistics.. Berkeley Linguistic Society, 5, pp. 63-76.

HOYOS ANDRADE, R. E., 1993. Elementos de una gramática para la enseñanza del español en el Brasil. *Anuario brasileño de estudios hispánicos*. Brasilia: Consejería de Educación de la Embajada de España, V. 3, pp. 11-15.

HURTADO ALBIR., 2007. Amparo. *Traducción y traductología: introducción a la traductología*, Madrid, Cátedra.

JESSNER, U., 1986. La transferencia en la adquisición de la segunda lengua. In Cenoz, J.; Valencia, J. F. (Org.). *La competencia pragmática: elementos lingüísticos y psicosociales.* Bilbao: Servicio Editorial de la Universidad del País Vasco, pp. 141-153. 1986.

MOLINA PLAZA, S., 2003. La traducción de las unidades fraseológicas inglés-español: el caso de las colocaciones y frases idiomáticas. In: *Las palabras del traductor,* Universidad Castilla- La Mancha, Disponible en: cervantes.es/lengua/esletra/035_molina.pdf. Acceso en: 05.agosto.2014.

NORD, Ch., 1991. *Text Analysis in Translation.* Amsterdam: Rodopi.

PENADÉS MARTÍNEZ, I., 1999. *La enseñanza de las unidades fraseológicas.* Madrid, Arco/Libros.

SEVILLA MUÑOZ, J., 1999. *Divergencias en la traducción de expresiones idiomáticas y refranes (francés-español).* Disponible en: www.deproverbio.com/Dpjournal/DP,5,1,99/SEVILLA/DIVERGENCIAS.html.

TAKEUCHI, N.N., 1980. *Um estudo de interferência lexical.* Curitiba. Tesis inédita, UFPR.

TOMASZCZYK, J, 1983. On bilingual dictionaries. The case for bilingual dictionaries for foreign language learners. In HARTMANN, R. R. K. (Ed.) *Lexicography: Principles and Practice,* London: Academic Press. pp. 41-52.

LINGUEE COMO HERRAMIENTA DE ENSEÑANZA-APRENDIZAJE DE LAS UNIDADES FRASEOLÓGICAS*

Mª Eugênia Olímpio de Oliveira Silva
Universidad de Alcalá
eugenia.olimpio@uah.es

Inmaculada Penadés Martínez
Universidad de Alcalá
inmaculada.penades@uah.es

Resumen

En este trabajo se presenta una propuesta para utilizar *Linguee* como herramienta de enseñanza y aprendizaje de las locuciones por parte de estudiantes brasileños de español como lengua extranjera. La actividad didáctica, dirigida a alumnos del nivel C1 o C2, se materializa en la elaboración de un diccionario bilingüe de este tipo de unidad fraseológica, realizado por los alumnos durante todo un curso a partir de las instrucciones explícitas del profesor sobre la clase de fraseologismos con los que se va a trabajar, sobre el tipo de información que se debe buscar para cada locución y sobre la utilidad, para llevar a cabo la tarea, de los diccionarios monolingües de español y portugués, así como de los bilingües en relación con estas dos lenguas.

1. INTRODUCCIÓN

La aplicación de la lingüística de corpus a la enseñanza del español como L2 es relativamente reciente y está poco desarrollada en comparación con el inglés en lo que se refiere a corpus producidos por hablantes nativos y no nativos. Junto a ello, todavía es menor la utilización por el profesor de ELE de las técnicas que proporciona el procesamiento del lenguaje natural para analizar unidades fraseológicas, aunque la fraseología computacional ya permite su detección, extracción, análisis y representación (Corpas Pastor, 2013). En este sentido, presentamos la aplicación a la enseñanza de ELE de *Linguee*, una herramienta considerada un corpus público de fácil acceso y, asimismo, un corpus paralelo (Alonso Jiménez, 2013), si bien para sus creadores es un diccionario inteligente. Este recurso ha recibido la atención de diferentes estudios que han examinado su potencial como instrumento para el traductor o para la enseñanza de la traducción

* Por limitaciones de edición, este texto recoge una versión muy reducida de la comunicación que, bajo el mismo título, se presentó en el *Congreso Internacional de la Sociedad Europea de Fraseología* (*EUROPHRAS 2015*), celebrado en la Universidad de Málaga del 29 de junio al 1 de julio de 2015. La versión completa de la comunicación, y además ampliada, está publicada en el número XIII de la revista *Lingüística en la Red* www.linred.es .

(Durán Muñoz, 2011; Patin, 2014). En menor medida, ha sido estudiada su aplicabilidad al proceso de enseñanza-aprendizaje de lenguas extranjeras (Buyse y Verlinde, 2013; Volk *et al*., 2014). Así pues, a partir del empleo pedagógico de un recurso tecnológico, puesto al servicio del aprendizaje de la fraseología (Coll, 2004; Solano Rodríguez, 2012), vamos a exponer los resultados que se podrían obtener del análisis en *Linguee*, por parte de aprendientes de español, de un conjunto de unidades fraseológicas de esta lengua y de sus equivalentes de traducción al portugués, análisis llevado a cabo desde el funcionamiento de las unidades en el discurso y que se plasmaría en la elaboración por los alumnos de un diccionario bilingüe español-portugués de locuciones, actividad didáctica que, realizada con más o menos rigor, resulta muy común en el ámbito de la enseñanza-aprendizaje de la L2.

2. LAS UNIDADES FRASEOLÓGICAS SELECCIONADAS

Los fraseologismos escogidos para examinar las posibilidades de *Linguee* se han distribuido en dos grupos claramente diferenciados. El primero comprende ocho locuciones verbales, concretamente: *caerse abajo*, *echar abajo*, *ir para abajo*, *ir para arriba*, *irse abajo*, *irse arriba*, *venirse abajo* y *venirse arriba*. Desde el punto de vista formal, participan de elementos idénticos y semejantes: los verbos *caerse*, *echar*, *ir*, *irse* y *venirse*, más los adverbios *arriba* o *abajo*. Además, para algunas de estas locuciones existe una combinación de palabras homónima no fraseológica (*caerse abajo*); otras constituyen variantes no marcadas (*irse abajo* ~ *venirse abajo*); y entre algunas se establece la relación semántica de antonimia (*ir para abajo* / *ir para arriba*).

Junto a ello, conviene indicar que las ocho locuciones están vinculadas por el mismo proceso cognitivo de formación. En efecto, la aparición de los adverbios *arriba* o *abajo* determina la existencia del esquema de imagen VERTICALIDAD, que presenta un carácter axiológico inherente; dicho de otra manera, los dos polos opuestos, ARRIBA y ABAJO, tienen distinto valor, en el sentido de que uno es positivo y el otro negativo, dada la tendencia a asociar la orientación ARRIBA con aspectos positivos como el control o el poder, frente a ABAJO, que suele conllevar los aspectos negativos contrarios a los anteriores. El esquema de imagen VERTICALIDAD, subsidiario de CAMINO, y la orientación ARRIBA-ABAJO con la que se asocia quedan ilustrados en los usos lingüísticos de las locuciones verbales tomadas como referencia, en las que las metáforas orientacionales CRECER o TOMAR ÍMPETU ES IR ARRIBA y DECRECER, HUNDIRSE o PERDER ÍMPETU ES IR ABAJO relacionan el dominio de la cantidad o el dominio mental del estado psicológico con el dominio físico de la posición.

El segundo grupo incluye seis locuciones adverbiales: *a dos velas*, *de capa caída*, *hasta el gorro*, *hasta las narices*, *por si las moscas* y *viento en popa*. Todas tienen la particularidad de ser idiomáticas, dos de ellas, además, son variantes con la misma marcación diafásica de informal: *hasta el gorro* y *hasta las narices*, y han sido obtenidas del vaciado de las locuciones registradas en un manual relativamente reciente: *Español ELElab*, del nivel B2, que sigue un enfoque comunicativo y basado en la acción (Prieto de los Mozos, Delgado Fernández, Escandell Montiel y Ghezzi, 2013).

3. PRESENTACIÓN DE *LINGUEE*

Linguee, según los datos aparecidos en su página web (<http://www.linguee.es/espanol-ingles/page/about.php>), es la combinación de un diccionario y un buscador; más concretamente, es una aplicación informática, fundada y dirigida por Gereon Frahling, que

permite buscar palabras y expresiones, en un par de lenguas previamente seleccionado, en miles de millones de textos bilingües. La búsqueda en *Linguee* se inicia con la introducción de la palabra o el término que se desea buscar en la otra lengua elegida. Los resultados de las búsquedas realizadas están estructurados de la siguiente manera. En la parte superior aparece la traducción o las traducciones resultantes, que proceden del diccionario que incluye *Linguee*, creado en colaboración con editores profesionales que trabajan constantemente para añadir nuevas entradas y mejorar su calidad. Los resultados aportan una visión general de las traducciones relacionadas con el término de búsqueda. Así, por ejemplo, para los pares de lenguas español-inglés y español-portugués se ofrece la siguiente información sobre la locución adverbial española *en un santiamén*:

en un santiamén *adverbio*
in no time *adv*
in a flash *adv* at the drop of a hat *adv*

en *prep* em *prep*
un *art* um *art*

La diferencia es notable, y la razón estriba en que el diccionario español-portugués, frente al de español-inglés, está, como se indica en la página web de *Linguee*, en elaboración, lo que no impide que pueda ser usado en el ámbito de la traducción y de la didáctica de lenguas.

A continuación de los datos anteriores, se muestran textos en la lengua origen junto a su traducción en la lengua meta elegida, procedentes de fuentes externas, para mostrar cómo se traduce el término buscado en un contexto determinado. Estos ejemplos de traducciones proceden de páginas web bilingües, especialmente de empresas, organizaciones internacionales, la Unión Europea o universidades, traducidas por profesionales; además, debajo de cada traducción, se muestra un vínculo que lleva al sitio web del que se ha extraído el texto, lo que confiere a esta herramienta un criterio de fiabilidad añadido:

Si el operador cree que el precio **irá para abajo**, se compra una opción de venta. *binaryoptionstradingguide.es*	Se o comerciante acreditar que o preço **vai descer**, uma opção de Put deverá ser então adquirida. *binaryoptionstr...ingguide.com.pt*

Por otra parte, los resultados que se encuentran a la espera de validación se indican con un triángulo amarillo de aviso:

⚠ Europa está **a dos velas**
europarl.europa.eu

⚠ A Europa está **encalhada**.
europarl.europa.eu

⚠ Esto significa que la lucha va **viento en popa:** tanto mejor.
europarl.europa.eu

⚠ Quer isso dizer que o combate **ganha terreno** e ainda bem.
europarl.europa.eu

Linguee se ha caracterizado, asimismo, como un corpus paralelo multilingüe por compilar gran cantidad de textos en distintas lenguas, junto con sus traducciones; en cualquier caso, sí conviene indicar que se ha examinado más su potencial como instrumento para el traductor o para la enseñanza de la traducción y, en menor medida, su aplicabilidad al proceso de enseñanza-aprendizaje de lenguas extranjeras, el caso que, justamente, nos ocupa aquí.

3.1. Ventajas

Desde la perspectiva de su creador, y en tanto que diccionario, *Linguee* presenta numerosas ventajas en comparación con otros diccionarios en línea:

- Permite búsquedas extremadamente rápidas: las traducciones aparecen a medida que se escribe.

- Se registran aproximadamente 1000 veces más traducciones.

- Constituye un diccionario de calidad revisado por el equipo de redacción de *Linguee*.

- Ofrece traducciones para cada término con una estructura clara e intuitiva.

- Proporciona las traducciones más frecuentes destacadas en negrita.

- Muestra traducciones contextualizadas en una gran variedad de textos.

- Presenta los textos acompañados de las fuentes de las que proceden.

Desde el punto de vista del uso de esta herramienta para el aprendizaje de una lengua extranjera, se ha subrayado:

- Su potencial para motivar a la llamada *generación internet*.
- Su alta accesibilidad.
- Su facilidad para la interacción del usuario con la propia herramienta.
- Su configuración, que le permite al usuario observar y analizar las unidades objeto de estudio en ejemplos de uso extraídos de textos auténticos, lo que facilita la comprensión de las unidades.

Si bien, tales aspectos positivos han sido matizados por la necesidad de que el profesor aconseje sobre su uso y lo facilite en la propia aula.

4. TAREA DE APRENDIZAJE PROPUESTA

4.1. Objetivos

La propuesta de enseñanza-aprendizaje de las unidades fraseológicas, específicamente locuciones, que aquí se plantea tiene como objetivos:

- Potenciar en los alumnos estrategias de aprendizaje, como el análisis de unidades y el razonamiento inductivo, para facilitar el proceso de comprensión y memorización de las unidades fraseológicas.

- Fomentar el uso de herramientas informáticas, como *Linguee*, para su aprendizaje.

- Promover la reflexión sobre la forma y el significado de los fraseologismos, atendiendo a su combinatoria sintáctica y a su valor pragmático, en suma, desde el punto de vista que ofrece su uso en el discurso.

De este modo, se espera que, una vez realizada la tarea que se le pide al estudiante, este sea capaz de:

- Utilizar *Linguee* como una herramienta de investigación porque facilita información sobre el potencial comunicativo de las unidades fraseológicas estudiadas.

- Emplear *Linguee* como instrumento de aprendizaje para obtener la competencia pertinente sobre el funcionamiento de los fraseologismos objeto de estudio.

- Aplicar las estrategias adquiridas con el uso de *Linguee* tanto a contextos de aprendizaje como a contextos profesionales, como el de la traducción en el caso de alumnos de ELE con un nivel C2 e interesados por este ámbito profesional.

4.2. Diseño

La propuesta didáctica consiste, concretamente, en que los estudiantes de un curso de ELE, de nivel C1 o C2, elaboren un diccionario bilingüe constituido por las unidades fraseológicas que aparezcan en el manual con el que estudian o que les hayan sido presentadas previamente en clase. El trabajo se desarrollará a partir de la consulta de diccionarios monolingües y bilingües, y con la ayuda del recurso electrónico *Linguee*, que puede propiciar informaciones relevantes que permitirían al alumno conocer el potencial discursivo y pragmático de las unidades estudiadas y de sus equivalentes de traducción en portugués, en este caso lengua meta. En este sentido, creemos que *Linguee* puede considerarse un valioso recurso, puesto que el tipo de información que ofrece no se encuentra disponible en los diccionarios bilingües, lo que restringe enormemente su utilidad en cuanto al estudio de la fraseología se refiere.

La propuesta tiene un carácter longitudinal, en el sentido de que, para ser efectiva, debe desarrollarse a lo largo de todo el curso. Esto significa que es necesario implementarla los primeros días de clase, empezando con su presentación al grupo de alumnos, que deben ser informados del tipo de trabajo que van a realizar y de los objetivos que se pretende alcanzar. Tras conocerla, los estudiantes deben recibir instrucciones explícitas sobre:

- Qué tipos de unidades van a recopilar y cómo deben organizarlas. Sugerimos que las locuciones sean incluidas en listados, ordenadas alfabéticamente, y que se registren en un espacio virtual (Wiki o Edmodo, por ejemplo), a fin de que todo el grupo tenga acceso a ellas. Con el objetivo de organizar la tarea, se debe asignar a cada alumno o grupos de alumnos las unidades con las que van a trabajar.

- Qué tipo de información deben recabar sobre cada una de las unidades analizadas. Para el español, esta información estaría constituida por el lema de la unidad, su clasificación categorial, su marcación diafásica, si es el caso, su combinatoria sintagmática, su definición y ejemplos de uso. Para el portugués, en cambio, solo se incluirían los equivalentes de traducción de la unidad, acompañados de ejemplos de uso. Una muestra del resultado de la tarea serían las siguientes entradas, no dispuestas de forma tradicional, pero donde se especifica el lema de la locución española; su clasificación gramatical; la marcación diafásica, es su caso, la combinatoria sintagmática; la definición; un ejemplo de uso; los equivalentes de traducción al portugués más frecuentes a partir de los datos obtenidos de *Linguee* (obsérvese que para la primera entrada los dos primeros equivalentes son unidades léxicas y el tercero una locución); y ejemplos de uso de los respectivos equivalentes:

echar abajo.
loc. tr.
[alguien/algo (nombre abstracto) *echa abajo* algo (nombre abstracto)].

Hacer fracasar un proyecto o un asunto.

he lamentado de nuevo todas las artimañas que han escenificado aquellos que querían impedir la votación porque quieren sencillamente *echar abajo* la sesión del viernes en Estrasburgo.

derrumbar.
O movimento xiita pretende *derrubar* o governo composto de cristãos, de sunitas e de drusos.
quebrar.
Como membros da Família da Luz, vocês podem simplesmente observar isto, sabendo que o caos e a confusão devem vir para *quebrar* o sistema para que ele possa ser reconstruído com luz.
deitar abaixo.
loc. tr.
Rauf Denktash e o seu círculo têm uma pesada responsabilidade por este fracasso, e compreendo a frustração e a cólera de uma grande parte da sociedade civil cipriota turca, assim como dos partidos da oposição. Evidentemente que as duas partes se comprometeram a continuar a negociações com vista a chegar a uma solução global antes do final do mês de Fevereiro; evidentemente que a possibilidade de *deitar abaixo* o último muro ainda de pé na Europa continua em aberto. Mas o betão presente nas cabeças dos privilegiados do regime de Rauf Denktash não desaparecerá como que por encanto pois, para eles, a ausência de solução é a solução.

viento en popa.
loc. adv.
infor.

De manera halagüeña o satisfactoria.

A pesar de ello, el diario reconoce textualmente que la economía india va *viento en popa*, el crecimiento puede sobrepasar este año el 7 %.

Se construye generalmente con los verbos *ir* o *marchar*.

Confío en que marche *viento en popa* en los próximos diez años y podamos acoger a nuevos miembros.

de vento em popa.
loc. adv.

Apesar disto, o jornal reconhece explicitamente que a economia indiana vai *de vento em popa*, o crescimento pode ultrapassar este ano os 7%.

- Cómo acceder a esta información, tanto a partir de la consulta de obras lexicográficas monolingües y bilingües como a partir de la consulta de *Linguee*. Así, por ejemplo, para las unidades anteriores es conveniente examinar los datos aportados por estos diccionarios:

DRAE:

> **echar.** […] **29.** Derribar, arruinar, asolar. *Echar abajo*, *en tierra*, *por el suelo.*

DFDEA:

> **echar abajo.** *v* Derribar, o hacer caer al suelo. ‖ CBonald *Noche* 20: Quiso que se preservaran con absoluto respeto los artesonados .., dejando solo que .. se echase a bajo algún tabique de mampostería. Mendoza *Tocador* 135: –¿Quién? –La policía .. Abre ahora mismo o echamos abajo la puerta y la escalera. **2 echar abajo.** *v* Derribar, o hacer caer. *En sentido no físico.* ‖ M. Abizanda *Sáb* 28.12.74, 65: Cada uno de ellos .. aportó ideas y planteamientos inéditos, echando abajo cuanto había construido el anterior. Pombo *Cielo* 77: Era fácil desde la posición de Leopoldo echar abajo periodismo [como carrera] o cualquier otra cosa.

DILEA:

> **echar abajo.** *tr.* [alguien/algo, algo] Hacer fracasar un proyecto o un asunto: –Hombre, es verdad que te afectan las críticas. Cuando te has tirado meses preparando un trabajo y, de repente, sale alguien que te lo *echa abajo*, pues dices que qué pasa; El dirigente palestino exigió que se ponga fin a las tropelías de los colonos, que según expresó «pueden *echar abajo* el proceso de paz».

AURÉLIO

> **Deitar abaixo.** 1. Deitar por terra.

HOUAISS

> **deitar abaixo.** Realizar a destruição de; deitar por terra; aniquilar. Ex.: a revolução deitou abaixo as antigas instituições.

Con este tipo de trabajo didáctico, es posible activar dos de las estrategias de aprendizaje que suelen utilizar los estudiantes de LE: el razonamiento y el análisis. Como ha propuesto Oxford (1990), los aprendices recurren a ellas al entrar en contacto con una nueva información y tienden a generar mentalmente un modelo formal basado en el análisis y la comparación con lo que ya saben o conocen; a crear reglas generales; y a revisarlas cuando la nueva información está ya disponible. Esta propuesta fomenta, asimismo, dos de las subestrategias que conforman estas estrategias de aprendizaje: el análisis contrastivo y la traducción, dado que incorpora al proceso de aprendizaje de las UF la consideración de la lengua materna del estudiante.

5. CONCLUSIONES

Como primera conclusión, es necesario mencionar que no todo son aspectos positivos en el uso de *Linguee*. El examen de las unidades seleccionadas ha mostrado que, en algunos casos, los resultados obtenidos, en cuanto al número de textos en que son objeto de traducción, son escasos. Así por ejemplo, de *por si las moscas* no se consigue ninguna ocurrencia y de *hasta el gorro* y *hasta las narices* solo una. La explicación está en el tipo de textos que constituyen esta herramienta informática: producidos por empresas, organizaciones internacionales, la Unión Europea o universidades, ámbitos poco proclives, en su producción textual, al uso de locuciones adverbiales como las señaladas, propias, más bien, del registro informal de la lengua.

Pero no solo la vinculación del tipo de texto a la situación comunicativa determina la falta de registros de ciertas unidades fraseológicas, también el género textual incide en este mismo aspecto. De este modo, de las locuciones *irse arriba* e *irse abajo* es relativamente fácil obtener, mediante el buscador Google, ejemplos en español insertos en textos periodísticos que relatan crónicas de corridas de toros o de competiciones deportivas, género textual que no parece tener acomodo en *Linguee*, lo que merma sus posibilidades de búsqueda y la

capacidad, por parte de los alumnos, de marcar diafásicamente las locuciones que lo requieran.

También la marcación diatópica, de registrarse, resultaría compleja, pues no hay manera de saber la variedad dialectal de las lenguas española y portuguesa a la que pertenecen los textos obtenidos, a menos que se recurra a la propia competencia del investigador. A ello habría que añadir que, para el caso del par de lenguas español-portugués, el diccionario todavía se encuentra en fase de elaboración.

Con todo, si consideramos la cantidad y la calidad de la información puesta al alcance del usuario, constatamos que *Linguee* constituye un instrumento interesante y útil. Y esta sería la segunda conclusión. Así, desde la perspectiva de la lengua española, el estudiante puede llegar a establecer el lema de la locución y su combinatoria sintagmática; asimismo, puede confrontar las definiciones de la unidad española que aparecen en los diccionarios de esta lengua, con los ejemplos de uso españoles para confirmar cuál o cuáles son más apropiadas al uso lingüístico. En relación con la parte en lengua portuguesa, hemos podido constatar que algunas de las locuciones que aparecen en los resultados de traducción para el portugués no están incluidas en los diccionarios bilingües portugués-español, aunque sí están registradas en diccionarios monolingües del portugués o aparecen en corpus de esta lengua. Por otra parte, hemos podido comprobar que los equivalentes suministrados, para los ejemplos consultados, son fiables e idóneos, lo que convierte a *Linguee* en un recurso electrónico eficaz. Y, además, se puede percibir la variedad de equivalentes de traducción que existen para una locución española: unidades léxicas, colocaciones y locuciones de la lengua portuguesa.

La tercera y última conclusión tiene que ver con la extensión de *Linguee* al ámbito de la praxis lexicográfica. Gracias a su doble faceta de diccionario y corpus paralelo, parece poder contrarrestar uno de los principales problemas achacados a la lexicografía bilingüe, el relacionado con los equivalentes lexicográficos. Como se sabe, la mayoría de los diccionarios bilingües se limita a ofrecer una lista indiscriminada de equivalentes que, por su falta de precisión, pueden inducir al usuario al equívoco y a la confusión. En este sentido, podemos decir que quien consulte *Linguee*, en lugar de encontrar meros equivalentes lexicográficos, establecidos en un nivel léxico, podrá conocer equivalentes funcionales o situacionales, de acuerdo con el contexto discursivo en que estén situadas las unidades. De este modo, le resultará más fácil reflexionar sobre la adecuación de los equivalentes encontrados. En definitiva, podrá constatar si las unidades léxicas equivalentes cumplen las mismas condiciones textuales o si se pueden emplear en situaciones similares.

Y ya para finalizar. Toda propuesta didáctica debe ser lógicamente experimentada para no quedar en el limbo de lo probable, alejada de todo punto del mundo real. Tal posibilidad existe, pues se dan las condiciones para que la propuesta sea llevada a la práctica por tres profesores en Brasil, en diferentes contextos y niveles de enseñanza. En el primero de ellos, se trabajará con alumnos que estudian español en la universidad, en un curso de formación para profesores de ELE; en el segundo, con alumnos que estudian español como lengua instrumental, en un curso técnico que se realiza una vez concluida la enseñanza secundaria si no se desea seguir estudios universitarios; finalmente, en el tercer contexto, se trabajará con alumnos que estudian español como lengua obligatoria en la etapa educativa de enseñanza media.

Bibliografía

ALONSO JIMÉNEZ, E., 2013. *Linguee* y las nuevas formas de traducir. *Skopos*, 2, pp. 5-28.

BUYSE, K. Y VERLINDE, S., 2013. Possible Effects of Free on Line Data Driven Lexicographic Instruments on Foreign Language Learning: The Case of *Linguee* and the *Interactive Language Toolbox*. *Procedia - Social and Behavioral Sciences*, 95, pp. 507-512.

COLL, C., 2004. Psicología de la educación y prácticas educativas mediadas por las tecnologías de la información y la comunicación: una mirada constructivista. *Sinéctica*, 25, 1-24, [en línea] Disponible en: http://www.virtualeduca.org/ifd/pdf/cesar-coll-separata.pdf [Fecha de consulta: 09/03/2015].

CORPAS PASTOR, G., 2013. Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. En: I. Olza y E. Manero Richard, eds. *Fraseopragmática*. Berlin: Frank & Timme. pp. 335-373.

DURÁN MUÑOZ, I., 2011. Recursos electrónicos para la búsqueda terminológica en traducción: clasificación y ejemplos. *Revista Tradumàtica*: *tecnologies de la traducció*, desembre, pp. 137-146.

OXFORD, R., 1990. *Language Learning Strategies*: *What Every Teacher Should Know*. Boston: Heinle.

PATIN, S., 2014. Del uso de los córpora paralelos en la enseñanza de la traducción: el caso del Europarl. En: F. Olmo Cazevieille y J.-M. Mangiant, eds. *II Coloquio franco-español de análisis del discurso y enseñanza de lenguas para fines específicos*. *Lenguas*, *comunicación y tecnología digitales*. Valencia: Universidad Politècnica de València. pp. 159-174.

PRIETO DE LOS MOZOS, E. (dir.), DELGADO FERNÁNDEZ, R., ESCANDELL MONTIEL, D. Y GHEZZI, M., 2013. *Español ELElab B2*. Salamanca: Ediciones Universidad de Salamanca.

SOLANO RODRÍGUEZ, M.ª Á., 2012. Fraseodidáctica basada en tecnologías digitales. En: Mª I. González Rey, ed. *Unidades fraseológicas y TIC* (Biblioteca Fraseológica y Paremiológica, Serie «Monografías», n.º 2). Madrid: Instituto Cervantes, Centro Virtual Cervantes. pp. 167-186.

VOLK, M., GRAËN, J. Y CALLEGARO, E., 2014. Innovations in Parallel Corpus Search Tools. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (*LREC-2014*), [en línea] Disponible en: http://dx.doi.org/10.5167/uzh-97282 [Fecha de consulta: 09/03/2015].

Diccionarios

AURÉLIO: FERREIRA, A. B. de H., 2004. *Novo Dicionário Eletrônico Aurélio* (versão 5.11a). Rio de Janeiro: Positiva.

DFDEA: SECO, M., ANDRÉS, O. Y RAMOS, G., 2004. *Diccionario fraseológico documentado del español actual. Locuciones y modismos españoles*. Madrid: Aguilar.

DILEA: PENADÉS MARTÍNEZ, I., en preparación. *Diccionario de locuciones idiomáticas del español actual.*

DRAE: REAL ACADEMIA ESPAÑOLA, 2014. *Diccionario de la lengua española*. 23.ª ed. Barcelona: Espasa Libros.

HOUAISS: HOUAISS, A., 2008, *Dicionário eletrônico Houaiss da língua portuguesa* (versão 2.0.1). Rio de Janeiro: Objetiva.

Linguee: *Linguee*, *diccionario inglés-español con mil millones de traducciones disponibles*, [en línea] Disponible en: http://www.linguee.es/espanol-ingles/page/about.php [Fecha de consulta: mayo-junio de 2015].

PHRASEOLOGY IN E-LEXICOGRAPHY AND E-TERMINOLOGY

LA INFORMACIÓN FRASEOLÓGICA EN LA LEXICOGRAFÍA Y LA TERMINOLOGÍA ELECTRÓNICAS

PHRASEOLOGY - CULTURAL CODE OF ETHNICITY
(On the material of French, English, and Georgian languages)

Tsiuri Akhvlediani
Professor-Emeritus
Tbilisi State University
tsiuriakhvlediani@yahoo.com

George Kuparadze
Associate Professor
Tbilisi State University
gkuparadze@yahoo.com

Abstract

The language is considered to be a cultural code of ethnicity; an informative intermediary with a national marker contemplating feelings, perception and presentation of the universe.

It always identifies the nation's characteristic features and culture. National-cultural signs of phraseology are formed by: *1. Peculiarities of linguo-creative thinking; 2. Ethno-lingual specific interpretation of the universe; 3. Secondary concepualisation and categorisation peculiarities of images reflected in the human consciousness based on the status of objects so important to the given ethnicity.*

It is known that a language, especially its lexicon, influences the speakers' cultural patterns of thought and perception in various ways.

The aim of the presented paper is to explore the ethno-cultural dimension of a wide range of pre-constructed or semi-pre-constructed word combinations focusing on the scope of their diachronic evolution in French, English and Georgian languages.

The corpora for investigation includes multiword units of the *kick-the-bucket* type, collocations, irreversible binominals, phrasal verbs, compounds, metaphorical expressions, similes, proverbs, familiar quotations, etc – all of which have been subsumed under phraseology.

On the basis of analysis of conceptual content space of national character, the following features verbilized by phraseological units have been singled out: 1) Basic, 2) Ethic, 3) Aesthtic, 4) Abnormal.

1. Basic or Natural Features, that are genetic for the character, express:

- *Features of Temperament types;* they are subdivided into 4 groups:

a) phlegmatic – characterised by indifference;

b) melancholic - characterised by pessimism;

c) sanguineous – boasting a fast response;

d) choleric - characterised by nervousness and hot-temper;

- *Level of Intelect* such as the highest mental abilities; limited mental abilities;

- *Arbitrary properties*, which, at the linguistic level are associated with firmness,

determination, targeting of attack;

2. Ethical Features of an individual in ethno-culture are formed in the process of socialisation. Person's status depends on his/her relationship to labour, social environment; they are defined by moral-ehtical standards. Negative ethical assessments have such features as: cunning, hipocrisy, flattery, lies, deception, rudiness, hurshness, talkativeness;

3. Aesthetic signs characterise an individual in accordance with his/her attitude to his/her self-appearance. They (the signs) denounce careless, negligent people; they encourage elegance, refinement, sophistication;

4. Abnormal signs are ascribed to those characteristis that constitute a deviation from the norm; Their basis are either the ideas of individual rationalism or, on the contrary, an absolute indifference to everything, including self-respect as well. Abnormal signs reflect a nerrow practitzism, without any tangible benefit to the aspirations or, on the contrary, indifference to everything:

- **egocentrism, arrogance, haughtiness,** uncontrollability, pamper, excessive pride;
- **involuntary, ungracious.**
- **Absence of any positive qualities of character;** such a person is assessed as "supernagative".

Thus, the signs of character are the same for people of any nationality but they are manifested differently according to various traditions, culture, national temperaments and mentalities. The importance of vital fragments characteristic of French, English and Georgian Ethnic cultures is, by no means, defined by those of the values that are so closely interwoven in them. The scrupulous observation has also revealed that a proper understanding of language accepts cultural, social and cognitive perspectives to develop a better understanding of what we do when we talk, listen, read, write and are engaged in thinking that activates the internalized linguistic system at any level.

1. INTRODUCTION

It is a stated fact that the language, at the modern stage of development of linguistic science, does not directly reflect the natural world, but it also makes this world conceptualized from a national viewpoint; that is why the language is considered to be not only a means of communication and world perception, but also as a cultural code of ethnicity; an informative intermediary with a national marker contemplating feelings, perception and presentation of the universe. (Bolly C. 2008).

The term 'language' has different connotations in different theoretical paradigms and different everyday contexts. Beliefs about language form part of a speech community's linguistic culture. Their origins reach from mythology and religion to state-of-the-art sociolinguistic theory and are a central concern of this project because they inform revitalisation strategies and the impact of related measures within society as a whole. Like the broader concept of culture, Western understandings of language span several levels of abstraction. Structural linguists treat languages as self- sufficient systems of concept-related signs and rules that allow for a virtually indefinite amount of empirically accessible speech. (Konstanze Glaser (nee Gebel), 2002)

2. CONNECTION OF PHRASEOLOGY AND ETHNO CULTURE

In recent years, phraseology in the broad sense has become a unifying theme for an increasing number of theoretical and practical linguistic studies. Among this broad palette of investigations into the meaning, structure or use of set-phrases, cross-linguistic research is one of the major and fascinating topics. An Englishman may sleep like a dog, but Frenchman will, among other possibilities, sleep like marmot (dormer comme une marmotte), a Dutchman like a rose, a German like a stone. This list might be extended to all the languages of the world and would reveal the amazing richness and diversity of language. Is the no rhyme or reason to the unbridled imagination underlying set phrases in different languages or is it possible to discover some universal principles? Will set phrases enable researchers to gain information about the cultural patterns and ways of life prevailing in other parts of the world? Set phrases in the broad sense have been identified in many languages.

It's well known that phraseological tradition originated in Russia and Germany. As a result, Russian and German were among the first languages to be fully described from the point of view of phraseology, although the movement later extended to English, French and most European languages. There is a close link between culture and phraseology. This is best revealed by proverbs and fully idiomatic set phrases, because they tend to rely heavily on images, traditions or habits, that are characteristic of a given culture. It's no easy matter, however, to draw a line between images that are related to more or less universal aspects of the human mind, and features of the specific culture. There is also a common idiomatic heritage to all European languages, originated from biblical, Latin and Greek expressions.

2.1. Language and Culture

The proposition that there is a correlation between language and culture or culture-specific ways of thinking can be traced back to the views of Herder and von Humboldt in the late 18th and early 19th centuries. It was most explicitly formulated, however, by the German-American linguist and anthropologist Edward Sapir in various publications from 1929 onward (republished posthumously in 1949 under the title Selected writings of Edward Sapir in language, culture and personality), and in the writings of his pupil Benjamin Lee Whorf (republished posthumously in 1956 as Language, thought, and reality: Selected writings of Benjamin Lee Whorf). The Sapir-Whorf hypothesis, as it came to be called, expresses the notion that different languages lead their speakers to different conceptualizations of the same extra-linguistic reality, which seems to be most evident in the way that reality is segmented by the lexicon.

Even though few linguists would fully agree with a strict reading of the Sapir-Whorf hypothesis today, it is generally accepted that a language, especially its lexicon, influences its speakers' cultural patterns of thought and perception in various ways, for example through a culture-specific segmentation of the extra-linguistic reality, the frequency of occurrence of particular lexical items, or the existence of keywords or key word combinations revealing core cultural values. Nevertheless, the exact workings of the link between language and culture are still poorly understood. The few specific theoretical frameworks that do exist are often felt to be inadequate, and the research methodology is only insufficiently developed (it is telling, in this context, that the methods employed by Whorf in particular seem to have had serious shortcomings). (Skandera P. 2007)

L. Elmslev noted that language as a system of signs and as a solid unit, always appeared to be the key to human semantic system and psychics. It always identifies the nation's characteristic features and culture. With its help, it is possible to percieve a modern style of an ethnic group as well as to penetrate into the living events of past generations. The norms, generated by the society, are concentrated in the language of the ethnic group that provide the smooth functioning of society. Thus, the bearer of the national language, national culture and mentality tends out to be a human being (an individual).

National-cultural signs of phraseology are formed by: *1. Peculiarities of linguo-creative thinking; 2. Ethno-lingual specific interpretation of the universe; 3. Secondary concepualisation and categorisation peculiarities of images reflected in the human consciousness based on the status of objects so important to the given ethnicity.* (Bally Ch. 1962)

In any ethno-culture the character is treated as a complex, multi-unit implying both positive and negative side in itself, which in average individual are accumulated in accordance of various degree of expression.

The proposition that there is a correlation between language and culture or culture-specific ways of thinking can be traced back to the views of many scholars in the late 18th and early 19th centuries (Herder and von Humboldt). It is generally accepted that a language, especially its lexicon, influences its speakers' cultural patterns of thought and perception in various ways.

The aim of the presented paper is to explore the ethno-cultural dimension of a wide range of pre-constructed or semi-pre-constructed word combinations focusing on the scope of their diachronic evolution in French, English and Georgian languages.

Anyone who wants to understand how spoken language works has to pay attention also to a range of suprasegmental and paralinguistic features, which in particular contexts may include tempo, volume, timbre, pauses and hesitations, fillers, false starts, laughter, feedback, accompanying gestures (Crystal and Davy 1969), they must try to construe the social context and motives of the participants.

2.2. Types of Phraseological Units

The corpora for investigation includes multiword units of the *kick-the-bucket* type, collocations, irreversible binominals, phrasal verbs, compounds, metaphorical expressions, similes, proverbs, familiar quotations, etc – all of which have been subsumed under phraseology.

Phraseological meaning contains background information. It covers only the most essential features of the object it nominates. It corresponds to the basic concept, to semantic nucleus of the unit. It is the invariant of information conveyed by semantically complicated word combinations and which is not derived from the lexical meanings of the conjoined lexical components. According to the class the word-combination belongs to, we single out:

a) idiomatic meaning;

b) idiophraseomatic meaning;

c) phraseomatic meaning.

The information conveyed by phraseological units is thoroughly organized and is very complicated. It is characterized by:

a) multilevel structure;

b) structure of a field (nucleus + periphery);

c) block-schema

It contains 3 macro-components which correspond to a certain type of information they convey:

a) the grammatical block;

b) the phraseological meaning proper;

c) motivational macro-component (phraseological imagery; the inner form of the phraseological unit; motivation).

Phraseological unit is a non-motivated word-group that cannot be freely made up in speech but is reproduced as a ready made unit. Reproducibility is regular use of phraseological units in speech as single unchangeable collocations. Idiomaticity is the quality of phraseological unit, when the meaning of the whole is not deducible from the sum of the meanings of the parts. Stability of a phraseological unit implies that it exists as a ready-made linguistic unit which does not allow of any variability of its lexical components of grammatical structure (Dribniuk V.2007).

The disclosure of meaning of phraseologisms containing implicitly expressed information undergoes either by means of the semantic and etymological analysis of the whole phraseological unit or by one of its components.

Implicit information is revealed through semantic analysis of one of the phraseoleksis of the phraseological unit, which is the independent component of PhU (expressed by the noun) refers to human qualities;. e.g. in PhU (Phraseological Unit) **homme de carton ‘**frailty', the meaning of phraseoleksis: '**carton**' ***'matière assez èpaisse, faite de pâte à papier'*** contains the seme 'false' that enables in the PhU to reveal the seme: "The man who does not lead to respect.".

In **English** the similar content is provided through the PhU containing the pronoun: „He is nobody" i.e. a person who does not deserve any respect from anyone".

Georgian equivalents are: 'არარაობა' [araraoba], 'ყალბი' [kalbi], 'ფუჭი' [futshi].

In PhU: **ménager ses mots**, if we analyze the components, **ménager** –*' dire avec mésure, avec économie'* and **mots** – ***'phrase, parole',*** the following seme will be revealed though the unit: 'Wisdom, foresight'.

Georgian: 'რაიმეს გულდაგულ აწონ-დაწონვა' [raimes guldagul atson-datsonva], 'თავისი სიტყვების ოქროს სასწორზე აწონვა' [tavisi sitkvebis okros sastsorze atsonva].

The implicit information is also revealed in the process of etymological analysis of the PhU. The PhU: **être le dindon de la farce** (meaning: 'Condition of being stupid'), discloses semes "confusion", "indecision".Etymological analysis shows that the formation of PhU is linked to the origins of the images of medieval comedy - farces of "trusting fathers" with nickname: ***pères dindons*** (word for word: 'father-turkeys'), who used to turn up to be the subjects of fun and making fooled by their naughty children. It should, by no means, be noted that in **French** ethno culture turkeys have long considered to be the symbol of stupidity.

English ethnic circles recognize: „be the butt of a joke"i.e. when someone is not taken seriously.

The **Georgian**: 'hen', 'goose' and 'turkey' appear to be the etalons of foolishness and stupidity: 'ქათამივით შტერი' [katamivit shteri], 'ინდაურივით ან ბატივით ყეყეჩი' [indaurivit an bativit kekechi].

Phraseological units as indirect produced nomination marks arise in the process of human interaction with the environment, the surrounding reality and at the crossing points of contours of the human as a social being and a representative of the various reflexive activities of the certain ethnic culture. The total social gains of knowledge for the certain cultural community is generally accepted and perceived by its members.

Phraseological units of all kinds permeate everyday language use throughout all linguistic registers: in everyday speech, journalism, academic prose, literature, political or diplomatic speech and writing, etc. Phraseology plays an important psychological and cultural role in language processing and social cohesiveness. Normal use of phraseological expressions is sometimes self-conscious and creative, at other times uncertain and can also be contentious. (Robo L. 2013).

Somatisms are considered to be the universal types of phraseologisms, but the national specifics of phraseological signs containing these components is manifested on their combinative level. Phraseolexis **cœur** 'heart' appears to be the constituent element of majority of PhU-s. It is quite natural that **(cœur)** 'heart' is the symbol of kindness and PhU-s: **un homme de cœur, avoir / porter un coeur d'homme** denote 'a kind person'.

In **French** phraseological realm 'heart' may be presented as item made of different materials: - **coeur de cire** - 'Candle –hearted, generous person' who is soft and plastic, 'a golden-hearted' person', though it may also happen when this phraseolexis has some negative connotations such as: **coeur d'or; cœur de marbre / pierre** 'a marble / stone hearted' person, which means: 'cold, ruthless-hearted human being'.

In **English** the same phraseolexis appears in the following PhUs: **„ Cross One's Heart and Hope to Die"** i.e. A person who says this is promising that what he or she is saying is true; **"Wear Someone's Heart on One's Sleeve"** i.e. to show one's feelings openly; **"Young at Heart"i.e.** A person who is advanced in age yet still enjoys doing things that young people do; **"Not Have the Heart to do Something"** i.e. when he or

she is afraid or unwilling to say something that might hurt or offend another person. (Kerlynb.hubpages.com 2011).

In **Georgian:** კეთილი გულისმქონე, ოქროს გულისმქონე, გულცივ, გულქვა, სასტიკ და შეუბრალებელ პიროვნებებს [ketili gulismkone, .okros gulismkone, gultsiv, gulkva, sastik da sheubralebel pirovnebebs].

In certain circumstances people with different nationalities, speaking different languages may have general conceptual image of the universe. Human parts of the body, which function analogously, have caused they are partly identical. The Georgian language belongs to the Georgian branch of the Iberian-Caucasian family. The English Language belongs to the Indo-European language family and French presents the Roman group of languages. The creation of the whole range of phraseological units. In some cases they are identical but sometimes partially identical.

In their works researchers have found out that it is lexical-semantic factors have the crucial importance on the formation of the phraseological units and not grammatical factors. Similarity of the lexical meaning is sufficient for the existence of the phraseological universals in the non-cognate languages and their similarity of the grammatical structure is not important. From the semantic point of view, these forms are absolutely identical but the grammatical structure is different.

The phraseological unit '**heart and soul**' is very interesting. It means **"with the whole body"** or **"with the whole essence".** The human is defined as a unity or ensemble of flesh and soul and the existence of both components compile the human unity respectively. This kind of approach exists in every nation so it caused the transformation of language units and as a result of this a phraseological universal took its shape. The lexical-semantic meaning of the phraseological unit: **heart and soul** is understood in the same way in different languages and means "with all one's feelings and essence"(Tsiskarishvili M. 2010).

In any ethno culture the character is treated as a complex, multi-unit, whose constituents are the advantages and disadvantages of each person, which varies from personality to personality.

The English authors Jennifer Seidl and W.McMordie in their book "English Idioms and How to Use Them" state: "An idiom is a number of words which, taken together, mean something different from the individual words of the idiom when they stand alone. The way in which the words are put together is often odd, illogical or even grammatically incorrect. We have to learn the idiom as a whole and we often cannot change any part of it." (Seidl, McMordie 1983)

Though there are differences in opinions, all linguists agree that phraseological units or idioms are probably *"the most picturesque, colourful and expressive part of the language vocabulary, which reflect nation's customs, traditions and prejudices, recollections of its past history, scraps of folk songs and fairy tales. But it is necessary to distinguish them from other words and phrases existing in the language"* (Jansone Anna 2013).

2.3. Features and Functions Verbalized by Phraseological Units

The socially-determined reflection of objects of the real world mediated by consciousness, promoting their cognition. The social determinacy is shown in the fact that, though potential phraseological units are created by separate individuals, these individuals are part of the society, and the realization of the cognitive function by them is possible only on the basis of previous/ background knowledge. Cognitive and nominative functions are realized within the limits of the communicative function, forming a dialectic unity, and all the other functions are realized within the limits of the given functions. The hierarchy of the functional aspect of the phraseological system is seen in it (Fedulenkova T. 2013).

On the basis of analysis of conceptual content space of national character, the following features verbalized by phraseological units have been singled out: 1) Basic, 2) Ethic, 3) Aesthtic, 4) Abnormal. (Kerlynb.hubpages.com 2011).

1. **Basic or Natural Features,** that are genetic for the character, express:

- *Features of Temperament types;* they are subdivided into 4 groups:

a) phlegmatic – characterised by indifference:

French: être froid comme une carafe d'orgeat

English: Cold fish

Georgian: ცივი, გულგრილი, ფლეგმატური, მისთვის ყველაფერი სულერთია [tsivi, gulgrili, flegmaturi, mistvis kvelaferi sulertia].

b) melancholic - characterised by pessimism:

French: tomber dans l'abattement

English: Fall into despondency ; "Be down in the dumps"- Be very sad and depressed

Georgian: ცუდ გუნებაზე ყოფნა [tsud gunebaze kofna].

c) sanguineous – boasting a fast response:

French: faire qch sur-le –champ/ trois coups de cuiller à pot

English: Do sth on the spot / three shots ladle

Georgian: დაუყოვნებლივ რაიმეს გაკეთება [daukovnebliv raimes gaketeba].

d) choleric - characterised by nervousness and hot-temper:

French: avoir l'air triste comme un bonnet de nuit

Georgian: ღვარძლიანი, ფხუკიანი, ბოღმიანი [gvardzliani, fkhukiani, bogmiani].

- *Level of Intellect* such as the highest mental abilities; limited mental abilities;

- *Arbitrary properties*, which, at the linguistic level are associated with firmness, determination, targeting of attack;

2. Ethical Features of an individual in ethno-culture are formed in the process of socialisation. Person's status depends on his/her relationship to labour, social environment; they are defined by moral-ehtical standards. Negative ethical assessments have such features as: cunning, hypocrisy, flattery, lies, deception, rudiness, hurshness, talkativeness:

French: un grand abatteur de besogne;

franc comme l'osier;

droit comme un arbre;

jouer d'adresse / user de detour;

faire le chien couchant;

être secret comme un coup de canon;

il a du crédit comme un chien à la boucherie;

poli comme une porte de prison.

English: Bend over backward/backwards;

Do as much as you can to help or please someone.

Georgian: გამრჯე, ძალიან მშრომელი [gamrdje, dzalian mshromeli];

მოუსყიდავი, უანგარო, პატიოსანი [mouskidavi, uangaro, patiosani];

ქლესაობა, მლიქვნელობა [klesaoba, mlikvneloba];

ფარული, როგორც ზარბაზნის გასროლა [faruli, rogorts

zarbaznis gasrola];

გულღია, ალალი, გულწრფელი, გულმართალი [gulgia, alali, gultsrfeli, gulmartali].

3. Aesthetic signs characterise an individual in accordance with his/her attitude to his/her self-appearance. They (the signs) denounce careless, negligent people; They encourage elegance, refinement, sophistication:

French: être ficelé / fichu comme l'as de pique / un sac;

être tiré à quatre épingles;

un Beau rummel.

English: Fine and Dandy – a man who affects extreme elegance in clothes and manners; a fop.

Georgian: ცუდად და უგემოვნოდ ჩაცმული [tsudad da ugemovnod tshatsmuli];

ახალთახლებშია გამოწყობილი [akhaltakhlebshia gamotskobili].

4. Abnormal signs are ascribed to those characteristis that constitute a deviation from the norm; Their basis are either the ideas of individual rationalism or, on the contrary, an absolute indifference to everything, including self-respect as well. Abnormal signs reflect a narrow practitzism, without any tangible benefit to the aspirations or, on the contrary, indifference to everything: (Büchmann G. 1978)

- **egocentrism, arrogance, haughtiness,** uncontrollability, pamper, excessive pride;
- **involuntary, ungracious.**

French: le coq du village / etre fier comme un pou;

se rengorger comme un dindon ;

être dorloté comme un roi ;

fier comme Artaban.

Georgian: მამალივით ამპარტავანი [mamalivit ampartavani],

ინდაურივით გაფხორილი [indaurivit gafkhorili];

ეშმაკივით ამაყი [eshmakivit amaki],

ხელმწიფესავით ნებიერი [khelmtsipesavit nebieri].

- <u>Absence of any positive qualities of character</u>; such a person is assessed as "super-negative".

French: abandon de soi-même;

n'avoir ni forme, ni couleur;

compter pour de la vaseline;

Monsieur le Bon;

couille molle.

Georgian: გულგრილობა საკუთარი თავის მიმართ [gulgriloba sakutari tavis mimart];

სათვალავში არ ჩაითვლება [satvalavshi ar tshaitvleba],

დოყლაპია [doklapia], დონდლო [dondlo].

French: bijou de la foire Saint-Ovide

villain bonhomme

maivaise graine

un villain coco

Georgian: არარაობა, ცუდი, საზიზღარი ადამიანი [araraoba, tsudi, sazizgari adamiani];

უსიამოვნო ტიპი [usiamovno tipi], უზნეო ადამიანი [uzneo adamiani].

3. CONCLUSIONS

In conclusion, the signs of character are the same for people of any nationality but they are manifested differently according to various traditions, culture, national temperaments and mentalities. The most distinct and peculiar features characteristic to the given ethnicity are depicted in Phraseology.

National and international elements are closely linked to a semantic structure of phraseological units. International elements manifest themselves in human mentality as well as in interaction of languages and cultures. National elements reflect originality of speech group, caused by the conditions of its life and peculiarities of the history.

Phraseology is connected with communication, as the exchange of meanings between individuals through a common system of symbols, concerned scholars since the time of ancient Greece. (Bekmuratova N. 2011)

The importance of vital fragments characteristic of French, English and Georgian Ethnic cultures is by no means defined by those of the values that are so closely interwoven in them.

A person's understanding of their own and others' cultural identity develops from birth and is shaped by the values and attitudes prevalent at home and in the surrounding community. This identity becomes more complex and fluid over time as people develop allegiances to different groups within the broader society. At the same time, cultures themselves are not static but develop and change as the belief systems and ways of life of different groups adapt under other cultural influences including mass media and popular culture to create new identities (Cornell, S. & Hartmann, D. 1998).

Phraseological meaning is a factor of a language and not a speech. It may be discovered in units which are characterized by phraseological steadiness. The steadiness of phraseological meaning is closely connected with the steadiness of lexical structure of a unit. The inner form may be explicit and implicit and it influences the meaning of a set phrase. The most important feature of phraseologisms is fully or partly reinterpreted meaning. The nature of the meaning is reinterpreted if there is any departure from the literal meaning (Negrych N. 2013).

The scrupulous observation conducted has also revealed that a proper understanding of language takes into account and accepts that cultural as well as social and cognitive perspectives on language knowledge and use are all needed if we are to develop a better understanding of what we do when we talk, listen, read, write and are engaged in thinking that activates the internalized linguistic system at any level.

References

BALLY CH., 1962. *Traité de Stylistique française. Genève.*

BEKMURATOVA N., 2011. *Communicative Peculiarities of Phraseological Units in Political Discourse.* Drogobitski Pedagogical Institute.

BOLLY C., 2008. *Les unités phraséologiques. Louvain-la-Neuve.*

CORNELL S., & HARTMANN D., 1998. *Ethnicity and Race: Making Identities in a Changing World*, Pine Forge Press, Thousand Oaks, California.

CRYSTAL D., and DEREK D., 1969. *Investigating English Style.* Cambridge: Cambridge University Press.

FEDULENKOVA T., 2013. *English Phraseological Units and Their Constant Functions,*Materials of the conference "EDUCATION AND SCIENCE WITHOUT BORDERS", № 2

KONSTANZE G., (nee Gebel), 2002. *Abstract as many other ethno-cultural identities in Europe.* Middlesex University.

JANSONE A., 2013. *Praseological Units with the Elements Referring to 'Life' or 'Death' in English and Russian,* Daugavpils University, Latvia

NEGRYCH N., 2013. *Semantic Modulations of Phraseological Units,* Yuriy Fedkovich Chernivtsi National University.

ROBO L., 2013. *A Diachronic and Source Approach of Phraseological Units – Theories of Definition,Criteria and Structure Analysis in English and Albanian* Language,Academic Journal of Interdisciplinary Studies MCSER Publishing-Rome,Italy.

SEIDL J., MCMORDIE W., 1983. *English Idioms and How to Use Them.* – Moscow: Vyšaya škola

SKANDERA P., 2007. *Phraseology and culture in English,* Walter de Gruyter GmbH & Co. KG, 10785 Berlin

(Websites)

DRIBNIUK V., 2007. *Typology of phraseological units in English,* Chernivtsi National University, <www.rusnauka.com/10. ENXXIV 2007/Philologia/>21605.doc.htm

KERLYNB.HHUBPAGES.COM. 2011. *10 Heart Idioms Explained to English as a Second Language Learners. British Columbia.*

STUDYMODE.COM. 2012. *Cultural Components in Phraseology.* Retrieved 10. <http://www.studymode.com/essays/Cultural-Components-In-Phraseology->1162436.html

TSISKARISHVILI M., 2010. *Phraseological Universals Related to the Word "Heart" (Based on the Georgian-English-Bulgarian and Bats Languages),*

conf.uni-ruse.bg/bg/docs/cp10/6.3/6.3-13.pdf, v.49

PHRASEOLOGICAL CORRESPONDENCE IN ENGLISH AND SPANISH SPECIALIZED TEXTS

Míriam Buendía Castro
University of Malaga
mbuendia@uma.es

Pamela Faber
University of Granada
pfaber@ugr.es

Abstract

This paper describes a method of selecting correspondences between verbal phraseological units in English to Spanish within the specialized domain of the environment. The underlying idea is that verbs in specialized texts and their argument structure can be classified and organized in a set of conceptual-semantic categories typical of a given specialized domain (Buendía 2013). In this context, when semantic roles and macroroles from the Role and Reference Grammar (Van Valin and LaPolla 1997) are specified as well as the resulting phrase structure, it is then possible to establish templates that represent this meaning for entire conceptual subdomains. Accordingly, the range of verbs generally associated with a certain category can be predicted within the frame of a specialized event (Faber 2012).

1. INTRODUCTION

The selection of suitable interlinguistic correspondences at the word, phrase, and even text level represents one of the most problematic aspects of translation. Although a great deal has been written about translation equivalence, little of it is very useful. Apart from the traditional opposition of faithful/free, other term pairs such as semantic/communicative (Newmark 1981) and formal/dynamic (Nida and Taber 1969) have also been proposed as descriptions for the degree of perceived similarity at the level of form and/or function between a source language text and a target language text. However, these changes of label have not been accompanied by significant new insights into the nature of interlinguistic or intertextual equivalence. Whatever the terms used, judgments of equivalence often tend to be depressingly based on linguistic form instead of any sort of shared conceptual reference or meaning representation (Faber and Ureña 2012) despite the fact that this type of mirror-reflection equivalence is a chimera.

In terminology and specialized translation, correspondences between specialized knowledge units (SKUs) in different languages are generally established at the conceptual level. In other words, two SKUs are regarded as equivalents if their conceptual properties coincide. Syntactic correspondence is only relevant to the extent that syntax can be

regarded as a reflection of semantic properties. An SKU can be a single term or a polylexical unit, which includes verbs and verb phrases. This paper describes a method of selecting correspondences between verbal phraseological units in English to Spanish within the specialized domain of the environment.

The underlying idea is that verbs in specialized texts and their argument structure can be classified and organized in a set of conceptual-semantic categories typical of a given specialized domain (Buendía 2013). In this context, when semantic roles and macroroles from the Role and Reference Grammar (Van Valin and LaPolla 1997) are specified as well as the resulting phrase structure, it is then possible to establish templates that represent this meaning for entire conceptual subdomains. Accordingly, the range of verbs generally associated with a certain category can be predicted within the frame of a specialized event (Faber 2012). In this regard, verbs belonging to the same frame normally have the same number and type of arguments. These arguments have similar semantic characteristics and constrain verb meaning within specialized texts. This makes it possible to establish frame-based correspondences between different languages.

In our study, we performed an in-depth contrastive analysis of the verb phrases in a Spanish-English corpus of environmental specialized texts, more specifically linked to the field of natural disasters. Our target audience consisted of English and Spanish native speakers, who might find themselves immersed in a number of communicative situations in which they would have to activate terminological phrasemes: (i) translation of Spanish environmental texts into English; (ii) translation of English environmental texts into Spanish.

2. APPROACH TO THE STUDY OF COLLOCATIONS

Our approach to collocations integrates insights from both the semantically-based approach and the frequency-oriented approach to collocations. As is well-known, the semantically-based approach (Mel'čuk et al. 1984–1999, Hausmann 1989, Benson et al. 1986, 2009) conceives collocations as mainly binary units with a semantically-autonomous base and a semantically-dependent collocate. In contrast, the frequency-oriented approach to collocation (Sinclair et al. 1970/2004) conceives collocations as statistically significant co-occurrences of two or more words.

As such, in our research, the concept of collocation is defined in its broadest sense. In other words, in contrast to authors who only focus on collocations of the type 'to make the bed', we also regard as collocations, combinations as transparent as 'the fire burns'. Our study focuses on combinations formed by verb+noun or noun+verb, in which the collocate is the verb, and the noun is the base. In addition, our approach also includes a certain degree of compositionality since each lexical unit in a collocation conserves its meaning. Accordingly, in the collocation 'fire burns', the predicate 'burn' only admits noun phrases designating combustible entities, whereas 'fire' requires a verb designating a combustion process (e.g. 'burn'). Finally, in order for a multi-word unit to be a collocation, it should also have a high frequency in texts that activate the same pragmatic context or situation.

3. THE SCOPE OF THE STUDY: THE ENVIRONMENTAL KNOWLEDGE BASE ECOLEXICON

EcoLexicon represents the conceptual structure of the specialized domain of the Environment in the form of a visual thesaurus in which environmental concepts are configured in semantic networks. The various terminological designations for a concept are offered in six languages: Spanish, English, German, French, Russian, and Modern Greek. In EcoLexicon it is assumed that up to a certain level, its potential users are familiar with scientific language and its usage in English or Spanish, since these are the interface languages. Potential users should thus possess a good command of any of the six languages in the knowledge base, as well as a minimum level of scientific knowledge (López, Buendía, and García 2012: 62).

EcoLexicon provides both conceptual and linguistic information. Conceptual information is reflected in EcoLexicon in four ways: (i) the semantic network for each concept, which is based on a closed inventory of semantic relations; (ii) the terminographic definition of the concept, which encodes the most basic conceptual relations in the category template; (iii) association with conceptual categories, which makes it possible to access the classes of the ontology to which the search concept belongs; (iv) resources, (images, documents, URLs, audiovisual material, etc.) associated with each concept.

EcoLexicon provides all the terms designating each concept in the six languages contained in the knowledge base. In addition, for each term, it offers the following linguistic information: (i) morphosyntactic information regarding grammatical category (noun, verb, adjective or adverb); term type (main term, synonym, geographical variant, and acronym); gender (masculine, feminine, and neuter); (ii) contexts of use; (iii) phraseology.

We decided to focus on verb collocations associatedwith terms since recent studies have highlighted the importance of verbs in specialized terminology (Lorente 2007, Lopez 2007, Buendia 2012, 1013, Buendia, Montero and Faber 2014), and there are currently few terminographic resources that incorporate them since terminology has generally played down the description of verbs (L'Homme 1998, Buendia 2012).

This paper continues the research begun in Buendía, Montero and Faber (2014), which proposes a methodology for classifying and describing English verb collocations in EcoLexicon according to meaning. We now target the establishment of English-Spanish equivalents within the specialized domain of the environment and more specifically within the subdomain of natural disasters.

3.1. The corpus

For our analysis, we used the English corpus of texts on natural disasters compiled in Buendía, Montero and Faber (2014). In addition, we compiled a corpus in Spanish with the same characteristics . The Spanish subcorpus included 44 texts with 449,416 tokens and 31,230 types. Although the initial objective was to compile subcorpora with a similar number of words for each language, in the end, the final size of the corpus was conditioned by the availability of texts, based on the selection criteria. This was particularly true of the Spanish subcorpus. Since most of the textbooks in the domain were written in English, it was fairly easy to retrieve textbooks on natural hazards in English in machine-readable format. However, it was considerably more difficult to find textbooks in Spanish in this same format. Therefore, the textbooks in Spanish had to be scanned and converted to txt format with an OCR. They also had to be manually revised. For this reason, it was not possible to have two subcorpora of the same size, and this also explains why the Spanish corpus was smaller than the English one.

The type/token ratio (TTR) and the standardized type/token ration (STTR), obtained with WordSmith Tools express the relation between the total number of tokens and the total number of different types found in a text. They provide valuable information concerning the lexical diversity of the texts. Of the two numbers, the most interesting and reliable for comparing texts of different size is the STTR. In this regard, in the subcorpus in English, there are on average 40.61 different words for every text sequence of 1,000 tokens, whereas for the subcorpus in Spanish the average is slightly higher (41.28). These data provide evidence that even though the English subcorpus is larger than the Spanish, the two corpora are comparable in lexical diversity.

4. ANALYSIS

4.1. Top-down analysis

The Spanish subcorpus was analyzed n line with the analysis described in Buendía, Montero and Faber (2014). First, candidate verbs were retrieved from the corpus with the term extractor, TermoStat[117], a tool developed by Drouin at the University of Montreal, which gave us 323 verbs. Then, these verbs were classified in lexical domains as proposed by the Lexical Grammar Model (Faber 2009), based on their definitions. The classification of verbs in domains and subdomains was based on the definition of the verbs. Definitions were formulated by consulting different dictionaries and other reference material, according to the Lexical Grammar Model and Frame-based Terminoloy (Faber 2009, 2011, 2012). Since each subdomain focuses on a particular area of meaning, this means that all verbs within a subdomain can be defined either directly or indirectly in terms of the same hypernym. As such, each verb has a *genus* that corresponding to the superordinate verb in the hierarchy. In this regard, the more specific the hierarchy becomes, the more specialized the information will be. Table 1 displays the lexical domains of the LGM:

- (i) *to be* [existence]
- (ii) *to become different* [change]
- (iii) *to have/give* [possession]
- (iv) *to say* [speech]
- (v) *to feel* [emotion]
- (vi) *to do/make* [action]
- (vii) *to use* [manipulation]
- (viii) *to know/think* [cognition/mental perception]
- (ix) *to move (go/come)* [movement]
- (x) *to become aware* (*notice/perceive)* [general perception]
- (xi) *to see/hear/taste/smell/touch* [sense perception]
- (xii) *to be/stay/put* [position]

Table 1. Lexical domains in the verbal lexicon

Our membership criteria were established according to the most prototypical meaning of the verb within the context of natural disasters. Finally, the meanings of the verbs were refined by studying their activation in texts as reflected in concordances. More specifically, both arguments and predicates were identified and analyzed. This allowed us to discard those verbs that were not directly associated with any kind of NATURAL HAZARD, i.e.

[117] http://termostat.ling.umontreal.ca/

verbs that did not have an atmospheric disturbance or an associated phenomenon as one of their arguments. Table 2 displays the verbs activated in Spanish by the subdomain *to come against sth with sudden force* within the domain of ACTION. As shown, it includes verbs such as *golpear, impactar, batir, azotar, sacudir, chocar, colisionar.*

To come against sth with sudden force (Spanish)
golpear: *producir choque repentino y violento de un cuerpo contra otro.*
impactar: *golpear violentamente.*
batir: *golpear.*
azotar: *(referido especialmente al viento y demás fenómenos atmosféricos), golpear repetida y violentamente, produciendo daños o destrozos.*
sacudir: *azotar.*
chocar: *(referido a un cuerpo), encontrarse violentamente con otro.*
colisionar: *(referido especialmente a un vehículo), chocar violentamente con otro.*

Table 2. Definition of the Spanish verbs for the subdomain *to come against sth with sudden force*

4.2. Bottom-up analysis

For each verb, we analyzed the total number of concordances retrieved in our corpus. All the concordance lines for each verb provided by TermoStat were recorded manually in an Excel file. Subsequently, the arguments of each verb were identified and assigned a semantic label. Then, all the verbs having the same semantic categories were grouped together and assigned a name or phraseological pattern. The semantic categories identified for the specialized field of natural disasters were the following: NATURAL DISASTER, ATMOSPHERIC AGENT, WATER AGENT, ATMOSPHERIC CONDITION, MATERIAL ENTITY, AREA, CONSTRUCTION, ENERGY, HUMAN BEING, LANDFORM, WATER COURSE, DEATH, DAMAGE, LOSS OF LIFE/PROPERTY, PLANT, and EXPLOSIVE.

This analysis of the Spanish subcorpus was performed using the same methodology in Buendía, Montero and Faber (2014). In other words, once the semantic categories in each Spanish concordance line were identified, they were directly associated with the phraseological pattern in English which had the same category labels. Consequently, it can be said that the Spanish analysis was somewhat easier. This methodology is in consonance with that of Pimentel (2012) in her assignment of specialized verb equivalents within the legal domain. In line with Pimentel (2012), verbs that share the same type and number of semantic tags in the two languages were potential correspondences for each other.

The set of semantic roles in our study largely coincides with the most general thematic relations provided by Role and Reference Grammar (Van Valin and LaPolla 1997). The inventory used was the following: AGENT, NATURAL FORCE, DESTINATION, EXPERIENCER, FREQUENCY, GEOGRAPHICAL LOCATION, MANNER, PATH, PATIENT, SITUATION/ EXPERIENCE, ORIGIN, THEME, TIME, RESULT. Along with semantic categories and semantic roles, an additional categorization in terms of macroroles, as proposed by RRG (i.e. ACTOR and UNDERGOER), was also provided.

In our template proposal, the morphosyntactic structure associated with each argument has been specified. In this regard, the following phrases were distinguished: (i) noun phrase (NP), for phrases having a noun as the head of the phrase and which function as a noun in the sentence; (ii) prepositional phrases (PP), for phrases headed by a preposition, specifying the specific preposition(s) activated; (iii) adverbial phrases (AVP), for those phrases whose head is an adverb and fulfills the function of an adverb in the sentence.

5. ESTABLISHMENT OF EQUIVALENTS

We strongly believe that when various linguistic realizations refer to the same argument, then they designate the same kind of entity, evoke the same kind of conceptual structure, and have a similar semantic and syntactic behavior. Semantic categories are generalizations for a set of terms that are assumed to have a similar semantic and syntactic behavior. In specialized language, verb meaning is more restrictive because of the constraints of specialized subject fields. Consequently, if arguments are classified and structured in a set of conceptual-semantic categories typical of a given domain, along with the semantic roles activated, the range of verbs generally associated with a certain category could be predicted within the framework of a specialized event.

In this regard, following with the example of the subdomain *to come against sth with sudden force,* after the analysis described in Section 4, it was proved that it generally has two arguments: (i) a natural disaster (either associated with water or wind) which fulfils the semantic role of natural force and the macrorole of actor and which takes the form of noun phrase; (ii) the a construction, area or human being with the role patient, macrorole of undergoer in a noun phrase. As such, whenever a verb, either in English or Spanish, fulfils these requirements, it will presumably activate the meaning of this subdomain. Table 3 displays the phraseological pattern of the subdomain *to come against sth with sudden force.*

Subdomain: **to_ come_against_sth_with_sudden_force**			
Semantic role	Natural force	*hit* *strike* *batter**	Patient
Macrorole	Actor	*blast2*** *crash** *golpear*	Undergoer
Semantic category	natural disaster [water]* [wind]**	*impactar* *batir** *azotar* *sacudir*	area, construction, human being
Phrase type	NP	*chocar**	NP

Table 3. Phraseological pattern of the subdomain *to come against sth with sudden force*

As shown in Table 3, not all the verbs that conform to this template are synonyms and can be used in the entire set of contexts. In other words, some of the verbs impose certain restrictions on their arguments. As such, the first argument of 'batter', 'crash' 'batir' and 'chocar' is a water-based natural disaster (e.g. 'A cold wave battered Bucharest over the weekend'). In line with this, the first argument of 'blast2' is restricted to wind-related natural disasters (e.g. 'Hurricane Isaac blasts the New Orleans levee'). In addition, the second argument of verbs such as 'hit', 'strike', 'batter', 'golpear', 'azotar', and 'sacudir' is a patient, which can be a person or a construction (e.g. 'These people were battered by very recently floods', 'Some 106 million people were hit by floods and 60 million by drought in the last decade in Europe'). Finally, even though the verbs in this frame generally have two arguments, sometimes the natural force appears by itself with the verb since the second argument is implicit (e.g. Please, do not wait until a hurricane strikes to ask for help). As with the other verbs, location, time, and manner can always be included.

6. CONCLUSIONS

This paper establishes a methodology of assigning verb collocation equivalents based on meaning for English-Spanish within the domain of the environment. Regarding future research, the idea is to eventually extend this methodology to the other languages in EcoLexicon. Subdomains could thus be generalized as structures underlying interlinguistic correspondence. It would thus be possible to establish a direct correspondence between collocations in different languages. This methodology would also be extended to the other subevents in EcoLexicon. This will be possible because the corpus will have been semantically tagged. In addition, the inventory of categories for the NATURAL DISASTER is only a preliminary classification. As such, our aim is to refine the inventory of categories and at the same time, to establish a complete set of categories for the domain of the ENVIRONMENT. This categorization will ultimately lead to a system of semantic tagging, which will be a step further towards the large-scale automatic retrieval of specialized knowledge. This will used in the implementation of the incipient ontology currently under construction in Ecolexicon.

Acknowledgements

This research was carried out within the framework of project FF2014-52740-P, *Cognitive and Neurological Bases for Terminology-enhanced Translation (CONTENT)* funded by the Spanish Ministry of Economy and Competitiveness.

References

BENSON, M., E. BENSON, and R. ILSON., 1986. *Lexicographic Description of English. Studies in Language Companion Series. Volume 14*. Amsterdam/Philadelphia: John Benjamins.

BENSON, M., E. BENSON, and R. ILSON., 2009. *The BBI Combinatory Dictionary of English*. 3rd edition. Amsterdam/Philadelphia: John Benjamins.

BUENDÍA CASTRO, M., 2013. *Phraseology in Specialized Language and its Representation in Environmental Knowledge Resources*. PhD Thesis presented at the University of Granada, Spain.

BUENDÍA, M., 2012. Verb Dynamics. *Terminology*, 18(2), p. 149–166.

BUENDÍA-CASTRO, M., S. MONTERO-MARTÍNEZ, Y P. FABER, 2014. Verb collocations and Phraseology in Ecolexicon. In: K. Kuiper (ed.) *Yearbook of Phraseology*. Berlin: De Gruyter Mouton. pp. 57-94.

FABER, P., 2009. The Cognitive Shift in Terminology and Specialized Translation. *MonTI. Monografías De Traducción e Interpretación*, 1(1), pp. 107–134.

FABER, P., 2011. The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction or the Perception-action Interface. *Terminology*, 17(1), pp. 9–29.

FABER, P. ed., 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin. Boston: Mouton de Gruyter.

FABER, P, AND MAIRAL, R., 1999. *Constructing a Lexicon of English Verbs*. Berlin/New York: Mouton de Gruyter.

FABER, P., and UREÑA, J.M., 2012. Specialized Language Translation. In P. Faber, ed. 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin, Boston: De Gruyter Mouton, pp. 73–92.

HAUSMANN, F.J., 1989. Le Dictionnaire Des Collocations. In: F. J. Hausmann, O. Reichmann, H.E. Wiegand, and L. Zgusta (eds.) *Wörterbücher/ Dictionaries/ Dictionnaires — Ein Internationals Handbuch Zur Lexikographie/ An International Enyclopedia of Lexicography/ Enyclopédie Internationale De Lexicographie*. Berlin and New York: Walter de Gruyter. pp. 1010–1019.

LÓPEZ, C.I., and BUENDÍA, M., 2011. En Busca De Corpus Online a La Carta En El Aula De Traducción Científica y Técnica. *Trans-kom*, 4(1), pp. 1–22. Available at: http://www.trans-kom.eu/bd04nr01/transkom_04_01_01_Lopez_Buendia_Corpus.20110614.pdf. [Accessed 12 June 2015].

LÓPEZ, C.I., BUENDÍA, M., and GARCÍA, A., 2012. User Needs to the Test: Evaluating a Terminological Knowledge Base on the Environment by Trainee Translators. *Jostrans. The Journal of Specialized Translation*, 18, pp. 57–76. Available at: http://www.jostrans.org/issue18/art_lopez.pdf. [Accessed 12 June 2015].

LORENTE, M., 2007. Les Unitats Lèxiques Verbals Dels Textos Especialitzats. Redefinició D'una Proposta De Classificació. In: M. Lorente, R. Estopà, J. Freixa, J. Martí, and C. Tebé (eds.), *Estudis De Lingüístics i De Lingüística Aplicada En Honor De Mª Teresa Cabré Castellví. Volum 2: De Deixebles*. Barcelona: Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra, pp. 365–380. Available at: http://ricoterm.iula.upf.edu/docums/16_lorente.pdf. [Accessed 2 March 2015].

L'HOMME, M.C., 1998. Le Statut Du Verbe En Langue De Spécialité Et Sa Description Lexicographique. *Cahiers De Lexicographie*, 73(2), pp. 61–84. Available at: http://www.ling.umontreal.ca/lhomme/docs/cahiers-lexico-98.PDF. [Accessed 2 March 2015].

MEL'CUK I. et al. 1984-1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherche lexico-sémantiques I, II, III, IV*. Montréal: Les Presses de l'Université de Montréal.

NEWMARK, P., 1981. *Approaches to Translation*. Oxford: Pergamon Press.

NIDA, E.A. and TABER, C.R., 1969/1982. *The Theory and Practice of Translation*. Leiden: E. J. Brill.

PIMENTEL, J., 2012. *Criteria for the Validation of Specialized Verb Equivalents: Application in Bilingual Terminography*. Ph.D. Thesis presented at the University of Montreal, Canada. [online] Available at: http://olst.ling.umontreal.ca/pdf/Pimentel_J_thesis_2012.pdf [Accessed 22 September 2013].

SINCLAIR, J. M., JONES, S., and DALEY, R., 1970/2004. *English collocation studies: The OSTI report* (R. Krishnamurthy, ed.). London: Continuum.

VAN VALIN, R.D.JR., and LAPOLLA, R., 1997. *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.

LA WEB COMO CORPUS Y BASE DE INVESTIGACIÓN CIENTÍFICA

Heloisa Fonseca
Universidade Estadual Paulista / UNESP.
FAPESP 2014/15385-8
heloisafonseca25@gmail.com

Resumen

Este artículo pretende demostrar el potencial productivo de las investigaciones fraseológicas que se basan en la World Wide Web, con objeto de analizar las ventajas y desventajas que presenta el uso de este medio de investigación; en concreto, este trabajo se centrará en el uso de la web como corpus lexicográfico y como fuente de búsqueda avanzada, de contextos y de indicadores de uso. Además, observaremos el uso de Google como instrumento para las investigaciones fraseo-lexicográficas, especialmente para confirmar los equivalentes. Para ello, hemos analizado algunas expresiones idiomáticas y proverbios brasileños y franceses (quem não tem cão caça com gato / faute de grives on mange des merles; matar dois coelhos com uma cajadada só / courir deux lièvres à la fois; dar bode / faire long feu; estar com a macaca / avoir bouffé du lion). Los resultados obtenidos en este trabajo revelan algunos aspectos desconocidos, sobre todo culturales, ya que es posible encontrar en Internet un lenguaje sin filtros y más cercano al uso cotidiano. En este sentido, nuestro objetivo consiste en comprobar si Internet, más específicamente los buscadores automáticos, puede ayudar a establecer y confirmar equivalencias mediante contextos y, sobre todo, mediante la cantidad aproximada de resultados de las búsquedas. Por otra parte, los contextos encontrados se analizan basándonos en la metodología de la Lingüística de Corpus, Berber Sardinha (2000, 2004), Baker (1995, 1999), Tagnin (2001, 2004), y en la herramienta *Concord* del programa *WordSmith Tools*, que es un conjunto integrado de programas para observar cómo las palabras o expresiones se comportan en los textos. Creemos que es posible desarrollar una investigación científica tomando como base fundamentalmente Internet y aplicativos o demos gratuitos. A partir de los resultados obtenidos en este trabajo, podemos adelantar que el uso tanto de los presupuestos del Procesamiento del Lenguaje Natural (PLN) como de la Lingüística de Corpus ayuda a la elaboración de proyectos lexicográficos al proporcionar usos auténticos de la lengua.

1. INTRODUCCIÓN

Las investigaciones basadas en *corpora* ocupan hoy día un lugar relevante dentro de los estudios lingüísticos. Gran parte de esa relevancia se debe al desarrollo creciente de las tecnologías, que están presentes en todos los campos del saber. El ordenador dejó de ser un artículo de lujo para convertirse en un artículo de primera necesidad. En efecto, la precisión científica ha encontrado en el ordenador una herramienta que favorece el almacenamiento y procesamiento de datos, superando la capacidad humana para gestionar gran cantidad de información. Por este motivo, las investigaciones a partir de *corpora* han

sido muy bien aceptadas por los investigadores, pues estos se dieron cuenta de que la intuición lingüística no era suficiente por sí sola.

La unión hombre/máquina posibilita un análisis cuantitativo más sistemático y un análisis cualitativo más dinámico, hecho que influye en el desarrollo de las investigaciones en diversas disciplinas como la traducción, la lexicografía, la terminología o la lingüística computacional. Para atender las necesidades de cada una de estas disciplinas, se han creado muchas plataformas electrónicas con el objetivo de perfeccionar los análisis de los *corpora* recolectados. Un ejemplo de herramienta para análisis léxico de *corpus* es el *WordSmith Tools*, desarrollado por Mike Scott, de la Universidad de Liverpool. Esta herramienta, bastante utilizada por los estudios del léxico y de la traducción, posibilita, entre otros muchos usos, el análisis de unidades fraseológicas y terminológicas mediante la observación comparada de dos lenguas, metodología que puede ayudar a aclarar matices desconocidos entre culturas diferentes.

Para Berber Sardinha (2004: 235-236), "a comparação da padronização entre duas línguas é muito importante para a área de tradução (...) pode iluminar várias questões relativas à adequação de itens equivalentes"[118]. Por lo tanto, algunas construcciones y expresiones, inicialmente equivalentes, pueden tener rasgos semánticos diferenciados dependiendo del discurso en que aparezcan.

Conforme a lo expuesto, este artículo analiza cuatro unidades fraseológicas en portugués y en francés siguiendo la metodología de la Lingüística de Corpus, y compara: "quem não tem cão caça com gato" / "faute de grives on mange des merles"; "matar dois coelhos com uma cajadada só" / "courir deux lièvres à la fois"; "dar bode" / "faire long feu"; "estar com a macaca" / "avoir bouffé du lion".

Antes de realizar el análisis propuesto en este estudio, se tratarán algunos conceptos de la Lingüística de Corpus y su teoría a partir de los trabajos de Biderman (2005), Berber Sardinha (2000, 2003, 2004), Magalhães (2001) y Mona Baker (1995, 1999) y, a continuación, los datos recogidos serán analizados con el *WordSmith Tools*, usando la herramienta *Concord*.

1.1. Lingüística de Corpus: algunos conceptos básicos

Antes de entrar en el análisis de los datos, es necesario precisar qué entiende la Lingüística de Corpus por *corpus*. Muchas son las definiciones, pero de acuerdo con la investigadora brasileña Tagnin (2004: 4), un *corpus* es "uma coletânea de textos em formato eletrônico, compilada segundo critérios específicos, considerada representativa de uma língua (ou da parte que se pretende estudar), destinada à pesquisa"[119].

En el campo de la traducción, los *corpora* sirven para analizar las traducciones y apuntar semejanzas y diferencias entre las lenguas; además, entre otras posibles aplicaciones, los *corpora* permiten observar el estilo de los traductores, desarrollar materiales para la enseñanza de una lengua extranjera y encontrar matices culturales muy específicos.

Mona Baker, en su estudio ya clásico *Corpora in Translation Studies: an overview and some suggestions for the future research*, de 1995, señala tres cambios importantes en la forma de concebirse el concepto de *corpus*:

---

118 *"la comparación de los padrones entre dos lenguas es muy importante para el campo de la traducción (...) puede aclarar diversas cuestiones relativas a la adecuación de ítems equivalentes"* (BERBER SARDINHA, 2004: 235-236).

119 "*una colección de textos en formato electrónico, seleccionada según criterios específicos, considerada representativa de una lengua (o de la parte que se pretende estudiar) y destinada a la investigación" (TAGNIN, 2004: 4).*

> (i) corpus now means primarily a collection of texts held in machine-readable form and capable of being analysed automatically or semi-automatically in a variety of ways; (ii) a corpus is no longer restricted to 'writings' but includes spoken as well as written text, and (iii) a corpus may include a large number of texts from a variety of sources, by many writers and speakers and on a multitude of topics. What is important is that it is put together for a particular purpose and according to explicit design criteria in order to ensure that it is representative of the given area or sample of language it aims to account for [120]. (1995: 225)

Si observamos la definición de Tagnin y la comparamos con la dada por Mona Baker, podemos deducir que las definiciones de *corpus* giran en torno a tres puntos principales. Primero, los datos deben ser electrónicos, o sea, en formato digital para que sean procesados por ordenador; segundo, los datos deben seguir criterios específicos de selección, es decir, seguir parámetros de búsqueda muy específicos y, tercero, los datos deben abarcar una parte representativa de la lengua que se quiere describir.

Para Berber Sardinha (2000), una de las definiciones más completas de *corpus* es la propuesta por Sánchez (1995):

> Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise. (SÁNCHEZ, 1995, apud BERBER SARDINHA, 2000: 8-9)[121].

Además de las discusiones que giran alrededor de la constitución de los *corpora*, la Lingüística de Corpus ha sido objeto de un acalorado debate sobre si es una metodología o una disciplina; esta consideración, que cambia según el punto de vista de cada teórico, no será tratada en este trabajo, aunque usemos sus presupuestos metodológicos, pues para nuestros objetivos es válida la definición que propone Berber Sardinha (2000: 349): "A Lingüística de Corpus trabalha dentro de um quadro conceitual formado por uma abordagem empirista e uma visão da linguagem enquanto sistema probabilístico".[122]

Un punto clave en el empirismo es focalizar los datos lingüísticos, aquí reunidos bajo la forma de *corpus*, como datos obtenidos por la observación de la lengua. Por su turno, la probabilidad indica que todas las construcciones lingüísticas, a pesar de permitidas por el sistema, tienen una frecuencia de ocurrencia y de unión distintas. Por este motivo, cuanto mayor sea la extensión del *corpus*, mayor será la posibilidad de observación de ocurrencias lingüísticas esporádicas, pero legítimas. (BERBER SARDINHA, 2000: 342).

---

[120] *(i) corpus hoy día significa primeramente una colección de textos mantenidos en formato que propicie el procesado por máquina y pueda ser analizado automáticamente o semi-automáticamente en una variedad de formas (ii) un corpus ya no está restringido a lo "escrito" sino que incluye tanto lo oral como el texto escrito, y (iii) un corpus puede contener un gran número de textos de una variedad de fuentes, de muchos autores y oradores y de una variedad de temas. Lo que es importante es que sea agrupado para un propósito específico y de acuerdo con criterios bien determinados para asegurar que sea representativo de una determinada aérea o muestra de lenguaje que se pretende tener en cuenta. (BAKER, 1995: 225)*

[121] *Un conjunto de datos lingüísticos (que pertenecen al uso oral o escrito de la lengua, o a ambos), que sigan determinados criterios, que sean lo suficientemente extensos en amplitud y profundidad como para ser representativos de la totalidad del uso lingüístico o de alguno de sus ámbitos, dispuestos de tal manera que permitan ser procesados por ordenador, con objeto de facilitar resultados variados y útiles para la descripción y el análisis. (SÁNCHEZ, 1995, apud BERBER SARDINHA, 2000: 8-9).*

[122] "*La Lingüística de Corpus trabaja dentro de un marco conceptual formado por un tratamiento empirista y una visión del lenguaje en tanto que sistema probabilístico*" (BERBER SARDINHA, 2000: 349).

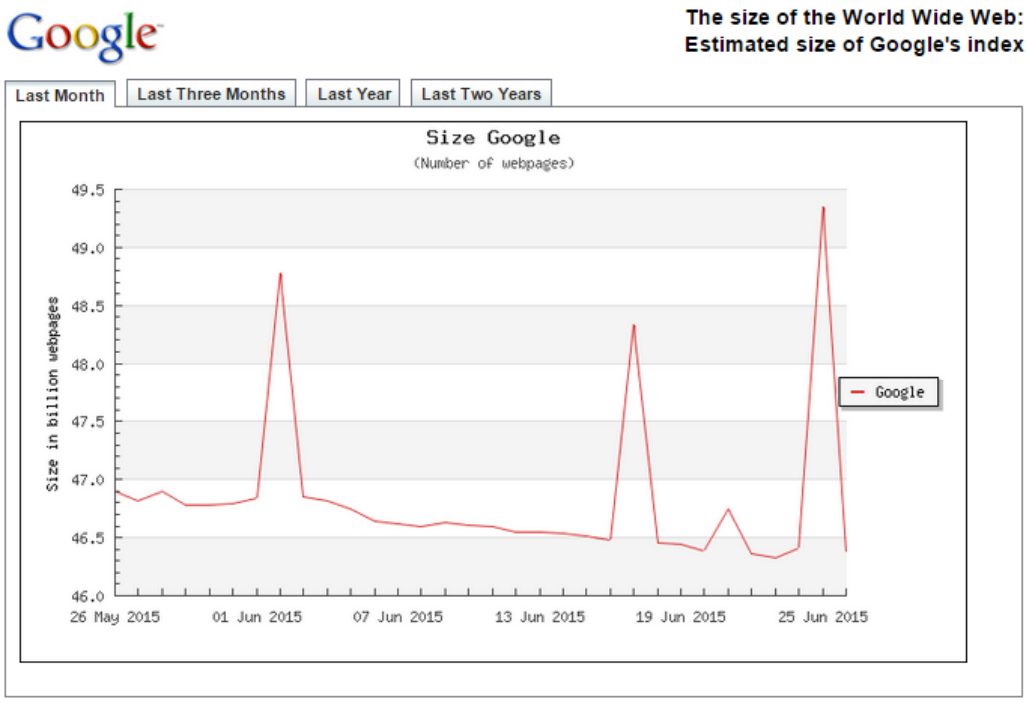
Con respecto a la extensión del *corpus*, creemos que la web puede funcionar como fuente de datos, como indica Xatara. Según esta autora, este recurso ofrece una gran cantidad de textos en diferentes formas de registro, o sea, tanto el lenguaje formal como el informal circulan libremente y, dependiendo del contenido de la página, la ocurrencia de usos eufemísticos será menor (2008: 772).

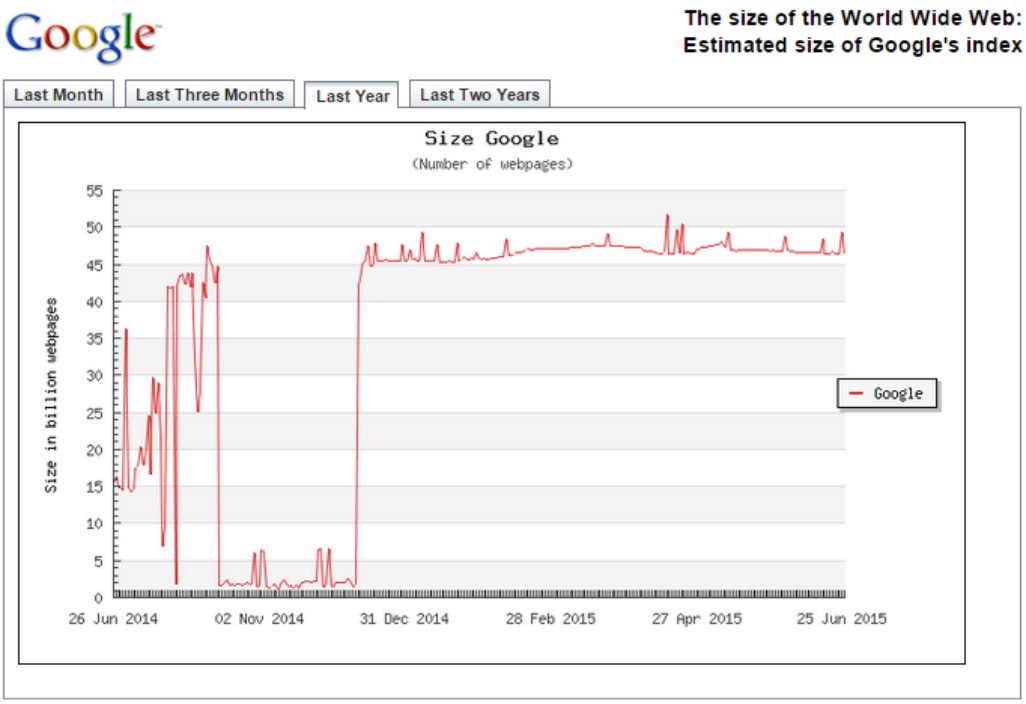
En relación a los fraseologismos, la web se presenta como un medio muy importante de investigación, pues condensa variaciones y ocurrencias legítimas que no observamos en el uso cotidiano del lenguaje, dado el grado de fijación y la espontaneidad del uso. Además del contenido y de los diferentes tipos textuales, podemos utilizar también la web con buscadores automáticos, estableciendo parámetros de búsqueda para que los datos sean lo más homogéneos posible, es decir, establecer parámetros y aplicar las mismas reglas de búsqueda a todos los textos que serán usados para crear el *corpus*.

1.2.La web como herramienta científica

Según investigaciones desarrolladas en la Universidad de Tilburg, Google mantiene la mayoría de las páginas indexadas en la web hasta contabilizar aproximadamente 49,5 billones de páginas en el mes de junio de 2015. En efecto, un buscador que tiene tal dimensión puede usarse también desde perspectivas muy diferentes y con diferentes objetivos, sin hablar de la posibilidad de hallar usos esporádicos para ciertas construcciones lingüísticas.

Para ilustrar el tamaño y la capacidad del Google, podemos observar a continuación una captura de pantalla del mes de junio de 2015 y otra del último año, correspondiente al período de junio de 2014 a junio de 2015. Como es de esperar, los números cambian, pues al tiempo que se crean muchas páginas, muchas otras se borran; a pesar de ello, la web sigue siendo un medio de búsqueda muy válido y una fuente de datos extraordinaria siempre que se fijen bien los parámetros de búsqueda.





Si comparamos Google con otras herramientas como Bing o Yahoo, Google destaca por su extensión y capacidad de almacenamiento, a lo que hay que sumar la facilidad y velocidad de las búsquedas. Como sugieren los investigadores de Tilburg, la web es una mina que permite hacer investigaciones con textos auténticos de la lengua. Sin embargo, constituye un terreno sin muchos parámetros, en el que no está clara la procedencia de los textos y la autoría, y donde hay muchos tipos de registros y no solo el escrito (audios, imágenes, links). No obstante, nos recuerdan que la palabra clave de las búsquedas basadas en este tipo de plataforma es "parámetros", es decir, todas las búsquedas deben obedecer a criterios claros para que los resultados sean homogéneos. Así, tomando como base los conceptos de la Lingüística de Corpus, queremos demostrar, en lo referente a los fraseologismos, que la web puede ser usada para la elaboración de corpus y también servir de base para la investigación científica.

En este sentido, es preciso especificar que este trabajo tiene dos fases muy claras: primera, la recolecta de los textos en que aparezcan los contextos de uso de cada unidad fraseológica y, segunda, la observación del *corpus* creado mediante la herramienta *Concord*, disponible en el *WordSmith Tools,* para analizar si las construcciones tienen los mismos usos en las dos culturas y descubrir matices desconocidos.

Así, el primer paso es desarrollar las búsquedas de los fraseologismos en Google, valiéndonos de la búsqueda avanzada, que garantizará la homogeneidad de los datos. En efecto, para la selección de los textos que iban a componer el *corpus* de análisis seguimos los siguientes parámetros de búsqueda:

1. Extensión Google: .br / .fr.
2. Caja de búsqueda: "esta palabra o frase exactas".
3. Páginas en idioma: portugués y francés.
4. Región: Brasil y Francia.
5. No protegidos por licencia.
6. Eliminar páginas de diccionarios y páginas explicativas de la procedencia y del significado mediante: -dicionário / -dictionnaire / -significado.

Para cada una de las unidades fraseológicas fueron seleccionados 50 textos, extraídos manualmente, para garantizar que todas las ocurrencias fueran legítimas y no contuviesen un uso no idiomático del fraseologismo. Por lo tanto, constituyen un pequeño *corpus*, según la definición de tipos de *corpora* de Berber Sardinha (2000), lo que es ideal para nuestros objetivos. Todos los textos fueron almacenados y trasladados al formato *txt (texto sin formato) para que pudieran ser procesados por el *WordSmith Tools*.

Como sabemos que en la web hay muchas páginas explicativas, fue necesario eliminar páginas de diccionarios usando la caja de exclusión "ninguna de estas palabras" y especificándose las palabras que no deberían aparecer en las páginas, como: -dicionário / -dictionnaire. El mismo procedimiento se aplicó para eliminar listas de proverbios y otras informaciones que tratan de explicar el significado de los fraseologismos en la búsqueda, sin contener textos con usos auténticos. Se permitió que hubiera textos en los cuales las unidades fraseológicas aparecieran en el cuerpo, en el título y, también, en los comentarios publicados por los usuarios de la red. En suma, con el método seguido, tenemos la garantía de que el objeto de estudio se contempla y se inserta en un contexto de uso realmente idiomático.

La herramienta utilizada para el análisis de los textos es *Concord*, que Teixeira (2006: 4) define así: "*Concord* muestra la palabra de búsqueda en KWIC (*Key Word in Context*)". Esta herramienta permite visualizar las búsquedas a partir de una palabra clave introducida por el investigador; a continuación, *Concord* busca la palabra clave en todos los textos que componen el *corpus* y forma una lista en la que la palabra aparece resaltada y viene acompañada del contexto de uso, es decir, de la frase en donde se inserta. Eso nos permite comparar las dos lenguas y determinar si las unidades fraseológicas se corresponden realmente.

Teniendo en cuenta lo expuesto, mostramos algunas unidades fraseológicas del portugués de Brasil y del francés de Francia con las correspondencias dadas en el *Dictionnaire d'expressions idiomatiques Français – Portugais – Français*, de Claudia Xatara, elaborado en colaboración con la Université de Nancy y que se encuentra disponible en línea (en adelante DEI). También comparamos las definiciones con el *Diccionario temático de locuciones francesas con su correspondencia española,* de Julia Sevilla Muñoz y Jesús Cantera Ortiz de Urbina (en adelante DT).

En el caso de uno de los fraseologismos, que no constituye una expresión idiomática y sí un proverbio, usamos el PIP, *Dicionário de Provérbios, Idiomatismos e Palavrões*, también de Claudia Xatara, y los diccionarios Le Robert (LR) y Michaelis (M), que también constan de versiones en línea.

1.3. El análisis del corpus

Tras las diversas etapas que seguimos para crear el *corpus*, usando la web como base de datos, lo sometimos a la herramienta *Concord* con el objetivo de observar matices lingüísticos y confirmar o rechazar las correspondencias dadas por los diccionarios que hemos mencionado.

LR - **Faute de grives, on mange des merles**. Faute de ce que l'on désire, il faut se contenter de ce que l'on a.
M - **Quem não tem cão caça com gato**. Diz-se de quem precisa prover meios eventuais para resolver problema.

El primer aspecto que llama la atención en este análisis es el matiz diferente de la definición de cada diccionario, ya que en francés este proverbio se relaciona con el deseo y en portugués con la solución de problemas; sin embargo, si comparamos los usos, encontramos en común la característica de suplir una necesidad (ligada a diversos ámbitos de la vida) con algo que en el momento de la elocución no era ideal o esperado.

No obstante, los usos que llamaron más nuestra atención fueron los relacionados con la política, como podemos observar:

democrática -, o regime em vigor promove no seio da população elevadas taxas de absenteísmo político e ideológico. O que fazer para camuflar a ausência do povo – esse vazio inexplicável e imperdoável? Como quem não tem cão, caça com gato, o regime abriu espaço para o ersatz representado pela participação política terceirizada. Por exemplo: as campanhas eleitorais, assim como a propaganda partidária entre as eleições, foram entregues a "pessoas estranhas ao serviço" – na base, milhares de biscateiros que, entre uma briga de galo e outra, comandam a fala dos candidatos e repetem na esfera pública

como o partido mais português de Portugal, o equivalente à aldeia de Monsanto do tempo dos concursos de outras épocas, se identifique inteiramente com um dito tão popular como aquele que nos diz que "quem não tem cão caça com gato". Como em tempo de oposição o PSD tende a assemelhar-se a um saco de gatos aquilo a que estamos a assistir nas directas do PSD assemelha-se a uma disputa de gatos de rua. Num partido onde só

isso é natural do processo político. Mas, não vamos tomar decisões afobadas. O que nos importa é tomar decisões certas e compor alianças com partidos que tenham afinidade com o DEM. Não vou fazer política do quem não tem cão caça com gato", destacou o deputado, numa sutil alusão a um recente episódio (ver nota). Neto também disse que, apesar de ter apoio do deputado Maurício Trindade, ainda luta pelo apoio irrestrito do PR: "Não vou definir meu

baixo que a criança porque ela, automaticamente, pára. Nunca no "Cosquinha" a gente teve que fazer isso. Elas acompanham a história e vêm brincar com a Amanda e com a Luiza. "NÃO VOU FAZER POLÍTICA DO QUEM NÃO TEM CÃO CAÇA COM GATO" Depois das definições de alguns adversários políticos, surgiu o comentário de que o DEM vive grande pressão para composição da sua coligação. Mas, o deputado ACM Neto descaracteriza a idéia.

En Brasil, esta unidad fraseológica cuando se emplea en ciertos contextos políticos adquiere la connotación de escoria, es decir, uso de materiales de segunda mano y de inferior calidad. Estos usos están relacionados con la corrupción, que en Brasil es con diferencia mayor que en Francia.

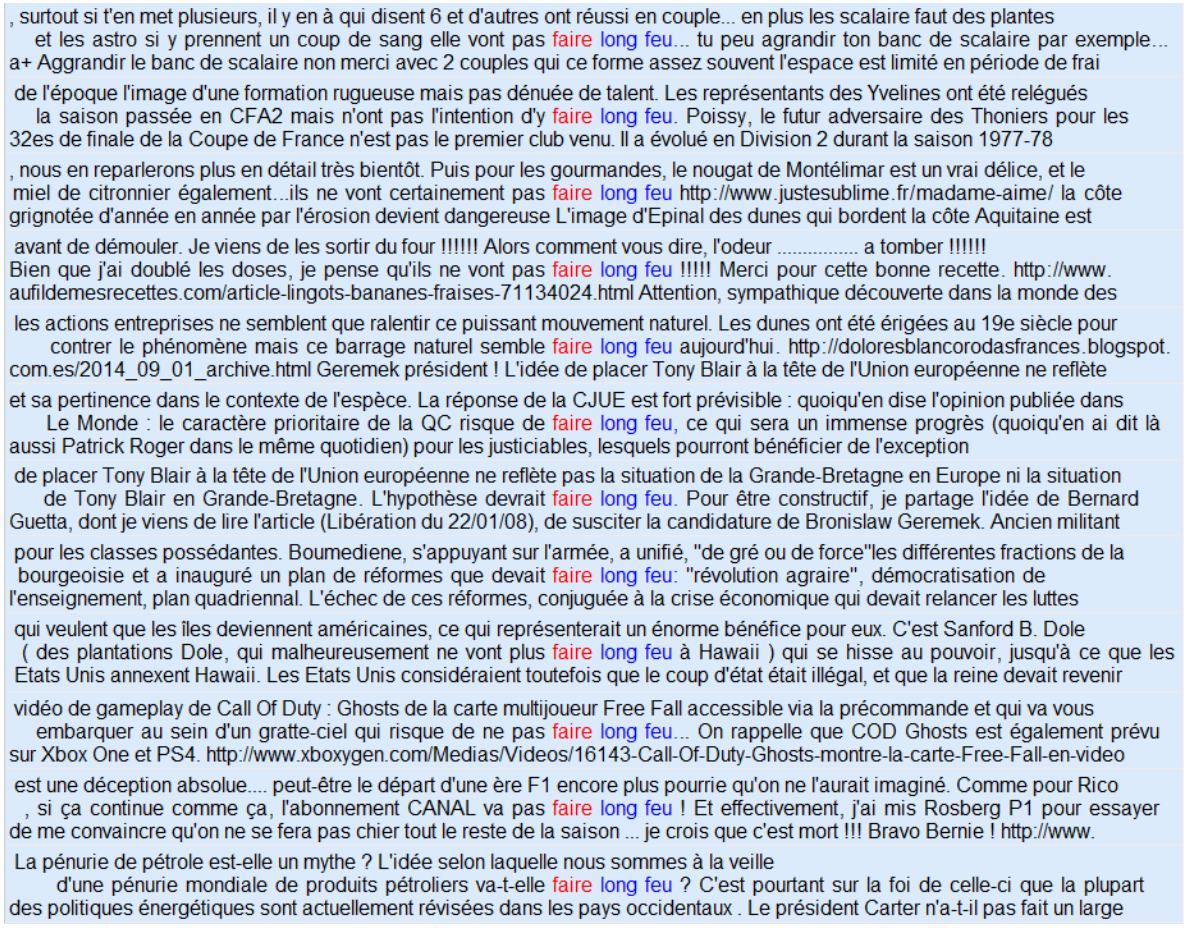
En el *corpus* seleccionado fue posible encontrar un contexto en francés relacionado con la política, que, como podemos observar, se refiere a la falta de opción de candidatos para asumir un determinado puesto.

anbitonnerait même la présidence du parlement européen. Mais là encore, tout est affaire d'équilibre. 3 commentaires: agriculteur a dit... Faute de grives, on mange des merles. Je croyais l'homme plus ambitieux. Il ne compte pas briguer la Présidence de la République flamande ? himself a dit... "Verhofstadt voit dans le Parlement européen, dont le rôle sera renforcé par le Traité de Lisbonne, l'occasion de

DEI - **Faire long feu**. Manquer son but, échouer. *Dar bode.*
DEI - **Dar bode**. Haver problema. *Faire long feu.*
DT - **Faire long feu**. 1. Errar el tiro con un arma. *Fallar.* 2. No alcanzar su objetivo en un negocio. *Fracasar.* Alusión a la mecha que arde lentamente y se apaga antes de que el fuego alcance la pólvora o el cartucho de un arma cuyo pistón (en francés *amorce*) arde demasiado lentamente de manera que el tiro falle su objetivo.

Con el análisis de esta unidad fraseológica se puso aún más de manifiesto la diferencia entre esta unidad y otra muy parecida, pero que usa la forma negativa y que no aparece en ninguna de las definiciones, lo que nos alertó sobre el cambio de uso. Está claro que estas unidades fraseológicas, al ser diferentes, deben aparecer en los diccionarios en entradas diferentes; ahora bien, si la obra fuese dirigida a extranjeros o tuviese una función didáctica, una buena opción sería incluir una referencia o una nota informando del cambio de sentido.

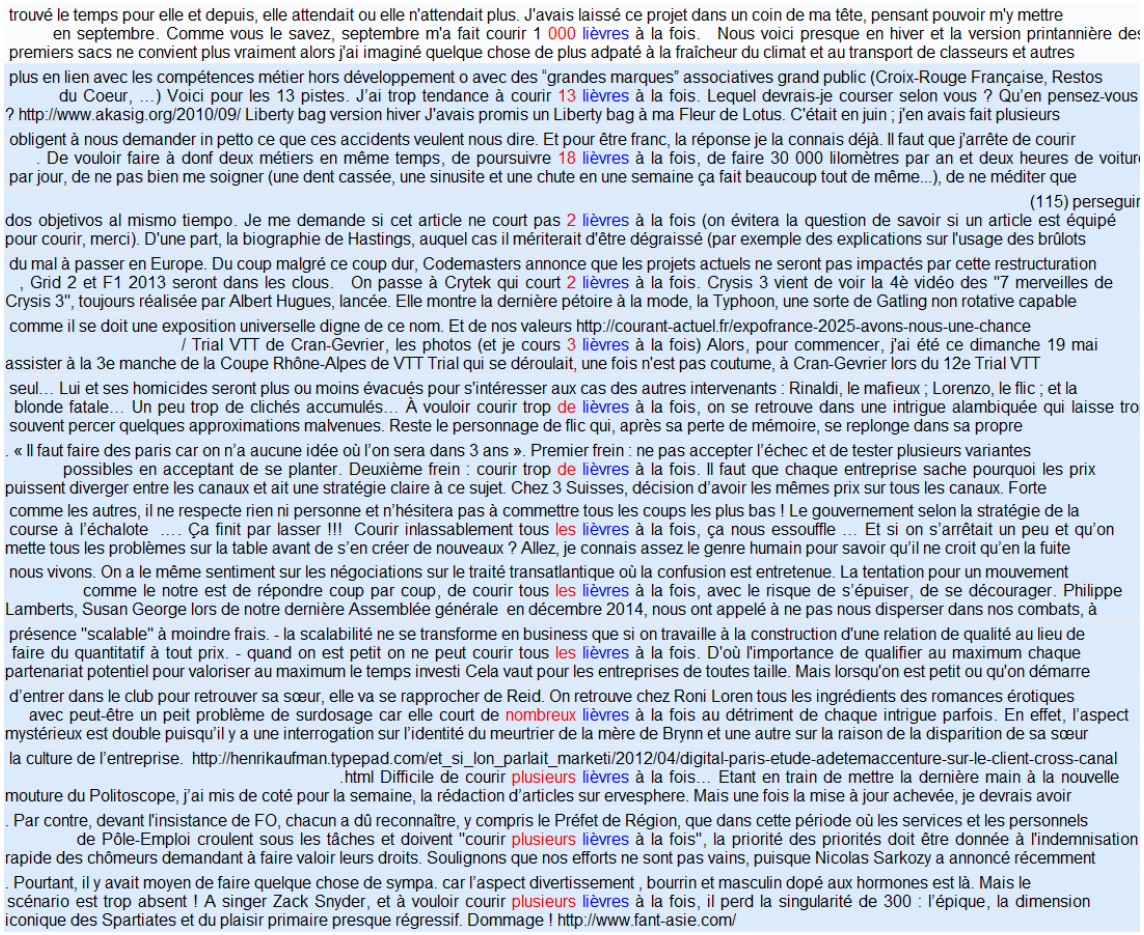
Así pues, "faire long feu", que significa "fallar", "no lograr éxito", no significa lo mismo que "ne pas faire long feu" que tiene el sentido de "no quedar mucho tiempo", tal como podemos ver en los ejemplos siguientes:



El segundo y el tercer caso de la lista ilustran bien la diferencia de sentido y de uso de estas dos unidades; por consiguiente, nos parece pertinente incluir una observación en la explicación semántica de los fraseologismos que informe de la diferencia de sentido que conlleva la forma negativa.

DEI - **Courir deux lièvres à la fois**. Atteindre deux objectifs par la même action. *Matar dois coelhos com uma cajadada só.*
DEI – **Matar dois coelhos com uma cajadada só.** Resolver dois problemas diferentes de uma só vez. *Courir deux lièvres à la fois.*
DT - **Courir deux lièvres à la fois**. Perseguir dos objetivos al mismo tiempo. *Correr dos liebres a la vez.*

Es posible observar con la ayuda del *corpus* cómo la partícula "deux", en los contextos del francés, se ha sustituido en la gran mayoría de los casos por otras unidades indicativas del sentido "más de uno", como "plusieurs lièvres", "trop de lièvres", "3 lièvres", "18 lièvres" y "1.000 lièvres", lo que también ocurre en portugués. La cantidad suele cambiar según el énfasis expresivo del contexto, como podemos ver a continuación, en el que hallamos una intensificación, como es el caso de 1.000, o una adaptación al contexto como en el ejemplo en que aparece el 13:


trouvé le temps pour elle et depuis, elle attendait ou elle n'attendait plus. J'avais laissé ce projet dans un coin de ma tête, pensant pouvoir m'y mettre en septembre. Comme vous le savez, septembre m'a fait courir 1 000 lièvres à la fois. Nous voici presque en hiver et la version printannière des premiers sacs ne convient plus vraiment alors j'ai imaginé quelque chose de plus adpaté à la fraîcheur du climat et au transport de classeurs et autres

plus en lien avec les compétences métier hors développement o avec des "grandes marques" associatives grand public (Croix-Rouge Française, Restos du Coeur, ...) Voici pour les 13 pistes. J'ai trop tendance à courir 13 lièvres à la fois. Lequel devrais-je courser selon vous ? Qu'en pensez-vous ? http://www.akasig.org/2010/09/ Liberty bag version hiver J'avais promis un Liberty bag à ma Fleur de Lotus. C'était en juin ; j'en avais fait plusieurs

obligent à nous demander in petto ce que ces accidents veulent nous dire. Et pour être franc, la réponse je la connais déjà. Il faut que j'arrête de courir . De vouloir faire à donf deux métiers en même temps, de poursuivre 18 lièvres à la fois, de faire 30 000 lilomètres par an et deux heures de voitur par jour, de ne pas bien me soigner (une dent cassée, une sinusite et une chute en une semaine ça fait beaucoup tout de même...), de ne méditer que

(115) perseguir

dos objetivos al mismo tiempo. Je me demande si cet article ne court pas 2 lièvres à la fois (on évitera la question de savoir si un article est équipé pour courir, merci). D'une part, la biographie de Hastings, auquel cas il mériterait d'être dégraissé (par exemple des explications sur l'usage des brûlots

du mal à passer en Europe. Du coup malgré ce coup dur, Codemasters annonce que les projets actuels ne seront pas impactés par cette restructuration , Grid 2 et F1 2013 seront dans les clous. On passe à Crytek qui court 2 lièvres à la fois. Crysis 3 vient de voir la 4è vidéo des "7 merveilles de Crysis 3", toujours réalisée par Albert Hugues, lancée. Elle montre la dernière pétoire à la mode, la Typhoon, une sorte de Gatling non rotative capable

comme il se doit une exposition universelle digne de ce nom. Et de nos valeurs http://courant-actuel.fr/expofrance-2025-avons-nous-une-chance / Trial VTT de Cran-Gevrier, les photos (et je cours 3 lièvres à la fois) Alors, pour commencer, j'ai été ce dimanche 19 mai assister à la 3e manche de la Coupe Rhône-Alpes de VTT Trial qui se déroulait, une fois n'est pas coutume, à Cran-Gevrier lors du 12e Trial VTT

seul... Lui et ses homicides seront plus ou moins évacués pour s'intéresser aux cas des autres intervenants : Rinaldi, le mafieux ; Lorenzo, le flic ; et la blonde fatale... Un peu trop de clichés accumulés... À vouloir courir trop de lièvres à la fois, on se retrouve dans une intrigue alambiquée qui laisse tro souvent percer quelques approximations malvenues. Reste le personnage de flic qui, après sa perte de mémoire, se replonge dans sa propre

. « Il faut faire des paris car on n'a aucune idée où l'on sera dans 3 ans ». Premier frein : ne pas accepter l'échec et de tester plusieurs variantes possibles en acceptant de se planter. Deuxième frein : courir trop de lièvres à la fois. Il faut que chaque entreprise sache pourquoi les prix puissent diverger entre les canaux et ait une stratégie claire à ce sujet. Chez 3 Suisses, décision d'avoir les mêmes prix sur tous les canaux. Forte

comme les autres, il ne respecte rien ni personne et n'hésitera pas à commettre tous les coups les plus bas ! Le gouvernement selon la stratégie de la course à l'échalote .... Ça finit par lasser !!! Courir inlassablement tous les lièvres à la fois, ça nous essouffle ... Et si on s'arrêtait un peu et qu'on mette tous les problèmes sur la table avant de s'en créer de nouveaux ? Allez, je connais assez le genre humain pour savoir qu'il ne croit qu'en la fuite

nous vivons. On a le même sentiment sur les négociations sur le traité transatlantique où la confusion est entretenue. La tentation pour un mouvement comme le notre est de répondre coup par coup, de courir tous les lièvres à la fois, avec le risque de s'épuiser, de se décourager. Philippe Lamberts, Susan George lors de notre dernière Assemblée générale en décembre 2014, nous ont appelé à ne pas nous disperser dans nos combats, à

présence "scalable" à moindre frais. - la scalabilité ne se transforme en business que si on travaille à la construction d'une relation de qualité au lieu de faire du quantitatif à tout prix. - quand on est petit on ne peut courir tous les lièvres à la fois. D'où l'importance de qualifier au maximum chaque partenariat potentiel pour valoriser au maximum le temps investi Cela vaut pour les entreprises de toutes taille. Mais lorsqu'on est petit ou qu'on démarre

d'entrer dans le club pour retrouver sa sœur, elle va se rapprocher de Reid. On retrouve chez Roni Loren tous les ingrédients des romances érotiques avec peut-être un peit problème de surdosage car elle court de nombreux lièvres à la fois au détriment de chaque intrigue parfois. En effet, l'aspect mystérieux est double puisqu'il y a une interrogation sur l'identité du meurtrier de la mère de Brynn et une autre sur la raison de la disparition de sa sœur

la culture de l'entreprise. http://henrikaufman.typepad.com/et_si_lon_parlait_marketi/2012/04/digital-paris-etude-adetemaccenture-sur-le-client-cross-canal .html Difficile de courir plusieurs lièvres à la fois... Etant en train de mettre la dernière main à la nouvelle mouture du Politoscope, j'ai mis de coté pour la semaine, la rédaction d'articles sur ervesphere. Mais une fois la mise à jour achevée, je devrais avoir

. Par contre, devant l'insistance de FO, chacun a dû reconnaître, y compris le Préfet de Région, que dans cette période où les services et les personnels de Pôle-Emploi croulent sous les tâches et doivent "courir plusieurs lièvres à la fois", la priorité des priorités doit être donnée à l'indemnisation rapide des chômeurs demandant à faire valoir leurs droits. Soulignons que nos efforts ne sont pas vains, puisque Nicolas Sarkozy a annoncé récemment

. Pourtant, il y avait moyen de faire quelque chose de sympa. car l'aspect divertissement , bourrin et masculin dopé aux hormones est là. Mais le scénario est trop absent ! A singer Zack Snyder, et à vouloir courir plusieurs lièvres à la fois, il perd la singularité de 300 : l'épique, la dimension iconique des Spartiates et du plaisir primaire presque régressif. Dommage ! http://www.fant-asie.com/

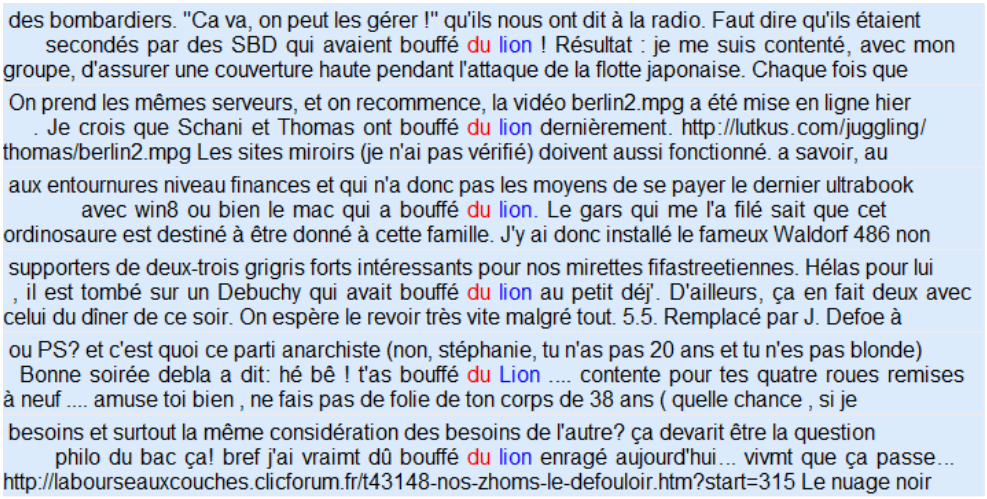
La mejor definición de esta unidad fraseológica quizá sea "una acción que tiene como finalidad alcanzar más de un resultado a la vez" y, después, debería incluir ejemplos que contengan otras partículas con sentido de más de uno.

DEI- **Bouffer du lion**. S'animer ou s'énerver beaucoup. *Estar com a macaca.*
DEI- **Estar com a macaca**. Estar muito excitado, alterado. *Bouffer du lion.*
DT- **Avoir bouffé du lion**. Mostrar una energía inusual, excepcional. *Ser valiente a prueba de bomba. Ser valiente a más no poder. Estar hecho una fiera / un jabato.*
Esta expresión se emplea sólo en pasado: *Il a bouffé du lion.*

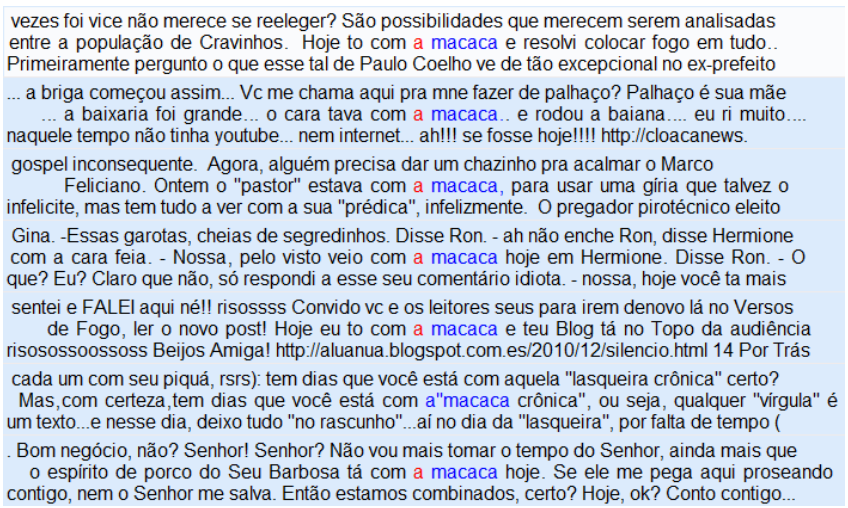
Tanto en los contextos del francés como en los del portugués es posible apreciar dos usos, ya sugeridos por los diccionarios, que pueden presentar prosodia semántica positiva o negativa. Para Berber Sardinha (2004: 236), la prosodia semántica puede ser "a associação recorrente entre itens lexicais e um campo semântico, indicando uma certa conotação (negativa, positiva ou neutra) ou instância avaliativa"[123]. Para este autor, el estudio de la prosodia semántica es importante en la medida que revela significados que los manuales de traducción o los diccionarios pueden no presentar.

[123] "la asociación recurrente entre ítems lexicales y un campo semántico, que indique cierta connotación (negativa, positiva o neutra) o instancia evaluativa" (BERBER SARDINHA, 2004: 236).

En los siguientes casos, hay una prosodia semántica positiva para los contextos que sugieren "animación", "excitación" y una prosodia semántica negativa para los que sugieren "rabia", "enfado".



En ambos casos las unidades fraseológicas pueden estar asociadas a unidades lexicales también positivas o negativas, como "enragé" o "a macaca crónica"; por supuesto, estas unidades conforman el contexto de uso y son de gran ayuda para la comprensión general.



2. CONCLUSIÓN

Creemos que este estudio puede ser de ayuda tanto para los investigadores del léxico como para los profesionales de la traducción, sobre todo si tenemos en cuenta la metodología utilizada, pues es accesible a todos los que tengan una conexión a Internet y, además, la versión demo del *WordSmith Tools* puede descargarse gratuitamente.

También creemos que con parámetros de búsquedas bien delimitados es posible usar la web como fuente de investigación científica, ya que esta surge como un *corpus* de textos y usos auténticos de la lengua actual.

Además, para quienes trabajan en la creación de materiales didácticos y diccionarios, y también para los traductores, esta metodología puede ser bastante eficiente, pues permite

observar las delicadas relaciones de sentido de las unidades fraseológicas con la cultura e incluso reconocer ciertas construcciones o variaciones de unidades ya fijadas.

Como hemos intentado mostrar, la Lingüística de Corpus se integra en la tendencia actual de dar mayor precisión a las investigaciones, y también más credibilidad, especialmente en el Procesamiento del Lenguaje Natural, pues permite la comparación y el análisis de una gran cantidad de datos, además de posibilitar el almacenamiento de parcelas muy específicas de la lengua. Por lo tanto, cada uno de estos bancos de datos se constituye y se usa para objetivos determinados.

Bibliografía

BAKER, M., 1995. *Corpora in Translation Studies*: an overview and some suggestions for the future research. *Target*, 7:2, pp. 223-243.

BAKER, M., 1999. Lingüística e estudos culturais: paradigmas complementares ou antagônicos nos estudos da tradução? In: MARTINS, M. A. P. (Org.), 1999. *Tradução e multidisciplinaridade*. Rio de Janeiro: Lucerna, pp. 15-34.

BERBER SARDINHA, A., 2000. *Linguística de corpus*: histórico e problemática. *D.E.L.T.A.*, 16 (2), pp. 323-367.

BERBER SARDINHA, A., 2003. Uso de corpora na formação de tradutores. *D.E.L.T.A.*, 19: Especial, pp. 43-70.

BERBER SARDINHA, A., 2004. *Linguística de Corpus*. São Paulo: Monole.

BIDERMAN, M. T. C., 2005. Unidades complexas do léxico. In: RIO-TORTO, G.; FIGUEIREDO, O.M; SILVA,F. (Org.*). Estudos em Homenagem ao Professor Doutor Mário Vilela*. 1. ed., v. II. Portugal: Faculdade de Letras da Universidade do Porto. pp. 747-757.

MAGALHÃES, C. M., 2001. Pesquisas textuais / discursivas em tradução: o uso de corpora. In: PAGANO, A. (Org). *Metodologias de pesquisa em tradução*. Belo Horizonte: FALE-UFMG.

SEVILLA MUÑOZ, J.; ORTIZ DE URBINA, J. C., 2004. Diccionario temático de locuciones francesas con su correspondencia española. Madrid: Gredos, 2004.

TAGNIN, S. E. O., 2001. *COMET: um Corpus Multilíngüe para Ensino e Tradução*. São Paulo: USP.

TAGNIN, S. E. O., 2002. Os *Corpora*: instrumentos de auto-ajuda para o tradutor. In *Cadernos de Tradução* IX. Florianópolis. Disponible en: http://www.periodicos.ufsc.br/index.php/traducao/article/view/5986/5690. [Acceso 15 mayo 2015].

TAGNIN, S. E. O., 2004. *Corpora: o que são e para quê servem*. Diapositivas. Disponible en: http://comet.fflch.usp.br/sites/comet.fflch.usp.br/files/u30/Lexicografia_2004.pdf. [Acceso 18 abril 2015]

TEIXEIRA, E. D., 2006. *Como usar o Wordsmith Tools*. V.3. São Paulo: Universidade de São Paulo.

XATARA, M. C., 2008. A Web para um levantamento de freqüência. UNESP, São José do Rio Preto. Disponible en: http://www.filologia.org.br/ileel/artigos/artigo_398.pdf. [Acceso 18 abril 2015].

DO WE NEED EQUIVALENCE-BASED E-TOOLS?

Maciej Paweł Jaskot
University of Social Sciences and Humanities (Warsaw, Poland)
mjaskot@swps.edu.pl

Abstract

The article addresses the question of how phraseological units (PUs) should be treated in e-tools dedicated to translation. Some models used in contrastive phraseology are presented. The equivalency of PUs is emphasized as well as their status as culturemes. Equivalent translation of PUs is possible if a *langue*-approach (with cultural-anchoring and pragmatic appropriateness considered as crucial variables) is adopted. However, unless a cross-linguistic definition of the boundaries of PUs is determined, a from-form-to-meaning approach will dominate together with a *parole*-approach such as the one used in word units translation. Defining the notion of equivalence, or taking into consideration one of the possible interpretations of this concept, is the first step to prepare a base for more extensive equivalent units, such as PUs. The second step would be to reconsider the notion of the PU, as equivalence seems to be possible at many levels, i.e. many variables of equivalence shall be taken into account.

Un ordine di parole può essere più micidiale di una formula chimica.

Gabriele D'Annunzio

1. INTRODUCTION

Recently, a lot of effort has been put into creating lexicographic e-tools such as bi- or multilingual dictionaries, phraseological lexicons and parallel corpora. In order to make them efficient it is necessary to address such theoretical issues as the formulation of linguistically grounded principles of selection of phraseological units (Pus), their translation, semantic interpretation based on diversity of the linguistic image of the world. We must not forget about linguistic ambiguities and, last but surely not least, the equivalence issue. Defining the notion of equivalence, or taking into consideration one of the possible interpretations of this concept, is the first step to prepare a base for equivalent units. The second step would be to reconsider the notion of the phraseological unit as equivalence seems to be possible at many levels (many variables of equivalence shall be taken into account), and the linguistic one does not solve the problem of translating (transposing) a more complex word unit (such as a PU, for example). Rethinking equivalence through the notion of cultureme would be useful and helpful to answer the question of what we should expect from modern e-lexical tools intended to help us translate PUs.

2. CONTRASTIVE PHRASEOLOGY CHALLENGE

The last decades in contrastive linguistics have been marked by the increasing interest in phraseology in its various aspects, especially the lexicographical one. However, despite the significant developments in the theory and practice of compiling translated lexicons, contrastive phraseology between some languages, often from the same group, such as Polish, Ukrainian or Russian, remains under-studied due to the lack of work that would systematically present the phraseology of these languages. As far as other languages are concerned, this particular interest in contrastive phraseology does not seem to have grown so much as yet so as to lead to the creation of phraseological e-tools, which poses a challenge for lexicographers.

The work carried out by the Lingua-Information Fund of the National Academy of Sciences of Ukraine deserves a mention. Its aim is to develop a set of instruments of the translational phraseological system of Ukrainian-Polish electronic phraseological dictionary (EPD). It involves the construction of lexicographic model of representation of phraseology in the system of the Ukrainian and Polish languages, as well as the formation of a lexicographical database of Ukrainian and Polish phraseological units (hereinafter – PU)[124]. To this end, software aimed at distinguishing the structural elements of the digital dictionary has been developed according to the structure of lexicographical system.

If a bi-lingual e-dictionary arouses a number of questions, the idea of a tri-lingual dictionary seems a far greater challenge. The team from the Institute of Slavic Studies of the Polish Academy of Sciences (ISS PAS), led by prof. Koseska-Toszewa, has been struggling to prepare a unique tri-lingual dictionary (Polish-Bulgarian-Russian) reflecting a new approach to research in semantics and contrastive linguistics (Koseska-Toszewa et alia, 2013).

It is widely known that lexical units are distinguishable from other types of phrases by their complicated semantics, which is strongly oriented towards a specific linguistic view of the world of a speakers' community. Therefore, the main problem about compiling a plurilingual dictionary is the selection of adequate translational equivalents with due account for differences in the worldview represented in the respective language systems. The same problem appears while dealing with PUs. In addition, discrepancies in determining what should and what should not be considered a PU often occur. Leaving this particular question aside, we still believe that one of the tasks of a comprehensive translational phraseological dictionary is to convey the phraseological system of one language by means of another language.

But what factors should be considered when choosing a PU to be compared? Since the frequency of use does not seem to be the best way to talk about the motivation for the use of a specific PU. As it has been demonstrated, while compiling a tri-lingual corpus at the already mentioned ISS PAS, the use of PUs is never mandatory and a PU can always be

[124] The basis for building the dictionary database of the registry of Ukrainian PUs is the phraseological subsystem of academic explanatory dictionary of the Ukrainian language in 20 volumes (SUM-20). The phraseological units of SUM-20 level include phrases that are of established composition and structure and lexically undividable, and express complete specific meaning that is not the sum of the values of its components, but occurs on the basis of rethinking mainly due to metaphorical and analogical processes (SUM-20, 2010, 24). Separated into the so-called "rhombed" zone of the Dictionary are also the collocations – set phrases that allow slight desemantization of one component (eg. *вовчий апетит*), word equivalents (eg. *до безмежжя*) and terminological phrases (eg. *топографічна анатомія*).

replaced by a different structure or circumlocutions. Is, however, the choice of PUs for a dictionary subjective?[125]

3. HOW TO ORGANIZE E-PU-DICTIONARIES? A POSSIBLE SOLUTION

One of the possible solutions for organizing a PU-multilingual-dictionary is the one adopted by the authors of Ukrainian-Polish phraseological dictionary. They pay special attention to the semantic characteristic of phraseological units by providing semantic context in which a phraseological unit is used and include dictionary definitions. Lexical compatibility of phraseological units presented as some PUs can be accompanied by different words, e.g.: *як по маслу* **(**рідше *по масли***),** перев. зі сл. *ими*, *ними* і т. ін. *jak po maśle; jak z płatka.*

The inclusion of definitions of the register of each unit helps to describe the nuances of meaning of PUs which are not conveyed by equivalents, and provide Ukrainian and Polish users with comprehensive lexicographic information for accurate translation:

In this particular case a dictionary entry includes various structural elements, or variables, such as the capital PU, alternatives, optional components, phraseological polysemy, interpretation formulas for some meanings of phraseological units, translational equivalents, grammar and stylistics remarks. The authors are aware that when translating PUs by taking into account their accurate meaning nuances, it is important to consider the expressive and stylistic features of a PU, since often such units of the two languages develop in different ways, have their specific features. That's why the principle of semantic and functional-stylistic adequacy of the translational phraseological equivalents is crucial.

4. THE EQUIVALENCY ISSUE

The problem of equivalence lies in the area, in which an interdisciplinary consensus has been achieved: lexico-semantic structures of lexis of a particular language are exceptional, specific to this language and, therefore, they are partially unique. It means that the lexical-semantic structures of two (or more) languages are non-isomorphic. Non-isomorphy of lexis forms the theoretical and observed empirical circumstances, examination of which leads to concrete manifestations of the problem of equivalence in different disciplines (cf. Jaskot & Ganoshenko, 2015). In this article, I am particularly interested in the metalexicographic aspect of this issue. Taking into account the non-isomorphy of languages seems, however, an aspect that generates a wide variety of interpretations within the same discipline.

4.1. The Ukrainian-Polish phraseological dictionary solution

The authors of the Ukrainian-Polish dictionary use three levels of equivalents:

The first one is called simply an equivalent. Such equivalents are the same by semantic, structural and stylistic characteristics. In this category international phraseological units can be found: *Піррова перемога* – *Pyrrusowe zwycięstwo*; *перейти рубікон* – *przejść, przekroczyć Rubikon.*

[125] In 2006, I published a small Spanish-Polish Dictionary of Idioms containing circa 100 PUs and I had to face the problem of chosing the most frequently used PUs.

At the second level the so-called adequate (or analogous) equivalents are found. Such equivalents partially or completely differ in connotation, have different grammatical structure, e.g.: *з одного міста – ulepiony z jednej, tej samej gliny*; *з горем [з лихом] пополам – z trudem; z [wielką] biedą.*

The third level is reserved for descriptive equivalents: a free, sometimes ad-hoc created phrase is used to translate them, eg.: *одна нога тут, а друга там – w podskokach; bardzo szybko; як в аптеці – ni mniej ni więcej; dokładnie.*

4.2. Equivalent according to Alla Luchyk

Alla Luchyk, professor at the National University of "Kyiv-Mohyla Academy" in Kiev is the co-author (together with O. Antonova) of the *Polish-Ukrainian Dictionary of Word Equivalents*. She understands (Luchyk & Antonova, 2012) the word 'equivalent' as one of the most common category of a fixed word combination (in case of Ukrainian and Polish). A. Luchyk claims that these elements (word equivalents) are lacking in both monolingual and plurilingual dictionaries. The spelling of these units, and not their metaphorical meaning which makes them closer to PU, is what contributed to the fact that in most cases word equivalents were not identified as a separate structural unit in the bilingual dictionary.

Let's look at an example:

ВЕСЬ (ВВЕСЬ, УВЕСЬ) ЧА́С *аналог прислівника* (постійно, безперервно) **przez cały czas, ciągle, w dalszym ciągu** or **ВІДТО́ДІ, ЯК** *аналог сполучника* (те саме, що **З ТИХ ПІР, ЯК**) **od tej pory<,> jak; od czasu<,> kiedy**.

4.3. Which type of equivalence is needed in dictionaries?

There still seem to be differences of opinion among lexicographers on what should be considered an equivalent (and equivalent to what?) and, consequently, what types of equivalency can be distinguished. The need for such a distinction is important, depending on the dictionary type and its function.

The two abovementioned approaches to the concept of equivalence take into consideration first of all the form of a word, or a more extended unit, but not its meaning. This type of approach will defend itself against criticism in practical lexicography when there is an urgent need to determine what a word is, because, according to A. Luchyk, one of the major problems of creating a modern dictionary is defining the notion of the word and its boundaries (cf. Luchyk & Antonova, 2012). I fully agree with Wiegand (2011), who points out that "bilingual dictionaries are not conceptualized as aids for contrastive studies of language systems (even though some advocates of contrastive lexicology use them in this way). They are rather meant, in the first place, to be a tool to understand and produce foreign texts and to make translations in both directions." Talking about equivalence in translation, I'd rather refer to whole texts and their translations. Apart from proverbs (or PUs), this is not at all the case in lexicography, as it needs to deal with the equivalence of meaning-bearing units below the level of sentences. Thereupon, a key to understanding the equivalence of words, free syntagmas, and units with several words which do not form sentences, especially idiomatic expressions and collocations, is needed (cf. Wiegand, 2011). Consequently, we cannot reduce the concept of equivalence to single lexical items by taking their form into account. If so, we would not be able to consider equivalent the Polish expression *stawić czoło* and the Spanish one *enfrentarse* (to front/face something), although they function as equivalents. That's why the equivalence relationship cannot be based on such unclear criteria as lexical units. If we look for equivalent items in a target language, we should consider its equivalence relationship, based on *parole* conditions, i.e. the equivalence

will occur within one of the item's meanings. Wiegand claims that "bilingual equivalence is a relationship between a source and target language item which is present when the threefold predicate *x is equivalent to y with regard to z* is true; "z" is the variable for the criterion of equivalence." According to this statement, the Polish word *pani* is equivalent to Spanish *señora* in the lexicographical understanding of equivalence. In this case the "z" variable is lexicographical, i.e. it is referentially-semantically equivalent. The same type of equivalency is observable between the Polish *pani* and Spanish *doña*. Thus, we could say that indeed, as stated in "classical" textbooks on translation studies "Translation starts with establishing equivalence on the word level." (cf. Ivanov, 2006). Nevertheless, what was claimed by Ivanov (2006), i.e. "functional compliance in a target language, transmitting expression on the similar level (words, collocations) to all relevant components within the given context, or one of the variants of meaning of the original unit in the source language" still has not been achieved. Indeed, we cannot assert that Spanish *señora*, Polish *pani* and English *lady* are equivalent without a clear definition of the abovementioned "z" variable, according to which the use of the three words of the example will be allowed in a particular context subject to translation. In this case we face pragmatically labeled units. In consequence, pragmatic equivalence is needed. Kisiel, Satoła-Staśkowiak, Sosnowski (2014) point out the difficulty one may encounter while compiling a dictionary:

> "The main problem in the trilingual dictionary is the presentation of adequate equivalents. Adequate selection indicates that 1. The range of meanings of equivalents is identical (compare Polish *grypsera*. "language of criminals", Bulgarian *гаменски жаргон* and Russian *блат***.** *'жарг*. ' code language, thieves argot', 2. These equivalents belong to the same part of speech (compare Russian adverb *авансом* and the Polish. secondary preposition *z góry*), 3. compatibility of equivalents coincides (compare Polish *komunikat* and Russian *Бюллетень*), 4. grammatically and semantically equal equivalents are also equal, pragmatically (...). It would certainly be naïve to rely on the principle that a language unit in one language corresponds to an identical unit in another language semantically, grammatically and pragmatically. Therefore, one needs to take into account the fact that in a multilingual dictionary a given language unit may correspond to two or more equivalents in another language, or that the equivalent in another language might not be a complete equivalent (that is, it will be different in one of the above mentioned characteristics) (...). The information about the difference of the equivalent must be given *expressis verbis*."

An obvious example of the absence of pragmatic equivalence is the translation of the forms of address. The choice of the appropriate form of address has remained one of the main difficulties in learning a foreign language at the level of its pragmatics. Therefore, we believe that an important task of modern dictionaries is to include pragmatic information in a dictionary entry. This aspect of lexicographic work is referred to by Koseska-Toszewa, Satoła-Staśkowiak, Sosnowski (2013):

> "The analysis of monolingual and bilingual dictionaries of Polish, Russian, and Bulgarian yielded the following conclusions. Firstly, monolingual dictionaries often describe the meaning of forms of address including their scope of use in a given culture. Nevertheless, the descriptions are insufficient to use given lexemes in real speech in accordance with the usage patterns sanctioned by a given culture (…). Secondly, in traditional bilingual dictionaries forms of address are usually disregarded and the user does not receive the information about the correct patterns of use of a given lexical unit, which would conform to the norms currently accepted by the society".

5. CULTUREMES: ALWAYS NON-EQUIVALENT?

These "patterns sanctioned by a given culture" belong to a whole group of lexical units, starting from clichés in a particular language, through PUs, stylistically marked vocabulary, stereotypes of linguistic consciousness and underlying word semantics we can find in some items. We face, however, the challenge of translating culturemes. The term "cultureme" itself was created outside the boundaries of linguistics, in the cultural theory of S. Lem, in which it describes, first of all, the minimal, indivisible units of culture: rituals, values, and stereotypes (cf. Jaskot & Ganoshenko, 2015).

In modern linguocultural research the term "cultureme" is a hotly debated topic and demonstrates various approaches to its content. V. Gak (1998) considers cultureme "as a sign of culture that also has a linguistic expression". A. Vezhbitskaia (1999) regards cultureme as "an integrated interlevel unit, the form of which is the unity of a sign and language meaning, while the content – the unity of language meaning and cultural value". The application of various translation techniques create a special linguistic and translation meaning of a cultureme, based on various relationships of equivalence (Gusarov, 2002): signification (methods of transcription, transliteration, calquing), semantisation (a method of descriptive translation), reference (a method of elimination of national cultural specificity, descriptive translation), syntagmatics (a method of translation periphrasis) and functionality (a method of approximate translation, descriptive translation).

Since separate words seem to be non-equivalent due to their cultural connotations, such as, for instance, the names of former Spanish currency *duro (5 pesetas), talego (1000 pesetas), boniato (5000 pesetas)* then how can more extensive units such as PU be translated, taking into account the need for equivalency?

6. THE EQUIVALENCE OF CULTURE-LABELED ITEMS: THE PHRASEOLOGICAL UNITS

Assuming that every PU in a particular language is at the same time a cultureme would lead us nowhere, as no translation of PUs would be possible. It's high time, however, we said what a PU is. Gläser (2001) finds it to be a very convenient term that covers several types of multi-word units, defined as:

> "(…) a lexicalised, reproducible bilexemic or polylexemic word group in common use, which has relative syntactic and semantic stability, may be idiomatised, may carry connotations, and may have an emphatic or intensifying function in a text."

Indeed, the possibility (but not the requirement) of idiomatisation as well as the ability to carry connotations (cultural ones, among others) is one of the main problems encountered during the process of making PU-dictionaries, when a cognitive approach is chosen. After all, not the form, but the semantic content is the priority. Here we come back to the main problem that has been already mentioned by A. Luchyk, the issue regarding a word: what are the boundaries of a word? The same question concerns a PU: what are its formal boundaries? Once we determine them (by choosing the form, which is so important for e-databases and corpora) the notion of equivalence appears. The equivalence variables, however, seem to be the key to an effective e-tool for translating PU. A clear description of equivalence variables (pragmatics, language register, inference, cultural background and so on) is a challenge. Nonetheless, is the translation of PUs really needed? These three

examples retrieved from the Polish-Bulgarian-Russian corpus show how translators coped with this problem. PUs are marked in bold:

(Polish) Najgorsze, że to, iż jest śmiertelny, okazuje się niespodziewanie, **w tym właśnie sęk!**
(Bulgarian) Лошото е, че той понякога е внезапно смъртен, това е неприятното!
(Russian) Плохо то, что он иногда внезапно смертен, **вот в чем фокус!**

and

(Polish) Matki i wychowawczynie - nie żadne lalkowate ślicznotki ze słodkimi ślepkami.
(Bulgarian) Никакви превзети дамички, **никакво вртене на очи**!
(Russian) Только не сентиментальные дамы, не те, что **строят глазки**.

Is the expressiveness and positive or negative connotation of the sentence lost where a phraseological unit is not working in the translated text? (cf. Kuzmin, 2007) Not always, but is the translated text a Polish text of a Russian author or a polonised Russian text in Polish? Thus, is there a link missing?

7. CONCLUSIONS

Translating PUs remains one of the most difficult task for the translator due to their

a) cross-linguistic definition being still unclear,

b) powerful culture-anchored meaning,

c) stylistic and connotative functions.

In the era of multilingual corpora and dictionaries based on these corpora, a cross-linguistic definition of a PU is needed. It seems that such a definition, because of technical limitations and a need for a clear formula for translation e-tools, is likely to start from the form (as clear PU boundaries are required) rather than from the meaning. A *parole*-approach to the translation of PUs, which has been adopted till now, helps to deal with the equivalence issue (as it works on the word level), but the formal (and therefore semantic) extension of PUs requires a different approach. However, a *langue*-approach to PUs seems to be jeopardized by their cultural-anchoring but, at the same time, the cultural background and their pragmatic load are the most important variables (the "z" factor as mentioned in 4.3.) that should be taken into account when we aim to equivalently translate a PU.

References

GAK, V. G., 1998. *IAzykovye prebrazovaniia*. Moskva: IAzyki russkoĭ kul´tury.

GLÄSER, R., 2011. The stylistic potential of phraseological units in the light of genre analysis. Oxford: Oxford University Press.

GUSAROV, D. A., 2002. *IAzykovoe vyrazhenie smysla v usloviiakh lingvoėticheskogo bar´era* (MA thesis). Moskva: MGU

IVANOV, A. O., 2006. *Bezėkvivalentnaia leksika*. Sankt Peterburg: Izdatel´stvo SpbGU.

JASKOT, M. & GANOSHENKO, Y., 2015. Culturemes and non-equivalent lexis in dictionaries. Presented during the scientific conference "Leksykografia polska, ukraińska, bułgarska: słowniki tradycyjne i elektroniczne" organized by the Institute of Slavic Studies of the Polish Academy of Sciences on 13 and 14 of November 2014 in Warsaw. [in press]

KISIEL, A., SATOŁA-STAŚKOWIAK, J., & SOSNOWSKI, W., 2014. O rabote nad mnogoiazychnym slovarëm. *Prykladna linhvistyka ta linhvistychni tekhnolohiï* (MEGALING-2012), p. 111-121.

KOSESKA-TOSZEWA, V., SATOŁA-STAŚKOWIAK, J., & SOSNOWSKI, W., 2013. From the problems of dictionaries and multi-lingual corpora. *Cognitive Studies/Études cognitives*, 13, p. 113.

KUZMIN, S. S., 2007. *Idiomatic Translation from Russian into English (Theory and Practice)*. Moscow: Flinta-Nauka.

LUCHYK, A., & ANTONOVA, O., 2012. *Pol's'ko-ukraïns'kyĭ slovnyk ekvivalentiv slova.* (A. Kisiel & V. Koseska-Toszewa, eds.). Kyïv: Ukraïns'kyĭ movno-informatsiĭnyĭ fond NAN Ukraïny, Natsional'nyĭ Universytet «Kyievo-mohylians'ka Akademiia», Instytut Slavistyky Pol's'koï Akademiï Nauk.

SUM-20 *Словник української мови у 20 т.*, 2010. В.А. Широков. ed. Київ: Наукова Думка.

VEZHBITSKAIA, A., 1999. *Semanticheskie universalii i opisanie iazykov*. Moskva: IAzyki russkoĭ kul´tury.

WIEGAND, H. E., 2011. Equivalence in Bilingual Lexicography: Criticism and Suggestions. *Lexikos* [online], v. 12, Oct. 2011. Available at: http://lexikos.journals.ac.za/pub/article/view/772/362. [Accessed 25 July 2015]. doi:http://dx.doi.org/10.5788/12-0-772

FRASEOGRAFÍA Y LINGÜÍSTICA DE CORPUS: SOBRE EL TRATAMIENTO DE LOCUCIONES VERBALES EN LA NUEVA EDICIÓN DEL *DICCIONARIO DE LA LENGUA ESPAÑOLA*

Jorge Leiva Rojo
Universidad de Málaga,
Andalucía Tech
Grupo de investigación
«Lexicografía y traducción»
leiva@uma.es

Resumen[126]

En los últimos años han aparecido multitud de trabajos en los que se aborda el estudio de la fraseología tanto en repertorios bilingües —caso de Corpas Pastor (1996*b*) o Santamaría Pérez (1998, 2001)— como monolingües —sirvan de ejemplo las contribuciones de Ettinger (1982), Carneado Moré (1985), Montoro del Arco (2004) u Olímpio de Oliveira Silva (2007)—. En todos ellos se pone de manifiesto con frecuencia la necesidad imperiosa de contar con el apoyo de la lingüística de corpus para crear diccionarios en los que el tratamiento fraseológico se base en evidencias reales y no en suposiciones o en ejemplos creados por el lexicógrafo de turno.

Las unidades fraseológicas, que según algunos estudiosos es el nivel más alto de conocimiento posible de toda lengua, han recibido tradicionalmente un tratamiento dispar en los repertorios lexicográficos, bien por no incluir lemas en los que se refleje la variación fraseológica, bien por no acotar con precisión su uso —por ejemplo, si han caído en desuso o si son propias de usos coloquiales o incluso vulgares de la lengua— o bien por ofrecer significados que no se ajustan a la realidad o que la acotan de forma parcial.

El objetivo de nuestro trabajo es doble: de una parte, pretendemos analizar la forma en que ha evolucionado el tratamiento lexicográfico que se da en la nueva versión del *Diccionario de la lengua española*, de la Real Academia Española, en comparación con la versión anterior, la

[126] El presente trabajo ha sido realizado en el seno de los proyectos *TRADICOR: Sistema de gestión de corpus para la innovación didáctica en traducción e interpretación* (PIE 13 054), *TERMITUR: Diccionario inteligente TERMInológico para el sector TURístico (alemán-inglés-español)* (HUM2754, 2014-2017. Junta de Andalucía), *INTELITERM: Sistema inteligente de gestión terminológica para traductores* (FFI2012-38881, 2012-2015, Ministerio de Ciencia y Tecnología) y *EXPERT: EXPloiting Empirical appRoaches to Translation* (317471-FP7-PEOPLE-2012-ITN, Comisión Europea).

vigésimosegunda. Para ello, seleccionaremos algunas las locuciones verbales, cuyo tratamiento se compararán a su vez con el que se les da en otros repertorios lexicográficos monolingües. De otra parte, mediante el uso de los corpus de textos electrónicos de libre acceso —el *Corpus del español del siglo XXI* (CORPES) y el *Corpus del español* (BYU-CE)— se intentará establecer cuál es el uso real que tienen tales locuciones verbales en la lengua que recogen los corpus analizados. Se pretende, en definitiva, comprobar si los corpus de textos *sancionan* en la práctica el uso teórico que se les presupone en los diccionarios a las locuciones verbales en cuestión. Finalmente, se propondrán soluciones de mejora, tanto en lo referente a la variación fraseológica como a las acepciones que se ofrezcan de ella y a los ejemplos que se proporcionen, con objeto de sacar el máximo partido a las situaciones reales de la lengua que contienen los corpus de textos empleados como soporte.

1. EL TRATAMIENTO FRASEOLÓGICO EN EL *DICCIONARIO DE LA LENGUA ESPAÑOLA*

Tal y como figura en las páginas introductorias del *Diccionario de la lengua española* (DRAE23), la nueva edición de esta obra, la vigésimotercera, ha supuesto con respecto a la anterior, aparecida en 2001, una «incesante labor de adición, enmienda y, en su caso, supresión de artículos y acepciones, así como de mejora de toda la información complementaria que incluyen» (DRAE23: XLIII). Dicha labor, en lo tocante a la fraseología, se traduce en varios elementos que se mencionan en las antedichas páginas introductorias:[127]

1.º. Como consecuencia de la adaptación a las nuevas normas ortográficas dictadas en la *Ortografía de la lengua Española*, las locuciones latinas «aparecen ahora en letra cursiva y sin las tildes que podrían corresponderles: ***corpore insepulto*** en lugar de **córpore insepulto**» (DRAE23: XLVI).

2.º. Se afirma que «[l]a marca "fr." (frase) se ha sustituido en esta edición por "loc. verb." (locución verbal)» (DRAE23: XLV), de forma que, a modo de ejemplo, podemos ver cómo la locución verbal *hablar por los codos* (que en la edición anterior era 'fr. coloq. Hablar demasiado' (DRAE22), ahora pasa a rezar de la siguiente forma: 'loc. verb. coloq. Hablar demasiado' (DRAE23). Consideramos acertada esta medida por ser la etiqueta *loc. verb.* mucho más específica que *fr.* De todas formas, a pesar de lo afirmado, sorprende comprobar que *fr.* sigue utilizándose en la lista de abreviaturas y signos del DRAE23 como desarrollo de dos voces: la ya sabida *frase* y *francés*.

3.º. El aspecto quizá más llamativo hace referencia a la forma de organizar las unidades fraseológicas en el DRAE23, pues es ahí donde se hace patente que, al menos en cuanto a catalogación, no se siguen unos criterios claros. Sirva como ejemplo de ello el que las locuciones nominales *aceite virgen*, *buena mano* y *agua de cerrajas* aparecen para ejemplificar lo que en el *Diccionario* denominan «combinaciones estables de un elemento sustantivo con otras palabras» (DRAE23, LII). Llama la atención, igualmente, el que, a continuación de tratar estas combinaciones de palabras, se hable de «locuciones y expresiones» (DRAE23, LII-LIII), dos denominaciones bajo las que se recogen, tal y como se desprende de los ejemplos con que se acompaña la explicación, lo que en realidad son locuciones adjetivas (*como el perro y el gato*), locuciones verbales (*tomar el pelo*, *distinguir lo blanco de lo negro*, *comer vivo* y *haberla hecho buena*), locuciones adverbiales (*por sí o por no*) y formulas rutinarias (*Tanto mejor* o *Estaría bueno*).

[127] La información que figura a continuación es la más relevante en lo referente a la información fraseológica que se da en las páginas introductorias del DRAE23, pero no es la única, ya que también aparecen menciones a elementos fraseológicos en el apartado «Artículos de muestra» (DRAE23: XLVII-L).

En lo que respecta a la lista de abreviaturas y de signos empleados (DRAE23: LV-LVIII), se observa que aparecen abreviaturas para *expresión* (*expr.*), la ya mencionada *frase* (*fr.*) y *locución* (*loc.*). También figura la abreviatura para *fórmula* (*Fórm.*), pero no parece que se emplee para fines fraseológicos. En otro orden de cosas, no hay una abreviatura para *colocación*, como tampoco la hay para *refrán*, *paremia* ni *enunciado fraseológico*, por citar algunos tipos más de unidades fraseológicas.

Como hemos apuntado, el tratamiento fraseológico de la nueva edición del *Diccionario de la lengua española* no parece diferir en demasía del de la vigésimosegunda edición: en lo que respecta a la macroestructura lexicográfica, los cambios parecen ser mínimos y, en cualquier caso, no totalmente coherentes (caso paradigmático es la supuesta eliminación de la abreviatura *fr.*, que finalmente es parcial solo).

El hecho de que no haya una abreviatura en el DRAE23 para *colocación* (como tampoco la había en el DRAE22) presagia que la esfera de las colocaciones ha quedado fuera de la nueva edición del *Diccionario*, como pasaba en la edición anterior. Así lo atestiguaba un trabajo de González Aguiar (2006: 226-227), en el que comprobaba, de las colocaciones empleadas por Corpas Pastor para ejemplificar su clasificación de colocaciones, cuáles estaban presentes en el DRAE22. De las seis colocaciones comprobadas, tres estaban ausentes[128] (*relación estrecha*, *fracasar estrepitosamente*,[129] *firmemente convencido*), mientras que de las tres restantes (*correr un rumor*, *desempeñar un cargo, tableta de chocolate*) se da información, aunque no se las considera unidades fraseológicas. En el DRAE23, por su parte, se advierten pocas variaciones con respecto a las seis colocaciones de las que hablaba González Aguiar: siguen estando ausentes las mismas que ya lo estaban en el DRAE22, mientras que, en lo referente a las tres que aparecían, la formulación de *correr un rumor* permanece invariable, mientras que solamente se han aplicado cambios menores a la expresión de las definiciones de dos locuciones (*desempeñar un cargo* y *tableta de chocolate*), aunque dichos cambios son irrelevantes desde el punto de vista fraseográfico.

En otro orden de cosas, las sucesivas ediciones del *Diccionario* a partir del año 1970 eliminaron de forma progresiva prácticamente todo rastro de paremias en general y de refranes en particular. Decimos *prácticamente* porque, a pesar de todo, siguen dándose casos de refranes en el DRAE.[130] Sirva como ejemplo de ello el refrán *La ocasión la pintan calva* ('expr. U. para indicar que se deben aprovechar las oportunidades cuando se presentan', DRAE23), que, según el Refranero multilingüe del Centro Virtual Cervantes es una variante del refrán *A la ocasión la pintan calva* (Centro Virtual Cervantes, 2015).

2. LAS LOCUCIONES VERBALES EN LA NUEVA EDICIÓN DEL *DICCIONARIO*: EL CASO DE *DIOS*

En un conciso pero interesante trabajo, Garrido Moraga (1990) realiza un estudio sobre el tratamiento fraseológico que se ha hecho del lema *dios* en catorce ediciones del *Diccionario*

[128] Son cuatro en realidad las colocaciones que González Aguiar consideraba ausentes: a las tres mencionadas añade *tableta de chocolate*, de la que, sin embargo, sí se da información en el DRAE22 ('2. Pastilla de chocolate plana y rectangular'), por lo que hemos enmendado esta ausencia.

[129] *Fracasar estrepitosamente* no aparece en el DRAE22, pero sí aparece «fracaso estrepitoso» como ejemplo de uso bajo el lema *estrepitoso* tanto en el DRAE22 como en el DRAE23.

[130] Llamativo era el caso del refrán *Mi hija hermosa, el lunes a Toro y el martes a Zamora* ('fr. que se decía de las mujeres andariegas y amigas de hallarse en todas las diversiones', DRAE21), que se eliminó del DRAE22 y no está presente tampoco en el DRAE23. Este era, según afirma Lledó Cunill (2004: 126), «uno de los pocos refranes que todavía se conservaban en el *DRAE*, desde que se puso en práctica en la edición de 1970 la expulsión sistemática de la gran cantidad que se recogía desde el *Diccionario de Autoridades* [sic]».

de la lengua española. La razón para seleccionar este lema, afirma, es la «escasa y nula variación en la definición del término a lo largo de los siglos» (Garrido Moraga, 1990: 11). A pesar de que sí se han producido cambios reseñables en la definición del lema, especialmente en la segunda de las acepciones (véase tabla n.º 1), consideramos que sigue siendo un lema interesante desde el punto de vista fraseológico.[131]

DRAE20 (1984)	**dios.** (Del lat. *deus*) n. p. m. Nombre sagrado del Supremo Ser, Criador del universo, que lo conserva y rige por su providencia. | **2.** m. Cualquiera de las falsas deidades a que dan o han dado culto las diversas religiones, como *el* DIOS *Apolo* o *el* DIOS *Marte* […], etc.
DRAE21 (1992)	**dios.** (Del lat. *deus*) n. p. m. Nombre sagrado del Supremo Ser, Criador del universo, que lo conserva y rige por su providencia. | **2.** m. Cualquiera de las deidades a que dan o han dado culto las diversas religiones, como *el* DIOS *Apolo* o *el* DIOS *Marte* […], etc.
DRAE22 (2001)	**dios.** (Del lat. *deus*). **1.** m. Ser supremo que en las religiones monoteístas es considerado hacedor del universo. ORTOGR. Escr. con may. inicial. **2.** m. Deidad a que dan o han dado culto las diversas religiones.
DRAE23 (2014)	**dios, sa.** (Del lat. *deus*) Escr. con may. inicial en acep. 1. c. nombre propio antonomástico). m. **1**. Ser supremo que en las religiones monoteístas es considerado hacedor del universo. **O** m. y f. **2**. Deidad a que dan o han dado culto las diversas religiones politeístas.

Tabla 1. Acepciones del lema *dios* en las cuatro últimas ediciones del *Diccionario de la lengua española*

En lo que respecta al tratamiento fraseológico que se da de este lema en la vigésimotercera edición del *Diccionario*, resulta llamativo el alto número de unidades fraseológicas presentes, 134, número algo inferior a las presentes en la edición anterior, donde aparecía un total de 136 unidades fraseológicas. Esta diferencia de unidades se debe a la desaparición en la última edición de cuatro unidades fraseológicas (las locuciones sustantivas *Dios Chico* y *Dios Grande*, la locución verbal *creer alguien en Dios a macha martillo* [o *a puño cerrado*] y la fórmula rutinaria *De Dios, el medio* y a la aparición de la frase elativa *que Dios tenga en la gloria* (o *en su* [*santa*] *gloria*) y de la locución nominal *ni Dios*. Resulta llamativo que ninguna de las cuatro unidades fraseológicas que desaparecen en este lema en el DRAE23 estuvieran etiquetadas como *ant.* (abreviatura de anticuado) en el DRAE22. Es más, la única del total de 136 unidades que llevaba esta marca en la edición de 2001 sigue estando presente en la edición de 2014: la locución adverbial *de Dios en ayuso* ('Loc. adv. desus. De Dios abajo', DRAE23).

[131] No es este el único punto de vista que resulta interesante. Aunque no es objeto de este trabajo hacer un análisis ideológico de las definiciones que se da de *dios*, consideramos muy destacable la *limpieza religiosa* que se da en las sucesivas ediciones del *Diccionario*, como puede verse en la tabla 1. A pesar de la revisión efectuada en las acepciones, una lectura más profunda desvela que los cambios en cuanto a la fraseología son menores, y que a ellos la limpieza no ha llegado aún. Sirva como ejemplo de ello la acepción de la locución verbal *poner a Dios por testigo* ('Invocar su santo nombre para aseverar lo que se dice', DRAE23), que ha permanecido con idéntica formulación en DRAE20, DRAE21 y DRAE22 (Real Academia Española, 2015), así como, por citar dos ediciones más, en las ediciones décimoquinta y duodécima, publicadas en 1925 y 1884 respectivamente (Instituto de Investigación Rafael Lapesa de la Real Academia Española, 2013).

2.1. ¿Son todas las que están?

En lo referente a las locuciones verbales, que es el objeto central de este trabajo, aparece en DRAE23 un total de 42 locuciones verbales.[132] En un primer momento, llama la atención que son escasas las locuciones que tienen un sentido negativo; por citar algunas, mencionamos aquí *dejar Dios de su mano a alguien*, *irse mucho con Dios*, *ser algo un contra Dios* o *llamar Dios a alguien por un camino* —esta última es la única locución en la que se advierte que su uso mayoritario es negativo—. Igualmente, son relativamente numerosas las que guardan relación con el ámbito eclesiástico estrictamente hablando, pero que tienen una frecuencia de uso, como indican los corpus consultados, muy reducida e incluso inexistente (véase tabla 2).

	CORPES	**BYU-CE**
dar a Dios a alguien	0	0
hablar con Dios	138	11
ofender a Dios	34	33
poner a Dios delante de los ojos	0	0
recibir a Dios	7	2

Tabla 2. Apariciones de cinco locuciones verbales en dos corpus electrónicos

Con respecto a otra de las locuciones verbales que registra el DRAE23, *descreer de Dios* ('Renegar de Dios'), consideramos que no se trata en realidad de una locución, como prueba el hecho, en primer lugar, de que su frecuencia de uso no es excesivamente elevada (dos apariciones en el CORPES y una sola en el BYU-CE) y, en segundo, a pesar de que no sea requisito imprescindible para considerarse unidad fraseológica, que su significado no es idiomático, sino la suma de sus significados evidentes —véase para ello, además, la definición de *descreer* ('Faltar a la fe, dejar de creer', DRAE23)—.

Finalmente, llama la atención el hecho de que se dé como locución verbal lo que en realidad debería considerarse locución adverbial: *hacer algo como Dios manda* ('Hacerlo bien, con exactitud y acierto', DRAE23). Para considerar que esta atribución es errónea —debida probablemente a las dificultades de delimitación que plantean las locuciones verbales y adverbiales (cf. García-Page, 2007: 137)— nos basamos en el resultado que arroja la consulta en el BYU-CE de la secuencia «como Dios manda» (campo Words) + «[v]» (campo Collocates) en hasta nueve posiciones a la izquierda. De esa forma, podemos ver cómo el verbo que acompaña a la cadena buscada no es solo el verbo *hacer* —que con cinco apariciones no es el más frecuente siquiera—, sino que también se dan casos con 71 verbos más, entre los que están los verbos *ser* (19 apariciones) o *vivir* (6 apariciones) y, en menor número, verbos tan dispares como *amar*, *celebrar*, *comer*, *esquiar*, *hablar*, *querer*, *rezar* o *trabajar*. Esta opinión que mostramos aquí se ve refrendada por el hecho de que dos diccionarios fraseológicos, DFDEA y DFEM, consideran *como Dios manda* una locución adverbial.

[132] Hemos decidido no hacer distinción entre las locuciones verbales y las clausales de la clasificación de Corpas Pastor (1996*a*), por no considerarla relevante para los fines de este trabajo. A este respecto, resulta de interés la argumentación ofrecida por Penadés Martínez (2006: 1201). En otro orden de cosas, para consultar el listado de todas las locuciones verbales registradas en el lema *dios* del DRAE23, véase el Apéndice n.º 1.

2.2. ¿Están todas las que son?

Como es de esperar, es improbable que un diccionario (y menos uno en versión impresa) comprenda todas las locuciones verbales del lema *dios* en lengua española. A pesar de esta limitación, consideramos que el DRAE23 ha dejado olvidadas algunas locuciones verbales que sí podrían estar. Una comparación con los dos diccionarios que acabamos de mencionar arroja los siguientes resultados: los dos comprenden muchas menos unidades fraseológicas bajo el ya sabido lema *dios* (60 en el DFDEA, de las que 10 son locuciones verbales, y 41 unidades —14 locuciones verbales— en el DFEM frente a, recordemos, las 134 unidades fraseológicas y 42 locuciones verbales del DRAE23).

De la comparación de los tres repertorios lexicográficos extraemos información relevante: de una parte, el alto número de unidades que contiene el DRAE23, sobre todo si se compara con diccionarios específicos de fraseología, como es el caso del DFDEA y DFEM. De otra, que la divergencia de cifras es aún más relevante si se tiene en cuenta que los dos últimos diccionarios citados incluyen locuciones verbales que no están presentes en el DRAE23. De esta forma, en el DFDEA se reflejan las locuciones verbales *acordarse Dios* [*de alguien*], [*costar*] *Dios y ayuda* y *llevarse Dios* [*a alguien*]; mayores son las divergencias con respecto al DFEM: son cinco las locuciones verbales que aparecen (*cagarse u. p. en Dios*, [*andar/estar u. p.*] *como Dios la echó/trajo al mundo*, *costarle u. c. Dios y ayuda a alguien*, [*ser u. p. un*] *alma de Dios* y [*andar/ir u. p.*] *por esos mundos de Dios*). Con respecto a esta última se podría alegar a favor del DRAE23 que, de acuerdo con los criterios de organización lexicográfica (DRAE23: LII-LIII), debería figurar bajo el lema *mundo*. Efectivamente, allí encontramos las locuciones adverbiales *por el mundo adelante/por esos mundos* ('Sin rumbo fijo o sin lugar determinado. *Irse*, *marcharse*, *andar por esos mundos*', DRAE23), pero, sin embargo, no está presente *de Dios* como complemento de *mundo*.[133]

3. CONCLUSIONES

Debido a limitaciones de espacio, y a modo de muy someras conclusiones, consideramos que, como consecuencia de lo relativamente inamovible que parecen ser las ediciones del *Diccionario* con respecto a sus antedecesoras —al menos en lo que al lema *dios* respecta—, sería necesario revisar las unidades fraseológicas que recoge, de forma que se pudieran incorporar algunas unidades que se emplean con frecuencia —*costar Dios y ayuda*, por ejemplo, con 7 y 14 apariciones en CORPES y BYU-CE respectivamente, podría ser uno de ellos— y se eliminaran aquellas que han caído en desuso. Se trata, por lo tanto, de un trabajo incipiente que esperamos poder continuar en futuros trabajos.

[133] Y eso a pesar de que, aparentemente, el sintagma *de Dios* aparece con frecuencia unido a *por esos mundos* (*por esos mundos*: 41 apariciones, *por esos mundos de Dios*: 17 apariciones en BYU-CE). En otro orden de cosas, la búsqueda en el corpus confirma que los verbos con los que suele aparecer expresan, habitualmente, movimiento (*andar*, *ir*, *correr*).

APÉNDICE N.º 1. LOCUCIONES ADVERBIALES REGISTRADAS EN EL LEMA *DIOS* EN EL DRAE23

amanecer Dios
bendecir Dios a alguien
clamar a Dios
dar a Dios a alguien
darse a Dios y a los santos
dejar Dios de su mano a alguien
dejarlo a Dios
descreer de Dios
dormir en Dios
estar alguien con Dios
estar de Dios algo
estar alguien fuera de Dios
gloriarse en Dios
gozar alguien de Dios
hablar con Dios
hablar Dios a alguien
hacer algo como Dios manda
herir Dios a alguien
irse alguien bendito de Dios
irse alguien con Dios
irse mucho con Dios
la de Dios es Cristo
llamar a Dios de tú
llamar Dios a alguien
llamar Dios a alguien a juicio, o *para sí*
llamar Dios a alguien por un camino
no haber para alguien más Dios ni Santa María que algo
no servir a Dios ni al diablo alguien o algo
no tener alguien sobre qué Dios le llueva
ofender a Dios
poner a Dios delante de los ojos
poner a Dios por testigo
ponerse bien con Dios
recibir a Dios
ser algo para alabar a Dios
ser algo un contra Dios
tener Dios a alguien de su mano
tentar a Dios alguien
tomar a Dios los puertos
tomarse con Dios
tratar alguien con Dios
venir Dios a ver a alguien

Bibliografía

BYU-CE = DAVIES, M.. (2002-). *Corpus del Español: 100 million words, 1200s-1900s*. [en línea]. Disponible en: http://www.corpusdelespanol.org [consultado el 23 de junio de 2015].

CARNEADO MORÉ, Z. V, 1985. *La fraseología en los diccionarios cubanos*. La Habana: Editorial de Ciencias Sociales.

CENTRO VIRTUAL CERVANTES, 2015. *Refranero multilingüe*. [en línea]. Disponible en: http://cvc.cervantes.es/lengua/refranero/ficha.aspx?Par=58079&Lng=0 [consultado el 23 de junio de 2015].

CORPAS PASTOR, G., 1996*b*. La fraseología en los diccionarios bilingües. En: Manuel ALVAR EZQUERRA, ed. *Estudios de Historia de la Lexicografía del Español*. Málaga: Servicio de Publicaciones de la Universidad, p. 167-182.

CORPAS PASTOR, G., 1996*b*. *Manual de fraseología española*. Madrid: Gredos.

CORPES-CE = REAL ACADEMIA ESPAÑOLA. *Corpus del español del siglo XXI (CORPES)*. [en línea]. Disponible en: < http://web.frl.es/CORPES/view/inicioExterno.view> [consultado el 23 de junio de 2015].

DFDEA = SECO, M.; ANDRÉS, O.; RAMOS,G. 2004. *Diccionario fraseológico documentado del español actual: locuciones y modismos españoles*. Madrid: Aguilar.

DFEM = VARELA, Fernando; Hugo KUBARTH, 1994. *Diccionario fraseológico del español moderno*. Madrid: Gredos.

DRAE20 = REAL ACADEMIA ESPAÑOLA, 1984. *Diccionario de la lengua española*, vigésima edición. Madrid: Espasa Calpe.

DRAE21 = REAL ACADEMIA ESPAÑOLA, 1992. *Diccionario de la lengua española*, vigésimoprimera edición. Madrid: Espasa Calpe.

DRAE22 = REAL ACADEMIA ESPAÑOLA, 2001. *Diccionario de la lengua española*, vigésimosegunda edición. Madrid: Espasa Calpe.

DRAE23 = REAL ACADEMIA ESPAÑOLA, 2014. *Diccionario de la lengua española*, vigésimotercera edición. Madrid: Espasa Calpe.

ETTINGER, S., 1982. Formación de palabras y fraseología en la lexicografía. En: G. Haensch *et al.*, eds. *La lexicografía. De la lingüística teórica a la lexicografía práctica*. Madrid: Gredos, p. 233-258.

GARCÍA-PAGE, M., 2007. Esquemas sintácticos de formación de locuciones adverbiales. *Moenia* 13, p. 121-144.

GARRIDO MORAGA, A. M. 1990. De nuevo sobre fraseología en los diccionarios: una cala en el DRAE. *Revista del Instituto de Lengua y Cultura Españolas*, 6(1), 7-18.

GONZÁLEZ AGUIAR, M. I., 2006. La definición lexicográfica de las unidades fraseológicas: la aplicación de modelos formales. En Margarita ALONSO RAMOS (ed.). *Diccionarios y fraseología*. Anexos de *Revista de Lexicografía* 3. A Coruña: Universidad, p. 221-234.

INSTITUTO DE INVESTIGACIÓN RAFAEL LAPESA DE LA REAL ACADEMIA ESPAÑOLA, 2013. *Mapa de diccionarios* [en línea]. Disponible en: http://web.frl.es/ntllet [consultado el 23 de junio de 2015].

LLEDÒ CUNILL, E., 2004. Presencia femenina suprimida, modificada (respecto a la edición de 1992 del diccionario) o nueva. En: Eulàlia LLEDÓ CUNILL (coord.), M.ª Ángeles CALERO FERNÁNDEZ y Esther FORGAS BERDER. *De mujeres y diccionarios. Evolución de lo femenino en la 22.ª edición del* DRAE. Madrid: Instituto de la Mujer (Ministerio de Trabajo y Asuntos Sociales, p. 97-196.

MONTORO DEL ARCO, E. T., 2004. La variación fraseológica y el diccionario. *De Lexicografía*, p. 1000-1014.

OLÍMPIO DE OLIVEIRA SILVA, M. E., 2007. *Fraseografía teórica y práctica*. Fráncfort del Meno: Peter Lang.

PENADÉS MARTÍNEZ, I., 2006. Las locuciones interjectivas en la fraseología española. En Antonio ROLDÁN PÉREZ, coord. *Caminos actuales de la historiografía lingüística: actas del V Congreso Internacional de la Sociedad Española de Historiografía lingüística*, vol. 2, Murcia: Universidad, p. 1197-1208.

REAL ACADEMIA ESPAÑOLA, 2015. *Nuevo tesoro lexicográfico*. [en línea]. Disponible en: http://www.rae.es/recursos/diccionarios/diccionarios-anteriores-1726-1992/nuevo-tesoro-lexicografico [consultado el 23 de junio de 2015].

SANTAMARÍA PÉREZ, María Isabel, 1998. El tratamiento de las unidades fraseológicas en la lexicografía bilingüe. *ELUA. Estudios de Lingüística* (12), p. 299-318.

SANTAMARÍA PÉREZ, M. I., 2001. *Tratamiento de las unidades fraseológicas en la lexicografía bilingüe español-catalán*. Alicante: Biblioteca virtual Miguel de Cervantes.

THE COMPILATION OF AN ONLINE CORPUS-BASED BILINGUAL COLLOCATIONS DICTIONARY

Adriane Orenha-Ottaiano
Universidade Estadual Paulista "Júlio de Mesquita Filho"
adriane@ibilce.unesp.br

Keywords: collocations, collocations dictionary, corpus-based dictionary, online dictionary

Abstract

This paper aims at giving an account of the compilation of an online bilingual collocations dictionary, in the English-Portuguese and Portuguese-English directions, based on learner, parallel and reference corpora. It will bring a full range of types of collocations (verbal, noun, adjectival and adverbial collocations), and it is designed for teachers and learners of English as a second language as well as learner and professional translators, among others, in order to help them use collocations in oral and written texts more accurately and productively. Being the first corpus-based bilingual collocations dictionary in the above mentioned directions, we hope to achieve the challenge of meeting learners' collocational needs as the collocations will be selected according to learners' difficulties regarding the use of collocations.

1. INTRODUCTION

The contribution of corpora to lexicology/lexicography, as well as to phraseology/ phraseography, has already been pointed out by many researchers (Altenberg and Granger, 2002; Burger et al., 2007; Halliday *et al.,* 2004*;* Meunier; Granger, 2008; Moon, 2008; Orenha-Ottaiano, 2013; Sinclair, 2007; Teubert, 2004, 2007, among others).

Moon (2008), for instance, stated that corpora is "the great facilitator for the description of phraseology, and its use in dictionary-making has heavily influenced the ways in which lexicographical attitudes towards phraseology have developed over the last twenty-five years". Moreover, it is also true that the quality of monolingual and bilingual dictionaries has also improved, due to the methodology provided by Corpus Linguistics. The use of corpora has enabled us to identify and extract phraseological units more easily and quickly, especially collocations.

Taking that into account, and considering the fact that collocations pose a serious problem to foreign language learners and trainee translators with regard to production (either oral or written), this paper will discuss the compilation of an online bilingual

collocations dictionary, in the English-Portuguese and Portuguese-English directions, based on learner, parallel and reference corpora.

The dictionary will focus on all types of collocations (verbal, noun, adjectival and adverbial collocations), and it is designed for teachers and learners of English as a second language as well as learner and professional translators, among others, in order to help the referred audience use them more accurately and productively.

2. LITERATURE REVIEW

The term 'collocation' is commonly ascribed to Firth (1957), defined in his article entitled *Modes of Meaning*, when explaining cases of lexical syntactic co-occurrences, that is to say, "words that usually go together". Nevertheless, Handl (2008) points out that both Bartsch (2004, pp. 30) and Mitchell (1971, pp. 35) claim that the term was actually used in the 1950s by H. E. Palmer and still earlier by Otto Jespersen in 1917. In spite of that, Firth gave his contributions to the understanding of the concept, especially regarding his well-known characterization of a collocation when he asserts "you shall know a word by the company it keeps" (Firth 1957, pp.11).

In what regards to the definition of collocations, Nesselhauf (2005, pp.1) characterizes them as "[…] arbitrarily restricted lexeme combinations". The author adds that they are made up of more than one word and are lexically and/or syntactically fixed to a certain degree. Heylen and Maxwell (1994, pp. 299) give a broader definition when they mention that collocations are "cohesive, recurrent, arbitrary combinations of words which are not idioms but in which the (figurative) meaning of one part is contextually restricted to the specific combination". Hausmann (1984), in turn, draws our attention to the fact that the most important aspect of collocations is their 'status of mental disponibility as a whole, not as a creation produced ad hoc by a speaker'.

On the strength of that, it is understood that collocations pose a problem to foreign language learners with regard to production, not comprehension. Under a cognitive viewpoint, it is known that native speakers rely on a more or less fixed repertoire of phrases stored in their mental lexicon. In order to simplify production, they automatically recover them as a whole block, according to their degree of linguistic competence – and not lexeme by lexeme. That means they will not need to produce them again at the moment of their speech. As Ter-Minasova (1992, 535) argues, "phraseological units function in speech in the same way as words". The author adds that, in the process of speaking, native speakers do not simply bring separate words together in linear succession, "they also use "ready-made" units, prefabricated blocks (i.e. phraseological units) that already exist in language as a global whole and function in speech as one word". Hence, what appears to be spontaneous is actually a stereotyped fixed and repetitive speech, and if the speaker does not have a vast repertoire of these stereotyped fixed units at their disposable, their speech may not sound natural.

Due to their importance to the teaching and learning of a foreign language, we claim that learners should count on dictionaries of collocations, in order to improve their fluency. The compilation of a corpus-based bilingual collocations dictionary (English-Portuguese and Portuguese-English directions) is also justified as the number of dictionaries which focus on this phraseological unit is still scarce, especially when it comes to bilingual versions. When it comes to monolingual dictionaries, there are four excellent works for learners of English as a second or foreign language we know of: *Macmillan Collocations Dictionary for Learners of English* (Rundell, 2010), *Oxford Collocations Dictionary for Students of English* (Mcintosh et al, 2009), *LTP Dictionary of Selected Collocations* (Hill; Lewis, 1999) and

The BBI Combinatory Dictionary of English (Benson et al, 1997). However, to our knowledge, there are not any corpus-based bilingual collocations dictionaries in the English-Portuguese and Portuguese-English directions, which may fulfill this gap.

In what concerns the taxonomy of collocations used in this study, our work is founded on Hausmann's (1985): verbal, nominal, adjectival and adverbial collocations. With respect to the elements of a collocation, Hausmann (1985) points out two of them: a basis and a collocate, each with different semantic status. Actually, there is a hierarchy between these two elements, as one of them (the basis) determines and the other (the collocate) is determined. Simply speaking, the basis is what we have already known and the collocate is the element we are looking for. The basis is an independent semantically autonomous element that determines which lexical patterns can combine with it. On the other hand, the collocate works as a modifier; it is semantically interpretable within a collocation, and it is chosen by a certain basis to form a collocation (Heid et al., 1991).

3. METHODOLOGY, MACRO AND MICROSTRUCTURE OF THE COLLOCATIONS DICTIONARY

Under the theoretical and methodological approach of Corpus Linguistics, the methodology of this research first relies on the extraction and analysis of collocations from the *Translation Learner Corpus* (*TLC*), a parallel corpus made up of university students' translations in the Portuguese-English direction, whose level of English varies from B2 to C1 - students' knowledge of language was identified according to the results of the *Oxford Placement Test* (Allan 2004). The original texts that comprise the corpus are newspaper articles taken from well-known Brazilian newspapers and magazines and the typology of the texts are related to current world news and the topics selected were *One year after Tsunami in Japan*; *Financial crises in Greece and in Europe*; *Unemployment*; *Elections in the US*; *Bullying*; *Marijuana Legalization* among others. The referred corpus was compiled at *Universidade Estadual Paulista* (UNESP), in Brazil, and the collocations were extracted with the help of *WordSmith Tools* (Scott 2008), which has enabled us to raise the most frequent collocational patterns and the most/least used types of collocations used by the translation learners in comparison to the original texts, the influence of the mother tongue on their choices, among other aspects.

In a second stage, based on the keywords and collocations analyzed in the *TLC*, more collocational patterns are extracted using *COCA*, *The Corpus of Contemporary American English* (Davies 2008-2012) and the *Sketch Engine* (Kilgarriff et al, 2004).

In order to include more collocations as well as to ensure dictionary users will have access to more frequent, recurrent and sophisticated collocations, we also use the frequency list from *COCA* with the purpose of extracting more patterns. The idea of having the proposed dictionary in online format will allow us to incorporate collocational information more qualitatively and quantitatively. Besides that, more examples can be included, different from conventional collocations dictionaries.

Below follows an example of some collocations extracted from the keyword *primaries*, from TLC, and some other collocational patterns from *COCA*:

PRIMARY (n.)
Verb + PRIMARY
Conduct
I took The Chronicle's advice and called an election official in New York to see how they **conduct** September **primaries**.
Enter
If you had decided to run, what do you think your chances would have been **entering** the Republican **primaries** and going through the primary process?
Establish
On Monday, former Phoenix mayor P. Johnson filed papers for a ballot initiative in 2012 that would **establish** open **primaries** in Arizona.
Hold
Because those states **held primaries**, both won by Hillary Clinton, in defiance of party rules, their delegates will not be seated at the Democratic convention.
Lose
It's a rare moment when members of Congress **lose** their **primaries**, and when they do, politicos want to know why.
Participate in
Nowhere is the support for big changes stronger than among voters who **participate in** Democratic presidential **primaries**.
Win
"We **won** primaries and people said we really didn't win them," Carville said.

Table 1. Collocations extracted from *COCA*, out of the keyword *primaries*

It is worth emphasizing that the collocations to be included in the dictionary are not only the ones generated from the keywords of the TLC, otherwise it would compromise the usefulness of the dictionary, as the most frequently used lexicon and collocations in the English language would not be fully represented in the work. Thus, we are also analyzing *COCA's* most frequent lemma list (100,000 words), based on the 450 million words that comprise the referred corpus, with a view to extract, out of the selection of the most frequent content words (nouns, verbs and adjectives), new collocational patterns from *COCA*. The purpose of this extraction and analysis is to make sure the dictionary has a wider range of collocations, taking into account the most frequent words used by English language speakers.

Besides *COCA*, we will also count on the help of *The Sketch Engine* (Kilgarriff et al, 2004), with the aim of extracting more patterns, on the strength of the list of the most frequent lemmas of *COCA*, as well as search for more contexts in which the collocations are used, in order to include them in the examples of the collocations that will comprise the dictionary.

The following methodological phase is related to the translation of the extracted English collocations into Portuguese. All the translations suggested will be checked in *Sketch Engine*, so that we again make sure the collocations translated into Portuguese are effectively used by native speakers. So far, the dictionary has more than 330 entries, and more than 3,500 collocations extracted, both in English and in Portuguese, all of them with their corresponding contexts – see some examples in table 2, or figure 1 below, taken from the online version.

As it can be noticed, the organization of the macro and microstructure of the dictionary, both printed and online version, are under construction phase and it is based on the methodology proposed by Orenha-Ottaiano (2004, 2009) for the compilation of corpus-based specialized collocations glossaries briefly presented here in this paper.

The entries, organized in alphabetical order, are being selected on the basis of frequency: the most frequent collocations extracted from the words with the highest

keyness from the TLC as well as from the most frequent words from *COCA*'s list. These entries are called basis as our work is anchored in Hausmann's theory (1985) and Heid et al's point of view (1991) in that the basis is usually what we already know and the collocate is the element we are looking for and, due to that, it should be the entry of a collocations dictionary. On that account, the entries will be the basis and the collocates will be organized right below them, also in alphabetical order, as shown in the table below:

INTERVIEW (n.) A meeting in which someone asks you questions to see whether you are right for a job or course; A meeting in which someone asks a famous person questions for a newspaper, television, etc.	**ENTREVISTA (s.)** Conferência marcada por pessoas em lugar determinado; encontro combinado; Prestação de informações ou de opiniões a um repórter de jornal, televisão etc., feita oralmente para publicação.
Verb + INTERVIEW	**Verbo + ENTREVISTA**
Conduct a ~	**Conduzir uma ~**
Last March, during the South by Southwest Music Festival in their hometown of Austin, I **conducted a** long **interview** with the " three Charlies " who lead C3: Jones, the mover and shaker behind the revitalized Lollapalooza; [...]	John Knowles do Metal Exiles, recentemente **conduziu uma entrevista** com o guitarrista/vocalista Michael Sweet, da banda de rock cristão, STRYPER.
Publish a ~	**Publicar uma ~**
Schirrmacher is chief cultural editor for the Frankfurt Allgemeine Zeitung, which **published the interview** in which Grass shares his long time secret.	Em abril de 1952, a revista LIFE **publicou uma entrevista** com Marilyn Monroe intitulado How I Stay in Shape, em português, Como eu Mantenho a Forma.
Record a ~	**Gravar uma ~**
Then, a British " Sunday Times " newspaper reporter got hold of Barayevs cell phone number, and after talking his way into the theater, he **recorded an interview** with him.	Na semana anterior àquela segunda-feira, Cott passara dias ao lado do casal. **Gravou uma entrevista** de nove horas para escrever uma reportagem de capa da Rolling Stone [...].
Get a ~	**Conseguir uma ~**
[...] I don't know if it's true is that he immediately called law enforcement once the defense investigators called to **get an interview**.	Segundo o site americano Jezebel, a candidata Lego já **conseguiu uma entrevista** por telefone.
Give a ~	**Dar uma ~**
Gordon Brown today **gave an interview** on the back of his allegations yesterday that The Sunday Times conspired with known criminals to gain access to his private accounts.	Andressa Urach **deu uma entrevista** para lá de polêmica para o "Pânico na Band" na última semana.
Adjective + INTERVIEW	**Adjetivo + ENTREVISTA**
Exclusive ~	**~ Exclusiva**
Barbara Walters sat down with Sir Elton and David for an **exclusive interview** in which they talk about the dramatic events that have made them proud fathers.	Em **entrevista exclusiva** a Folha, FHC disse que a irrigação de 100 mil hectares por ano poderá gerar 1,2 milhão de empregos [...].
Rare ~	**~ Rara**
A former CIA official is speaking out in a **rare interview**, defending controversial interrogation techniques like waterboarding.	[...] a desconfiança da cantora em relação a esses e outros temas podem ser conferidos em **entrevista rara**, concedida por Elis [...].
Live ~	**~ Ao vivo**
The show's host called in sick, and a desperate station manager recruited Diane to conduct a **live interview** with a spokeswoman from the Dairy Council about the wonders of milk and cheese.	Uma repórter desmaiou durante uma **entrevista ao vivo** em um programa americano.

Table 2. Example of a collocation entry whose entry is a noun

The table above is meant to illustrate how the entries will be organized. In what more specifically concerns this entry, we have already extracted 7 verbal collocations (verb + interview), 11 adjectival collocations (adjective + interview), 6 noun collocations (noun + interview). It can also be noted that each entry will have a definition. This decision may help in the cases that an entry has more than one meaning, and hence, generates different collocations.

As for the examples, there will be an icon ⓘ, indicating the sources of the examples for the collocations, so that the dictionary gets visually cleaner. If the user click on it, he or she will have access to the source of the example:

Figure 2. Icon indicating the source of the example

Still regarding the microstructure of the dictionary, it will bring all types of collocations, as previously mentioned: verbal, adjectival, nominal and adverbial and, the user will have the chance to choose from each type, in case he or she wants to restrict the query, as figure 3 shows, taken from the prototype of the online version of the dictionary:

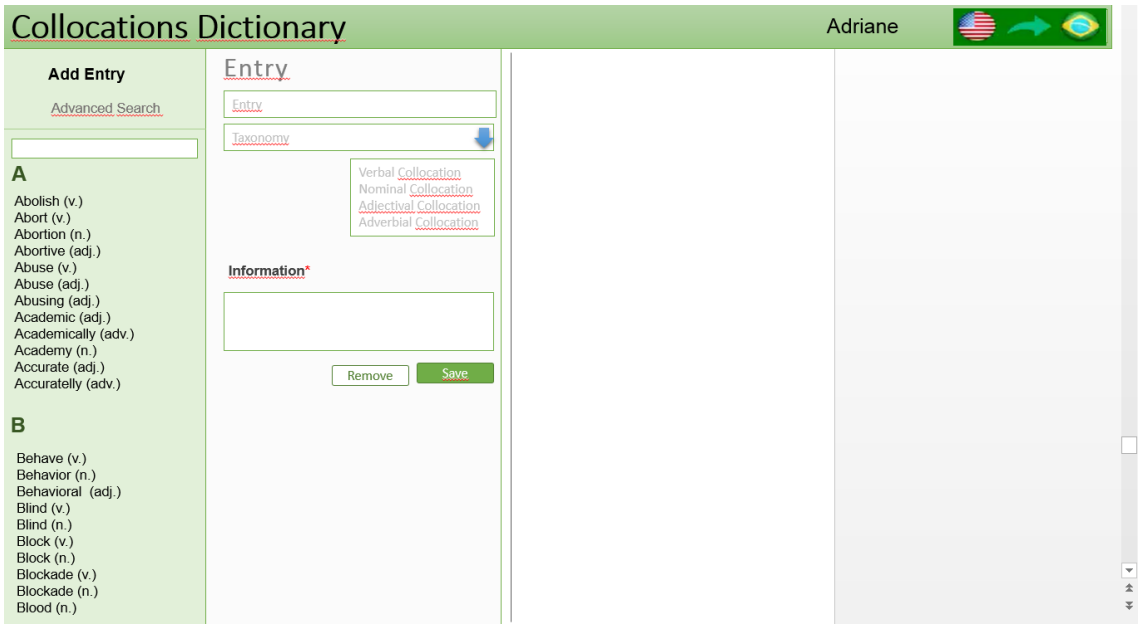


Figure 3. Screenshot of the prototype of the online version of the dictionary

On the right of the screen, it can be seen the entries of the dictionary in English. If the user clicks on the flag (Brazilian and English flag) on the top right, he or she can switch languages and see the collocations in the Portuguese direction. The screen will show one language on the left and the equivalent collocations in the other language, as illustrated below:

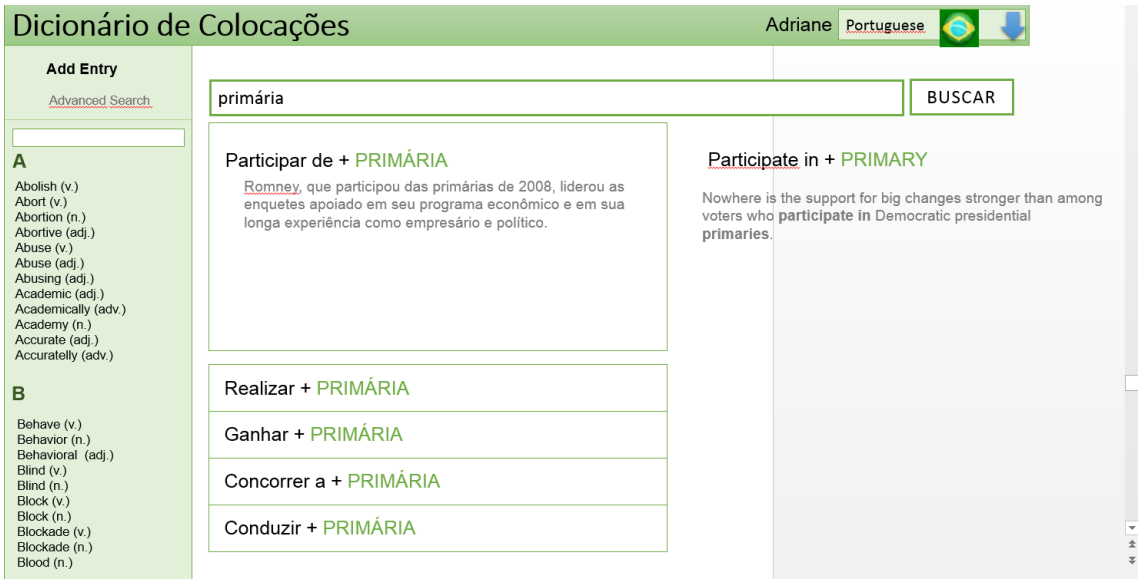


Figure 4. Screenshot showing the collocation from the Portuguese-English direction

Grounded on the importance of this work to the above mentioned target audience, and taking into account the potential it has to generate a two-version product, printed and online, we strongly believe that, considering the way it is methodologically organized, the bilingual collocations dictionary will boost the access of more people, from different parts of Brazil and abroad.

FINAL REMARKS

The theoretical and practical discussions raised from this investigation on the compilation of an online corpus-based bilingual collocations dictionary as well as its future perspectives may contribute to prompt debates and encourage research on corpus-based teaching and learning of collocations.

They may also serve to stress the relevance of support materials on collocations and the advantages of using phraseographical works in the classroom, with a view to fulfilling the needs of learners of English as a foreign language. The challenges are how to systematize pedagogical aspects concerning corpus-based teaching of phraseological lexicon, and how to develop phraseological competence and a more natural language and translation learning environment, consistent with the most frequent lexicon of a language.

Having shed some light on the potential benefits of an online corpus-based bilingual collocations dictionary and considering it to be the first collocations dictionary in the English-Portuguese and Portuguese-English directions, we hope the publication of the dictionary, printed and online version to be widely accessed and potentially useful.

References

ALLAN, D., 2004. *Oxford Placement Tests 2*. Oxford: OUP.

ALTENBERG, B. AND GRANGER, S., 2002. *Lexis in Contrast*. Amsterdam/ Philadelphia: Benjamins.

DAVIES, M., 2008-2012. *The Corpus of Contemporary American English:* 425 million words, 1990-present. [online] Available at: <http://corpus.byu.edu/coca/> [Acessed: April 20th, 2014].

BURGER, H., DOBROVOLSKIJ, D., KUHN, P. AND NORRICK, N. R. eds., 2007. *Phraseologie/Phraseology*. Ein internationales Handbuch zeitgenössischer Forschung/An International Handbook of Contemporary Research. 2 Halbbände (=HSK 28.1/2). Berlin/New York: de Gruyter.

HALLIDAY, M.A.K., TEUBERT, W., YALLOP, C. AND ČERMÁKOVÁ, A. eds., 2004. *Lexicology and Corpus Linguistics*: an introduction. London: Continuum.

HANDL, S. 2008. Essential collocations for learners of English. In F. Meunier; S. Granger, eds., 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins. 43-66.

HAUSMANN, F. J., 1985. 'Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie

des lexikographischen Beispiels'. In: H. Bergenholtz; J. Mugdan, eds., 1985. *Lexikographie und grammatik*. Tübingen: Niemeyer.

HEID, U., MARTIN, W. AND POSCH, I., 1991. An Overview of approaches towards the description of collocations. *Feasibility of standards for collocational description of lexical items*. *Eurotra 7-Report*, Stuttgart/Amsterdam.

MEUNIER, F., AND GRANGER, S. eds., 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins.

MOON. R., 2008. Dictionaries and collocations. In: F. Meunier and S. Granger, eds., 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins, pp. 247-252.

ORENHA-OTTAIANO, A., 2004. *A compilação de um glossário bilíngue de colocações, na área de jornalismo de Negócios, baseado em corpus comparável*. Unpublished Master's thesis, Universidade de São Paulo (USP), Brazil.

ORENHA-OTTAIANO, A., 2009. *Unidades fraseológicas especializadas: colocações e colocações estendidas em contratos sociais e estatutos sociais traduzidos no modo juramentado e não-juramentado*. Unpublished Ph.D. Thesis, Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Brazil.

ORENHA-OTTAIANO, A., 2013. The proposal of an electronic bilingual dictionary based on corpora. In *X International School on Lexicography*, Florence, Italy. Life Beyond Dictionaries. Ivanovo: University of Ivanovo. 1: p. 405-408.

SINCLAIR, J. McH., 2007. Data-derived multilingual lexicons. In: W. Teubert, ed., 2007. *Text corpora and multilingual lexicography*. Amsterdam/Philadelphia: John Benjamins, pp. 69-82.

TEUBERT, W., 2004. Language and Corpus Linguistics. In M.A.K. Halliday; W. Teubert; C. Yallop; A. Čermáková, eds., 2004. *Lexicology and Corpus Linguistics*: an introduction. London: Continuum.

TEUBERT, W. ed., 2007. *Text corpora and multilingual lexicography*. Amsterdam/ Philadelphia: John Benjamins.

SCOTT, M., 2008. *WordSmith Tools*, version 5.0. Liverpool: Lexical Analysis Software Ltd.

POUR UN CONTINUUM DES PHRASÈMES NON-COMPOSITIONNELS

Marie-Sophie Pausé
Université de Lorraine
ATILF-CNRS
marie-sophie.pause@univ-lorraine.fr

Abstract

S'il est aujourd'hui largement admis qu'une unité phraséologique comme BRISER LA GLACE est une locution, tandis qu'une unité comme LANCE-FLAMMES est un mot composé, de nombreux cas, comme POMME DE TERRE restent moins consensuels, classés tantôt parmi les locutions, tantôt parmi les mots composés. Par ailleurs, l'appartenance des mots composés au domaine de la phraséologie n'est pas communément acceptée. Cet article a pour objectif de présenter une conception phraséologique du mot composé, situé sur un continuum des phrasèmes non-compositionnels allant des unités les plus morphologiques aux plus syntaxiques. La compositionnalité est appréhendée dans une vue encodante de la langue, dans le cadre du développement d'une ressource lexicographique basée sur la Lexicologie Explicative et Combinatoire.

1. INTRODUCTION

On s'accorde généralement pour dire qu'une unité phraséologique (phrasème) comme BRISER LA GLACE 'mettre fin à une situation gênante' est une locution, tandis qu'un phrasème comme LANCE-FLAMMES 'arme destinée à lancer du feu' est un mot composé. Il existe cependant de nombreux cas où le traitement varie selon l'approche adoptée: POMME DE TERRE est classé tantôt parmi les locutions, tantôt parmi les mots composés.

Dans le cadre d'un projet de recherche portant sur la description lexico-syntaxique des locutions au sein du Réseau Lexical du Français (RL-fr) (Lux-Pogodalla and Polguère, 2011), nous souhaitons développer un modèle qui nous permettra à terme de prévoir les variations des locutions françaises en discours, en nous appuyant sur les principes de la Lexicologie Explicative et Combinatoire (Clas, Mel'čuk et Polguère, 1995). Il nous est pour cela indispensable de distinguer les locutions dont l'origine syntagmatique est encore fortement présente et laisse des marques sur leurs emplois, de celles qui se comportent comme des unités simples. En partant de la classification des phrasèmes opérée par I. Mel'čuk (2013), nous proposons un continuum des phrasèmes non-compositionnels, qui va des unités les plus morphologiques aux plus syntaxiques, en passant par des cas intermédiaires comme QU'EN DIRA-T-ON ou CHEZ SOI.

Nous préciserons tout d'abord la conception de la non-compositionnalité que nous adoptons (2.1.), ainsi que ce que nous entendons par *phrasème* (2.2.). Nous en arriverons ensuite à la présentation de notre continuum, en présentant les phrasèmes morphologiques (3.1.), puis syntagmatiques (3.2.).

Les unités lexicales seront notées en petites capitales: MAISON; les sens apparaîtront entre guillemets simples: 'bâtiment destiné à loger une famille'; les unités instanciées seront notées en italique: *une maison*. Certains exemples sont issus de Frantext, du FrWac et de l'Est Républicain[134].

2. NON-COMPOSITIONNALITÉ ET PHRASÈMES

2.1. La non-compositionnalité

La non-compositionnalité sémantique a fait l'objet de nombreux débats, comme en témoigne notamment Svensson (2008). Deux grandes approches de la notion peuvent être envisagées (Polguère, à paraître): une approche décodante et une approche encodante. Faute de place, nous ne pouvons rappeler les deux approches que dans les grandes lignes.

L'approche décodante, ou interprétative, consiste à dire qu'un énoncé est compositionnel si son sens global peut être reconstitué à partir des sens de ses constituants. Dans le cas contraire, il sera – comme l'explique Martin (1997) – non-compositionnel:

> « Une locution prototypique est caractérisée […] par sa non-compositionnalité. On a beau comprendre tous les mots qui entrent dans *tirer le diable par la queue*, cela ne suffit pas pour comprendre ce que cette locution veut dire. » Martin (*ibid.,* p.293).

L'approche interprétative se situe donc du côté de la réception d'un énoncé, idée notamment soulignée par Numberg et coll. (1994, p.498):

> « […] compositionality – that is, the degree to which the phrasal meaning, **once known**, **can be analysed** in terms of the contribution of the idiom parts » (nous soulignons).

La non-compositionnalité est parfois associée avec la notion d'opacité. On trouve en effet chez Gross (1996) un renvoi à *opacité* à partir de la définition de *compositionnalité*. L'opacité y est ensuite définie comme le fait, pour une séquence de mots, d'avoir un sens non reconstituable à partir des sens des mots qui la composent.

Nous pouvons, à l'inverse, nous situer du côté de la production de l'énoncé, ayant ainsi une approche encodante comme celle suggérée par Jönsson (2008):

> « The meaning of a complex expression is determined by the meanings of its parts and **its mode of composition**. » (*ibid.*, p.21) (nous soulignons).

---

[134] Frantext: http://www.frantext.fr
FrWac: http://nl.ijs.si/noske/wacs.cgi/first_form
Est Républicain: https://arcas.atilf.fr/cqpweb/ (accessible par mot de passe uniquement).

Un énoncé compositionnel répond alors aux règles de formation de tout énoncé libre. Plus précisément, il est conforme à la règle de l'union linguistique selon laquelle une combinaison libre est obtenue par addition des signifiés, signifiants et syntactiques des signes linguistiques de façon appropriée (Mel'čuk, 2004). Le syntactique ou la combinatoire d'une unité indique la manière dont elle se combine à d'autres (Mel'čuk, 1993; Fradin, 2003). Ainsi, l'énoncé *Paul a pris Marie par la main* correspond à l'addition des sens 'Paul' 'prendre' 'Marie' 'par' 'la' et 'main'. Il répond également au syntactique des signes linguistiques correspondants: pour être instanciés, PRENDRE réclame un sujet et un complément nominaux et MAIN appelle un déterminant, etc. Si, en revanche, une combinaison de mots ne correspond pas à la somme des signifiés, ou des syntactiques des composants, alors elle est non-compositionnelle[135]. PRENDRE LE TAUREAU PAR LES CORNES 'attaquer un problème de front' n'est pas la somme de 'prendre' 'le' 'taureau' 'par' 'les' et 'cornes'. Il s'agit donc d'une combinaison de mots non-compositionnelle.

D'une approche à l'autre, une combinaison de mots sera ou non jugée comme compositionnelle. Par exemple, d'un point de vue décodant, FER A REPASSER (1) peut être, dans une certaine mesure, envisagé comme compositionnel, si l'on considère que son sens peut être calculé à partir de chaque unité constituante. Nous pouvons en effet retrouver le sens 'instrument comportant une semelle en **fer** qui sert à **repasser**'[136]. En revanche, d'un point de vue encodant, on ne peut pas le considérer comme compositionnel. Si le sens de chaque constituant est compris dans le sens global de FER A REPASSER, il ne s'agit pas pour autant à proprement parler d'un fer, mais bien d'un instrument[137]. Le monde de combinaison des unités ne répond ainsi pas aux règles de l'union linguistique, contrairement à *chemise à repasser* (2). Autrement dit: 'chemise à repasser' = 'chemise' ⊕ 'à' ⊕ 'repasser', mais 'fer à repasser' ≠ 'fer' ⊕ 'à' ⊕ 'repasser'.

(1) J'ai un **fer à repasser**.

(2) J'ai une **chemise à repasser**.

Bien que, d'un point de vue encodant, FER A REPASSER n'est pas compositionnel, le sens des unités est tout de même compris dans le sens global. La locution a alors un sens dit *transparent*. Par opposition, PRENDRE LE TAUREAU PAR LES CORNES, dont le sens global 'attaquer un problème de front' ne contient aucun des sens des unités constituantes, est opaque. La visée encodante nécessite alors de distinguer deux dichotomies – 1. compositionnalité/non-compositionnalité; 2. transparence/opacité. En somme, PRENDRE LE TAUREAU PAR LES CORNES est un syntagme non-compositionnel sémantiquement opaque, et FER A REPASSER est un syntagme non-compositionnel sémantiquement transparent.

L'objectif de notre travail est de décrire les variations des locutions, dans le soucis de donner à un locuteur ou à une machine les possibilités qui lui sont offertes pour produire un énoncé. Nous adoptons en conséquence une vue encodante de la non-compositionnalité.

---

135 Pour la non-compositionnalité due à la non addition des syntactiques, voir 2.1.1. infra.

136 FER est ici à considérer dans le sens de 'partie d'un instrument, d'un outil, qui est en fer ou en un métal ayant l'apparence du fer' (définition empruntée au TLFi).

137 Sur le lien entre sens global et sens des constituants nous renvoyons notamment à Polguère (à paraître).

2.2. Les phrasémes non-compositionnels

2.1.1. Des morphèmes phraséologiques ?

La conception de la phraséologie comme l'étude des combinaisons de mots (Cowie, 1994, p.3168), implique que la polylexicalité est le critère définitoire premier d'une unité phraséologique (cf. les critères présentés par Gross, 1996, pp.9-10). La nature des combinaisons reste cependant à établir: s'agit-il uniquement de combinaisons syntaxiques, ou bien ces combinaisons peuvent-elles être d'ordre morphologique ? La plupart des typologies phraséologiques considèrent uniquement des combinaisons de mots d'ordre syntagmatique, qu'il s'agisse de combinaisons compositionnelles ou bien non-compositionnelles[138]. La locution est donnée comme l'unité phraséologique par excellence:

> « Idioms form the majority and may be regarded as the prototype of the phraseological unit. » (Gläser, 1998, p.126).

En revanche, la conception des mots composés comme unités phraséologiques n'est pas systématique (Granger et Paquot, 2008, p.32). Par *mot composé* nous entendons, en accord avec F. Villoing (2003, p.185), tout « […] lexème construit à partir de lexèmes selon un mode d'organisation qui n'est pas syntaxique », comme LAVE-LINGE, BOIT-SANS-SOIF ou RENDEZ-VOUS. La principale raison qui écarte les mots composés du domaine de la phraséologie est leur forme non syntagmatique. Néanmoins, si on regarde nos exemples, les mots composés ont des propriétés des unités phraséologiques:

19 la polylexicalité: LAVE-LINGE = LAVER + LINGE

20 l'association contrainte: **nettoie-linge*

21 une base syntaxique potentielle: *lave-linge* = *machine qui lave le linge* (cf infra)

22 la non-compositionnalité.

La question de la non-compositionnalité se pose autrement dans ce cas. Elle est en fait liée non seulement à la somme irrégulière des sens, mais également à la somme irrégulière des syntactiques des unités constitutives. En effet, le mode de composition de LAVE-LINGE ne correspond pas à l'union linguistique de LAVER et LINGE, étant donné que la combinatoire des composants n'est pas respectée: un verbe ne peut pas avoir d'article comme dépendant syntaxique (**un lave la vitre*), et un nom, pour être instancié, nécessite un déterminant (**il lave vitre*).

Il convient d'ajouter que dans certaines langues, notamment le coréen et le chinois (Nguyen, 2006), le nombre d'unités multilexémiques morphologiques est très important, comparativement à celui des locutions. Nous pouvons ainsi avoir une définition élargie du phrasème en tant qu'énoncé multilexémique non libre (Mel'čuk, 2013), qui est sous-catégorisable en phrasèmes syntagmatiques et phrasèmes morphologiques (ou quasi-morphes, cf. Mel'čuk, 1997, pp.52-54).

---

138 Pour les unités phraséologiques compositionnelles on pourra notamment se référer aux formulæ de Cowie (1988), actes de langage stéréotypés de Kauffer (2011) et clichés linguistiques de Mel'čuk (2013). Pour un tour d'horizon sur les principales typologies, nous renvoyons notamment à Granger et Paquot (2008).

La question de l'origine des mots composés se pose encore. Certains sont construits sur des syntagmes identifiables, comme MEURT DE SOIF (3a), dont le patron V Prép NC donne lieu à des constructions libres en français (3b).

(3) a. Antoine est dans les bars du soir au matin, un vrai **meurt-de-soif** !
b. En général, quand Jules rentre de son footing, il **meurt de soif**.

D'autres rappellent des structures syntaxiques incomplètes. Un ouvre-boîte est un instrument qui permet d'ouvrir des boîtes de conserve. OUVRE-BOITE peut alors être considéré comme formé à partir d'un syntagme verbal incomplet; l'élément manquant étant le déterminant: *ouvre une boîte*. De même, nous pouvons gloser que EN-CAS 'repas léger' est construit sur le syntagme *en cas de besoin*.

Deux conceptions s'opposent quant à l'explication du mode de composition des mots composés de type V NC. Selon une première approche (Benveniste, 1974, pp.145-162), les mots composés seraient issus d'une construction syntaxique elliptique. Une seconde approche (Villoing, 2003) considère quant à elle que le verbe présent dans les mots composés n'est en aucun cas conjugué. Il s'agirait plutôt d'un thème verbal et les relations entre les unités constitutives du mot composé seraient d'ordre sémantique (relations prédicat–argument).

Nous n'entendons pas trancher cette question épineuse. Nous estimons cependant que, de par leur polylexicalité et leur non-compositionnalité, les mots composés relèvent de la phraséologie et se trouvent à une extrémité d'un continuum (cf. infra). Ils s'opposent aux locutions par les dépendances morphologiques, plutôt que syntaxiques, entre leurs constituants lexicaux.

2.1.2. La frontière entre mot composé et locution

Notre discussion sur le caractère phraséologique des mots composés permet de proposer un critère formel de distinction entre locution et mot composé: une locution est un syntagme, et un mot-composé est un morphème. Pour l'illustrer, nous nous intéressons à présent au cas de POMME DE TERRE, dont le statut n'est pas unanimement partagé par la communauté scientifique. Nous l'étudions conjointement à BRISER LA GLACE, UN BON COUP et UN PEU.

BRISER LA GLACE et POMME DE TERRE sont toutes deux formées d'une manière analogue à des combinaisons libres (Kahane, 2008, p.2539), respectivement comme *casser le vase* et *livre de français*. Ce sont donc des syntagmes non-compositionnels. Si l'on s'en tient à ces deux caractéristiques, nous en induirons que ces unités sont toutes deux des locutions. Néanmoins, elles n'admettent pas le même nombre de variations syntagmatiques. Si BRISER LA GLACE peut être passivée (4a) et subir l'insertion de modifieurs (4b), POMME DE TERRE n'admettra pas ou peu de variations, se comportant ainsi comme un nom simple.

(4) a. Dans le fond, Marie-Thérèse (Charlotte pour les intimes) est d'une grande sensibilité et timidité qu'elle compense en se montrant un peu fière avec les gens qu'elle ne connaît pas. Mais une fois que **la glace est brisée**, elle se révèle réellement adorable.

(FrWac: http://pilleul.sylvain.club.fr/les_enfants_de_france.html) (nous soulignons)

b. Cassie tourna la tête, et vit Peyton… une jeune fille qu'elle connaissait de loin. Elle lui fit tout de même un sourire, histoire de **briser** un peu **la glace**.
(FrWac: http://do-your-life.jeun.fr/plage-f47/seule-au-crepuscule-liiibre--t303.htm) (nous soulignons)

Cette différence de comportement entre les deux unités explique l'absence de consensus au sujet de leur statut. Une approche privilégiant le comportement de l'unité dans la phrase considérera POMME DE TERRE comme un mot composé (Gross, 1996, pp.25-59; Corbin, 1997), au même titre que LAVE-LINGE.

> « Les expressions *cul de sac*, *pomme de terre*, se présentent avec la structure productive *N de N* de *chef de groupe* ou de *bouteille de lait*. En fait, l'usage revient à considérer des noms composés, qui seraient mieux orthographiés avec des traits d'union: *cul-de-sac*, *pomme-de-terre*. Ces noms composés deviennent alors des entrées ordinaires de dictionnaire. Elles apparaissent dans des distributions nominales où elles ne sont pas distinguables des noms simples comme *impasse* ou *patate*. » (Gross, 1982, pp.152-153).

Une approche contraire considérera l'unité comme une locution (Kahane, 2008; Mel'cuk, 2013; Polguère, 2008, pp.51-58).

L'approche que nous nommerons *comportementale* peut être appuyée par la présence d'unités ayant la forme de syntagmes, mais se comportant comme des unités appartenant à une partie du discours différente de celle de la tête de syntagme. Par exemple, UN BON COUP (5) est formellement un syntagme nominal, mais se comporte comme un adverbe: *soigne-toi un bon coup*.

bien

(5) Écoute, Bes: soigne-toi **un bon coup** et reviens guéri. Ta façon de t'esquiver en pleine réunion est très déplaisante.
(Frantext: SCHREIBER Boris, *Un silence d'environ une demi-heure*, 1996, p.977)(nous soulignons)

Nous pourrions être tenté de considérer ces unités comme des mots composés. Néanmoins, certaines admettent des modifications qui réactivent leur origine syntagmatique, comme UN PEU (6).

(6) a. **Un peu** déçu, je pénétrai sur le terrain de football et jusqu'à six heures nous tapâmes dans un ballon.
(Frantext: JOFFO Joseph, *Un sac de billes*, 1973, p.228)

b. Il n'a jamais joué au football mais se passionne pour l'arbitrage. « Tout s'est fait **un petit peu** par hasard » explique-t-il « je regardais un jour une rencontre télévisée avec mon père, et j'ai eu envie d'en savoir plus sur la fonction d'arbitre ».
(*L'Est Républicain* 15 juin 1999)

Le cas de UN PEU indique que considérer une frontière entre locution et mot composé basée sur un critère formel opposant un syntagme à un morphème, nécessite de considérer des catégories intermédiaires, basées sur le comportement de l'unité dans la phrase.

3. PRÉSENTATION DU CONTINUUM

3.1. Phrasèmes morphologiques non-compositionnels: noms composés

Comme nous venons de le voir, les phrasèmes morphologiques non-compositionnels se caractérisent par un mode d'organisation qui n'est pas syntaxique. De plus, ils se comportent comme une unité simple et peuvent subir, à ce titre, seulement des modifications d'ordre morphologique:

> « Si la lexie suspecte a un arbre SyntS[139] "bizarre" qui ne peut pas être traité par les règles syntaxiques standard du modèle linguistique, L est décrite comme une lexie synchroniquement **mono**lexémique, c'est-à-dire […] comme un seul mot-forme, peu importe son orthographe (avec ou sans le trait d'union) ou son étymologie (historiquement, L peut être dérivée ou composée ou encore être un syntagme lexicalisé) […]. » (Mel'čuk, 2006, p.55).

Formellement, certains phrasèmes ont des structures qui rappellent des structures syntaxiques, complètes ou incomplètes (cf. 2.1.1.). Ainsi, nous établissons deux sous-catégories de phrasèmes morphologiques, suivant le type de dépendances qui lient les unités constituantes.

- Dépendances syntaxiques morphologisées. Ces dépendances s'appliquent aux phrasèmes formés sur un syntagme, dont les dépendances syntaxiques ne sont plus actives, comme BOIT-SANS-SOIF$_{NC}$ ou SANS-DOMICILE-FIXE$_{NC}$.

- Dépendances morphologiques potentiellement d'origine syntaxique. Ces dépendances s'appliquent aux phrasèmes tels que OUVRE-BOITE ou TOURNEVIS (cf 2.1.1).

À noter aussi, concernant les phrasèmes construits sur des syntagmes nominaux, que la frontière entre morphologie et syntaxe est parfois difficile à établir. Certaines unités sont clairement des mots composés. Par exemple, *un rouge-gorge* ne peut être, qu'il y ait tiret ou non, considéré comme un syntagme, étant donné qu'il y a incompatibilité entre le genre du nom et celui du déterminant. De même, il y a des cas d'incompatibilité entre le nombre du nom et celui du déterminant dans *un micro-ondes*. Nous pouvons aussi observer des cas d'ordre non-canonique Adj-NC, comme *une plate-bande*, des cas d'absence d'accord Adj-NC, comme *une grand-mère*, ou bien encore des variations de prononciation, comme *un croc-en-jambe* [krɔk]. Néanmoins, pour certaines unités, le seul critère disponible est l'intonation; il s'agit alors de déterminer si l'unité se prononce comme un seul mot forme, ou a la structure intonative d'un syntagme140.

---

139 SyntS = syntaxe de surface (cf. Mel'čuk, 2009, p.6).

140 Sur quelques comparaisons entre noms composés et locutions voir Mel'cuk (2006).

Concernant les dépendances morphologiques potentiellement d'origine syntaxique, il faut également considérer des unités telles que TROTTE-BEBE ou PENSE-BETE. Si l'ordre des éléments n'apparaît pas comme canonique (*le bébé trotte*, *le bête pense*), nous pouvons considérer ici que le nom a en fait un rôle de patient, et non d'agent:

> « Ces données se rapprochent […] des VN prototypiques dans la mesure où le procès, dans cette construction, a une valeur sémantique de causatif et, qu'en conséquence, le N répond aux propriétés d'un patient (« fait couler le sang, fait penser les bêtes, fait pisser le chien, fait sauter le bouchon, fait trotter le bébé »). En outre, deux participants sémantiques sont bien impliqués, le patient et l'instrument (pense-bête, saute-bouchon) ou le patient et l'agent (trotte-bébé). » (Villoing, 2009, pp.181-182).

Notons, par ailleurs, que si la majeur partie des phrasèmes morphologiques sont caractérisés par une stabilité interne, certains mots composés d'origine syntagmatique admettent des variations liées à leur mode de formation. Par exemple, CHEZ-SOI (7), voit l'un de ses constituants changer suivant le possesseur (Cheminée, 1999).

(7) a. Enfin on reparla de rentrer. […] Dommage que ce soit fini on commençait vraiment à s'y mettre, hélas les meilleures choses n'ont qu'un temps. D'ailleurs dans le fond on aime bien retrouver son petit **chez-soi**. On est content de partir mais on est content aussi de revenir.
(Frantext: ROCHEFORT Christiane, *Les Petits enfants du siècle*, 1961, p.99) (nous soulignons)

b. Et toi, t'as pas honte d'être dehors à une heure pareille ? T'as bien un **chez-toi** ?
(Frantext: SABATIER Robert, *Trois sucettes à la menthe*, 1972, p.295) (nous soulignons)

Faut-il considérer qu'il s'agit d'une seule ou de plusieurs unités lexicales correspondant à chacune des variantes ?

3.2. Phrasémes syntagmatiques non-compositionnels: locutions

De même que les mots composés, les locutions peuvent être classées suivant la nature des dépendances qui lient leurs unités constitutives.

1. Dépendances syntaxiques actives, dans le cas de locutions comme BRISER LA GLACE qui se comportent dans la phrase de la même manière que leurs têtes de syntagmes et qui sont sujettes à des variations syntagmatiques.

2. Dépendances syntaxiques inactives, dans le cas des locutions qui se comportent comme des unités simples, ne subissant ni variations paradigmatiques, ni variations syntagmatiques. Leur combinatoire peut être identique (POMME DE TERRE) ou différente de celle de la tête de leur syntagme (UN BON COUP).

3. Dépendances syntaxiques anciennes, dans le cas des locutions dont les dépendances syntaxiques entre les constituants appartiennent à un état antérieur de la langue (SANS COUP FERIR, TAMBOUR BATTANT).

En prenant en compte toute la diversité des cas que nous venons de discuter, nous sommes à présent en mesure de proposer un continuum sur lequel ils se répartissent entre deux pôles:

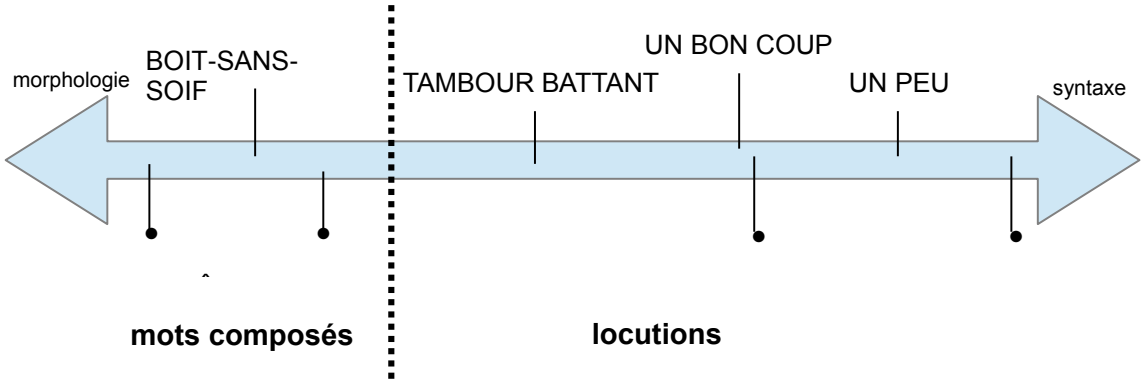


Figure 1. continuum des phrasèmes non-compositionnels.

Il reste à discuter de la place qu'occupe la construction appositive NC NC dans le continuum. S. Kahane (2008) dénombre deux sous-types de cette construction:

1. Modificative.

2. Coordinative.

La construction modificative est donnée comme la réduction d'une construction NC Prép NC (Fradin, 2009): *une borne incendie* (= une borne pour incendie), *des pommes vapeur* (= des pommes à la vapeur). La construction coordinative correspondrait quant à elle à l'association symétrique de deux noms: CANAPE-LIT, HOTEL-RESTAURANT, MOISSONNEUSE-BATTEUSE, CHIEN-LOUP. La symétrie n'est en réalité qu'apparente: un canapé-lit est un canapé qui fait également office de lit, mais en aucun cas on ne pourra le qualifier de **lit-canapé*. Pour définir le statut de tels phrasèmes, il faut déterminer si la construction NC NC relève de la morphologie, ou de la syntaxe. Autrement dit, la combinaison NC NC permet-elle à tout locuteur de construire des syntagmes ? Un premier élément de réponse est que tous les noms ne peuvent pas s'employer en apposition. Seul un petit nombre trouve cet emploi – citons en exemple *une tâche clé*, *un dessert maison* – qui bien souvent fait basculer le nom dans la catégorie des adjectifs: *un cas très limite*, *une taille tout à fait standard*. De plus, l'association de deux noms en apposition est bien souvent contrainte. Si on peut dire *confiture maison*, on ne pourra pas dire **confiture usine* (Kahane, 2008, p.2541).

En dernier lieu, certains phrasèmes comme A QUI MIEUX MIEUX, CUL CUL LA PRALINE, A BRAS-LE-CORPS ont des structures analogues à des structures syntagmatiques non productives. Nous ne pouvons, faute de place, discuter du statut de ces unités qui semblent se situer du côté de la morphologie.

4. CONCLUSION

Nous avons argumenté en faveur d'une conception des noms composés comme unités phraséologiques, en raison de leur polylexicalité, de leur base syntaxique potentielle, et de

leur non-compositionnalité; la non-compositionnalité étant envisagée suivant une approche encodante de la langue.

Nous avons de plus défendu l'idée selon laquelle les phrasèmes non-compositionnels peuvent être décrits suivant un continuum allant des unités les plus morphologiques aux plus syntaxiques. Les mots composés, unités morphologiques, peuvent être rapprochés de la syntaxe suivant leur forme plus ou moins syntagmatique: dépendances morphologiques potentiellement d'origine syntaxique (OUVRE-BOITE), dépendances syntaxiques morphologisées (BOIT-SANS-SOIF). Les locutions, unités syntaxiques, peuvent être quant à elles rapprochées de la morphologie suivant leur forme plus ou moins fixe: dépendances syntaxiques inactives (pomme de terre), dépendances syntaxiques actives (BRISER LA GLACE). Cette étude préliminaire a pour but de proposer une échelle de classement des locutions, suivant qu'elles admettent ou non des variations formelles: UN PEU s'emploie comme un adverbe mais doit être considéré comme une locution nominale afin de décrire la variation *un petit peu*.

Il reste néanmoins des cas ambigus comme la construction appositive, qui dans certains cas peut être envisagée comme une construction syntaxique libre, alors que dans d'autres, elle sera considérée comme une construction non productive, ne pouvant pas relever de la syntaxe.

Dans le prolongement de cette étude consacrée principalement aux variations syntagmatiques des locutions, il conviendra de vérifier si les variations paradigmatiques autorisées vont de pair avec les variations syntagmatiques.

Bibliographie

BENVENISTE, É., 1974. *Problèmes de linguistique générale II*. Paris: Gallimard.

CHEMINEE, P., 1999. Statut lexical, statut lexicographique: le nom composé dans le dictionnaire. *Linx*, 40, [online] Available at: http://linx.revues.org/784 (accessed 20 december 2013).

CLAS, A., MEL'CUK, I. AND POLGUERE, A., 1995. *Introduction à la lexicologie explicative et combinatoire*. Bruxelles: Duculot.

CORBIN, D. 1997. Locutions, composés, unité polylexématiques: lexicalisation et mode de construction. In: M. Martins-Baltar, ed. *La locution entre langue et usages*. Fontenay Saint-Cloud: ENS éditions. p.55-102.

COWIE, A. P., 1988. Stable and creative aspects of vocabulary use. In Carter, R. and M. J. McCarthy, ed. *Vocabulary and Language Teaching*. London: Longman. pp.126-137.

COWIE, A. P., 1994. Phraseology. In: R. E. Asher, ed. *The Encyclopedia of Language and Linguistics*, Oxford: Oxford University Press. pp.3168-3171.

FRADIN, B., 2003. *Nouvelles approches en morphologie*. Paris: Presses Universitaires de France.

FRADIN, B., 2009. Compounding in French. In: R. Lieber and P. Atekauer, ed. *Oxford Handbook on Compounding*. Oxford: Oxford University Press. pp.417-435.

GLÄSER, R., 1998. The stylistic potential of phraseological units in the light of genre analysis. In: A. P. Cowie, ed. *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press. pp.125-143.

GRANGER, S. AND PAQUOT, M., 2008. Desentangling the phraseological web. In: S. Granger and M. Paquot, ed. *Phraseology. An interdiscplinary perspective*. Amsterdam/Philadelphia: John Benjamins. pp.27-49.

GROSS, G., 1996. *Les expressions figées en français noms composés et autres locutions*. Paris: Ophrys.

GROSS, M., 1982. Une classification des phrases « figées » du français. *Revue québécoise de linguistique*, 11(2), p.151.

JÖNSSON, M. L., 2008. *On Compositionality. Doubts about the Structural Path to Meaning*. Thèse de doctorat. Université de Lund.

KAHANE, S., 2008. Les unités de la syntaxe et de la sémantique: le cas du français. *Actes du Congrès Mondial de Linguistique Française 2008,* p.2531.

KAUFFER, M., 2011. Actes de langage stéréotypés en allemand et en français. Pour une redéfinition du stéréotype grâce à la phraséologie. *Nouveaux Cahiers d'allemand*, 1, p.35.

LUX-POGODALLA, V. AND POLGUERE, A., 2011. Construction of a French Lexical Network: Methodological issues. *Proceedings of the First International Workshop on Lexical Resources. WoLeR 2011, An ESSLLI 2011 Workshop*, p.54.

MARTIN, R., 1997. Sur les facteurs du figement lexical. In: M. Martins-Baltar, ed. *La locution entre langue et usages*. Fontenay Saint-Cloud: ENS editions. pp.291-305.

MEL'CUK, I., 2009. Dependancy in natural language. In: A. Polguère and I. Mel'čuk, ed. *Dependency in Linguistic Description*. Amsterdam/Philadelphia: John Benjamins Publishing Company. pp.1-101.

MEL'CUK, I., 1993. *Cours de morphologie générale, vol. 1: Le mot.* Montréal: Les Presses de l'Université de Montréal — Paris: CNRS éditions.

MEL'CUK, I., 1997. *Cours de morphologie générale, vol. 4: Signes morphologiques*. Montréal: Les Presses de l'Université de Montréal — Paris: CNRS éditions.

MEL'CUK, I., 2004. La non-compositionnalité en morphologie linguistique. *Verbum*, 26(4), p.439.

MEL'CUK, I., 2006. Parties du discours et locutions. *Bulletin de la Société de linguistique de Paris*, 101(1), p.29.

MEL'CUK, I., 2013. Tout ce que nous voulions savoir sur les phrasèmes, mais... *Cahiers de lexicologie*, 102, p.129.

NGUYEN, É. V. T., 2006. *Unité lexicale et morphologie en chinois mandarin. Vers l'élaboration d'un Dictionnaire Explicatif et Combinatoire du chinois*. Thèse de doctorat. Université de Montréal.

NUNBERG, G. D., SAG, I. AND WASOW, T., 1994. Idioms. *Language*, 70(3), p.491.

POLGUERE, A., 2008. *Lexicologie et sémantique lexicale. Notions fondamentales.* Montréal: Les Presses de l'Université de Montréal.

POLGUERE, A., à paraître. Non-compositionnalité: ce sont toujours les locutions faibles qui trinquent. *Verbum*.

SVENSSON, M. H., 2008. A very complex criterion of fixedness: Noncompositionality. In: S. Granger and M. Paquot, ed. *Phraseology. An interdiscplinary perspective*. Amsterdam/Philadelphia: John Benjamins. pp.81-93

VILLOING, F., 2009. Les mots composés VN. In: B. Fradin, F. Kerleroux and M. Plénat, ed. *Aperçus de morphologie du français*. Saint-Denis: Presses Universitaires de Vincennes. pp.175-198.

VILLOING, F., 2003. Les mots composés VN du français: arguments en faveur d'une construction morphologique. *Cahiers de Grammaire*, 28, p.183.

PRAGMATIC INFORMATION AND UNPREDICTABILITY IN LEARNER'S DICTIONARIES

Irena Srdanović
Faculty of Arts, University of Ljubljana
irena.srdanovic@gmail.com

Abstract

This paper addresses the phenomenon of unpredictability of expressions in a second language from the viewpoint of second language learners. The problem of unpredictability of various expressions is closely related to differences between the learner's mother tongue and the foreign language in the process of acquiring by the learner (Nation, 2001). The differences and with it closely related (un)predictability can appear on different linguistic levels, such as morpho-syntactically different but semantically matched, syntactically matched but with different combination of words, or equivalent combinations of words in two target languages but partially or completely different meaning of words in the languages. The results indicate that dictionaries targeted at language learners need to emphasize information on word combinations that are difficult to predict by language learners, and suggest the representation of such information as one type of pragmatic information.

1. INTRODUCTION

Linguistic and lexicographic tradition has long been neglecting the pragmatic aspects of language. In lexicography, dictionaries designed for foreigners pay attention on pragmatic information, especially the dictionaries that used language corpora as material, starting with the Cobuild dictionary (Sinclair, 1987). In addition to the selection and method of presentation of this type of information, the dictionaries also differ in defining the pragmatic information. They are talking about pragmatic, encyclopedic and cultural information on the level of description of nonlinguistic elements. On more narrow level of language use, some dictionaries consider examples as pragmatic information, while others descriptions of the use of certain lexical items, or classification into categories such as free combinations, collocations and idioms (cf. Sharpe, 1989, Nuccorini, 1993, Inoue, 1998).

Nation (2001) believes that a collocation is often grammatically and lexically unpredictable, so students cannot easily produce native-like expressions. The problem of unpredictability of collocations is closely related to differences between the learner's mother tongue and the foreign language being learnt by the learner. For example, the adjective 'cold', which is typically used as one of the two different words in Japanese, *tsumetai* or *samui* 'cold' can be as well regarded as unpredictable, since the first one refers to feeling of coldness on touch, while the second one refers to the overall feeling of coldness,

which is not linguistically differentiated in English or some other languages. For that reason, for example, the collocation 'cold water' could easily be mistaken with *samui mizu* instead of the correct usage *tsumetai mizu* 'cold water' by learners. Sometimes this kind of unpredictability is actually stimulated by socio-cultural, cognitive or widely speaking, pragmatical aspects of differences. The research takes Japanese adjectives and nouns as an example and indicates the types of unpredictable collocations comparing them in Japanese and English based on the corpus and dictionary analysis.

2. PRAGMATIC INFORMATION IN DICTIONARIES

2.1. Pragmatic information: scope and representation

Pragmatics can be defined as "the study of how language is used to express meaning in context" (Archer et al., 2012:11). In the context of lexicography, pragmatic information as a term refers broadly to information on word usage that is incorporated or considered for inclusion into dictionaries. Especially bilingual dictionaries for learners of a foreign language have been advised to take into consideration pragmatic information seriously and systematically, due to their important role in intercultural communication (Gharaei, 2011:84).

Apresyan (1988, in Burkhanov, 2003) states that pragmatic specifications in lexicography are those dealing with "the speaker's attitude to reality, the message and/or the interlocutor, which is encoded in linguistic signs as units of language system". According to Landau (1989), pragmatic information in dictionaries refers to uses of spoken or written language, standard and non-standard uses of language, as well as geographic, social, or temporal limitations on use. Landau (ibid.) further classifies this information into the following categories: currency/temporality, frequency of use, geographic variations, specialized terminology (field labels), restricted/taboo usage, insult, slang, style or register/functional variety, status label (standard/nonstandard/illiterate).

Šorli (2011) uses as a model COBUILD dictionary project and talks about pragmatic information as an integral part of an (extended) unit of meaning, which is identifiable only by examining its repeated occurrences in corpus data. She presents pragmatic information in meaning descriptions of the Slovene Lexical Database (SLD focusing on pragmatic components that can be abstracted from contextual features. However, the databases addresses also other types of pragmatic information, such as word connotation and emotive and attitudinal meaning that can be associated with words per se.

There have been also other approaches and inclusion of additional distinguished information, but with no consensus so far on content as well as on representation of pragmatic information in dictionaries, due to complexity and variety of information.

As for the representation of such kind of information, Zgusta (1988) records three different ways: a) cultural settings, b) equivalence in bilingual dictionaries, c) definitions in monolingual dictionaries. Zgusta (ibid.) especially points out the quality of descriptive definitions in the COUBILD dictionary, which use personal pronoun *you* as an explicit pragmatics means and involvement of readers in a described situation.

Furthermore, Nuccorini (1993:215-223) distinguishes between explicit and implicit aspects in recording pragmatic information in dictionaries. Alongside phonological, syntactic, semantic elements, explicit pragmatic information appears. For example, in the case of LDOCE, they appear in Usage Notes (sections appended to the relevant entry), Language Notes (independent sections located at their alphabetical place within the

dictionary), glosses (comments for definitions and examples) and labels (grouped into areas »showing situation in which a word is used«, »showing attitude« and »showing other limitations on use«). As for implicit usage of pragmatic information, it appears in definitions and examples.

2.2. Pragmatic information in Japanese dictionaries

One of early examples of pragmatic considerations in Japanese language dictionaries can be seen in *Kenkyusha's New School Dictionary* (Japanese title: *Kenkyusha Shin Eiwa Chujiten*, edited by Yoshisaburo Okakura), which devoted around 200 pages to appendices about various information including idioms, daily conservation, letter-writing, full-page pictures with vocabulary explanation. In the same period, the marking of word frequency ranks based on Thorndike's list of basic words appeared as a new feature in *A Standard English-Japanese Dictionary* (1929), edited by Tsuneta Takehara. Also, the first monolingual English learner's dictionary *Idiomatic and Syntactic English Dictionary* (appeared in Japan in 1942 and later on in Oxford 1948, edited by Hornby) was innovative with the marking of countable and uncountable nouns, and major syntactic patterns of verbs.

The second half of the 20th century was marked with development of new technologies and approaches in lexicography, when publishers in Japan tried to implement various innovations in usage of empirical data and implementing them in the form of notes. One successful example from the period was *Taishukan Genius English-Japanese Dictionary* (Japanese title: *Jiten Taishukan Genius Eiwa Jiten*, 1988) with strong emphasis on the rich information about the usage. In 2001, its larger version was published under the name *The Unabridged Genius English-Japanese Dictionary*.

In the same period English lexicography outside of Japan was flourishing in corpus-based and corpus-driven approaches to lexicography with a wide scope of usage notes, restricted defining vocabulary, differences in frequencies in spoken and written language, sorting senses based on their frequency, register descriptions, examples carefully selected from corpora, collocation data, frequent syntactic pattern and so on. Some of the new corpora features were considered or implemented for the first corpus-based dictionary, COBUILD English dictionary (1987). In 1995, the corpus-based "Big Four", which refers to OALD, LDOCE, CIDE and COBUILD were published or revised. For detailed consideration of pragmatic information in some of the dictionaries, refer to Nuccorini (1993). In Japan, *The Wisdom English-Japanese Dictionary* claimed to be the first corpus-based medium-sized learner's dictionary (Tono, 2006:20).

As for Japanese dictionaries targeted at language learners of Japanese, a few dictionaries appeared but there are still no sufficient lexicographic resources. Some of the resources are specialized, and they take care of pragmatic information to some extent. For example, *Kodansha's Effective Japanese Usage Dictionary* explains different usages of synonymous terms, or *Dictionary of Sentence Patterns* (Japanese title: *Nihongo bunkei jiten*), incorporates pragmatic markers related to functional role of expressions.

Recently, different dictionary interfaces (web, mobile, CD, hand-held electronic dictionaries) have appeared as an outcome of rapid development of internet and technologies. This provides new opportunities for coverage and representation of various types of pragmatic information.

2.3. The phenomenon of unpredictability as pragmatic information for language learners

There is a consensus of lexicographers and researchers on the fact that there are different needs and methods in creation of learners' dictionaries vs. monolingual non-learner dictionaries. As described by Tono (2006:19), the main focus of general-purpose monolingual dictionaries is on meaning and spelling, while other types of information are less needed because "native speakers know how to construct a phrase or a sentence intuitively". On the other side, encoding information becomes very important in learner's dictionaries and collocations are especially helpful for that.

As stated above in the introduction, the phenomenon of unpredictability is closely related to pre-experience of language learners with mother tongue and other foreign languages. It is highly possible that this language experience of word combinations and their phonetic, morphosyntactic, lexico-semantic, pragmatic and other properties influences the production and understanding in a learning language, as well as possible grammatical, semantical or pragmatical failures. More these properties are different and more these properties are on case by case basis, the learning burden is bigger and native-like understanding and production is more difficult. As a result, mentioned differences in these kind of properties evoke more mistakes by language learners.

(Un)predictability of word combinations is not an integrated part of a word or its usage and as such is almost of no relevance from the point of view of an adult native speakers' reception and production of a language. However, the phenomenon of (un)predictability becomes very relevant as an attribute of a word usage from the viewpoint of the language user who is at the same time a learner of the language. The notion is related to word usage of at least two languages, mother tongue of a learner and the learning language, although other languages spoken by learner can intervene as well. As such, this research considers it as pragmatic information in a broad sense of a word relevant for learners' dictionaries or learner's user guides. This is also in line with the statement by Burkhanov (2003:109) that "usage is broadly understood and includes any linguistic phenomenon that, from the lexicographer's viewpoint, may present difficulties for the intended user."

The attribute of (un)predictability doesn't have an absolute value, it is relative as dependent on various factors, mainly related to mother tongue of a learner or other languages spoken by the learner. So, syntactic and morpho-syntactic structures in the language(s) and the usage of specific word combinations within the structures, semantics and pragmatics of the word combinations in question will influence how (un)predictable the word combinations are in a target language.

3. CASES OF UNPREDICTABILITY IN ADJECTIVE NOUN COMBINATIONS

Srdanović (2014) describes how the »Collocation data of adjectives and nouns« resource was created for most frequent 500 adjectives in two large-scale Japanese corpora, BCCWJ and JpTenTen. The most frequent noun collocates were extracted for all the adjectives using the Sketch Engine tool. The raw resource is then used for developing a model for dictionary of collocations targeted at Japanese language learners. The model presents the following steps: a) Defining main entries for the dictionary, b) Comparing collocations extracted from the two large-scale corpora, c) Arranging the list of collocates, d) Grouping collocates based on their difficulty level, e) Providing translations to a target language and furigana, f) Detecting collocations that are difficult to predict or unpredictable for learners,

g) Discovering complex patterns, typical usages, genre specifics etc, h) Providing lexical maps of noun collocates.

In the analysis conducted, especially steps d), f), g) and h) refer to usage of lexical items and their collocations and aim to provide richer information to learners. In this section, the focus is on step f) and presenting cases of unpredictability in adjective-noun combinations. 200 different adjective noun combinations for highly frequent adjectives *kawaii* 'cute', *ookii* 'big', *tsuyoi* 'strong', *nagai* 'long', *warui* 'bad', *hikui* 'low', *hayai* 'quick', *omoshiroi* 'interesting' (25 noun collocates for each adjective) were observed for unpredictability from the viewpoint of English native speakers as learners of Japanese in encoding situation. Besides that, a thorough analysis of adjective *takai* 'high, tall, expensive',[141] which served as a model for the dictionary, were conducted.

The analysis revealed a few different types of unpredictable or difficult to predict collocates for learners who are encoding language.

The first group are collocations that are morpho-syntactically different from the combinations in the native language. For example, *sei ga hikui hito/kata* 'a short person' is unpredictable since the additional elements *sei ga* 'the back is' must be added to make the collocation complete in Japanese (lit. *a person with a low back*). This collocation cannot be simply used as *hikui hito* with an intention to denote 'a short person', since without the necessary element *sei ga*, the combination of words is incomplete and has literary meaning 'lit. 'a person with low/high …<something>'. The same is valid for the *sei ga takai hito/kata* 'a tall person'. Similarly, *kawaii kanji* 'cute/a kind of cute', lit. *cute feeling* used to express positive attitude of appreciation/attraction is more difficult to produce by foreign learners than the sole expression *kawaii*, at least until the learners get used to it through frequent exposure.

The second group are collocations that do not have the same combination of words (lexical constituents) in the native language and have partially or fully different meaning. For example, *nagai me* 'in the long run', lit. *long eyes* is difficult to predict for encoding purpose, but also for decoding one. Another example is *takai okane* 'a lots of money', lit. 'high money '.

The third group are collocations with equivalent combination of words in mother tongue, but partially or fully different meaning. Although the collocation *takai kabe* has the same constituent 'high' (*takai*) and 'wall' (*kabe*) as in English, the same expression has a second more abstract meaning 'high barrier' with different and unpredictable constituents. These unpredictable usages, different from expressions in learners' mother tongues, need to be explicitly explained.

In the case of unpredictable collocations, the corpus-based analysis are further performed in search for complex patterns and typical usages. In addition, the analysis reveals usages specific to a particular sub-corpus/genre. Finally, the model dictionary provides images of lexical maps of adjective+noun collocates, which depict how collocates can be grouped together as concepts observed from a cognitive linguistics perspective.

[141] For more information on the analysis of *takai*, refer to (Srdanović, 2013, 2014, Srdanović and Sakoda, 2013).

4. CONCLUSION AND FURTHER WORK

As shown in the examples, it is harder to predict the following types of collocations:

1) collocation that is morpho-syntactically different., 2) collocation that in the native language do not have the same combination of words and have partially or fully different meaning, and 3) collocation with equivalent combination of words in mother tongue, but partially or fully different meaning. The results indicate that dictionaries targeted at language learners need to emphasize information on combinations that are difficult to predict by language learners, which can be regarded as one type of pragmatic information.

Tono (2012:21-22) discusses about misconceptions of user-friendliness and selection of an appropriate dictionary for learners. Often all-inclusive dictionaries are preferred, which appear to be too complicated and not appropriate for the level of a learner. He states that we "need to sort our necessary from unnecessary information, and tune the amount of information to the level of prospective users" (ibid.:22). In relation to that, this research suggests to consider putting emphasis on unpredictable information over the predictable one in the dictionaries targeted at language learners. To achieve this, further analysis of various types and levels of (un)predictability for a variety of word combinations in correlation with different languages is needed.

References

ATKINS, B. T. S., RUNDELL, M., 2008. *The Oxford Guide to Practical Lexicography*. Oxford: OUP

BURKHANOV, I., 2003. Pragmatic Specifications: Usage Indications, Labels, Examples; Dictionaries of Style, Dictionaries of Collocations. In Sterkenburg P. (ed.), *A Practical Guide to Lexicography*. Amsterdam: John Benjamin Publishing Company. pp. 102-113.

GALLY, T., 2012. Kokugo Dictionaries as Tools for Learners: Problems and Potential. *Acta Linguistica Asiatica* 2/2. pp. 9-19.

GHARAEI, Z., 2011. The Pragmatics of English-Persian Dictionaries: Problems and Solutions. *International Journal of English Linguistics* 1/2, pp. 83-90.

IKEGAMI, Y., 2006. English Dictionaries in Japan: Past, Present and Future. *English Dictionaries in Japan*. Taishukan

INOUE, N., 1998. Gakushuu eiwa jiten ni okeru gohou jouhou to korokeeshon jouhou – koopasu de nani ga dekiru ka. *Eigo kyouiku to eigo kenkyuu*, 15, pp. 71-86.

LANDAU, S.I., 1989. *Dictionaries: the Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

NAGASHIMA, D., 1966. The Influence of Noah Webster's Dictionary on Dr. Ōtsu-ki's Genkai. *Kokugogaku*, 64. pp. 71-80.

NATION, P., 2001. *Learning vocabulary in another language*. Cambridge University Press.

NUCCORINI, S., 1993. Pragmatics in learners' dictionaries. *Journal of Pragmatics*, 19(3), pp. 215-237.

SINCLAIR J.M. ed., 1987. *Looking up: An Account of the COBUILD Project in Lexical Computing*. Collins, London.

SHARPE, P. A. 1989. Pragmatic Considerations for an English-Japanese Dictionary. *International Journal of Lexicography*, 2, pp. 315-323.

SRDANOVIĆ, Irena, 2014. Corpus-based collocation research targeted at Japanese language learners. *Acta linguistica asiatica*, 4(2), pp. 25-35. (available at: http://revije.ff.uni-lj.si/ala/article/view/2866/2598)

SRDANOVIĆ, Irena, SAKODA, Kumiko, 2013. Analysis of learner's production of adjectives using the Japanese language learner's corpus C-JAS: the case of takai. *Acta linguistica asiatica*, 3(2), pp. 9-24. (available at: http://revije.ff.uni-lj.si/ala/article/view/968)

SRDANOVIĆ, Irena. Daikibo koopasu wo mochiita keiyoushi to meishi no korokeishon no kijutsuteki kenkyuu: nihongo kyouikuyou no jisho sakusei ni mukete. *Kokuritsu kokugo kenkyujo ronshu (NINJAL Research Papers)*, 6, pp. 135-161.

ŠORLI, M. 2011. Pragmatic Components in the Slovene Lexical Database Meaning Descriptions. *Electronic lexicography in the 21st century: new applications for new users: proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 251-259.

TONO, Y., 2000. On the effects of different types of electronic dictionary interfaces on L2 learners' reference behaviour in productive / receptive tasks. In: U. Heid, S. Evert, E. Lehmann and C. Rohrer (eds.), *Proceedings of the Ninth Euralex International Congress, EURALEX 2000,* pp. 855-861. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

TONO, Y., 2009. Pocket electronic dictionaries in Japan: User perspectives. In Henning Bergenholtz, Sandro Nielsen and Sven Tarp (eds.), *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, pp. 33-67. Bern: Peter Lang.

MULTILINGWIS – A MULTILINGUAL SEARCH TOOL FOR MULTI-WORD UNITS IN MULTIPARALLEL CORPORA

Simon Clematide
Institute of Computational Linguistics
University of Zurich
siclemat@cl.uzh.ch

Johannes Graën
Institute of Computational Linguistics
University of Zurich
graen@cl.uzh.ch

Martin Volk
Institute of Computational Linguistics
University of Zurich
volk@cl.uzh.ch

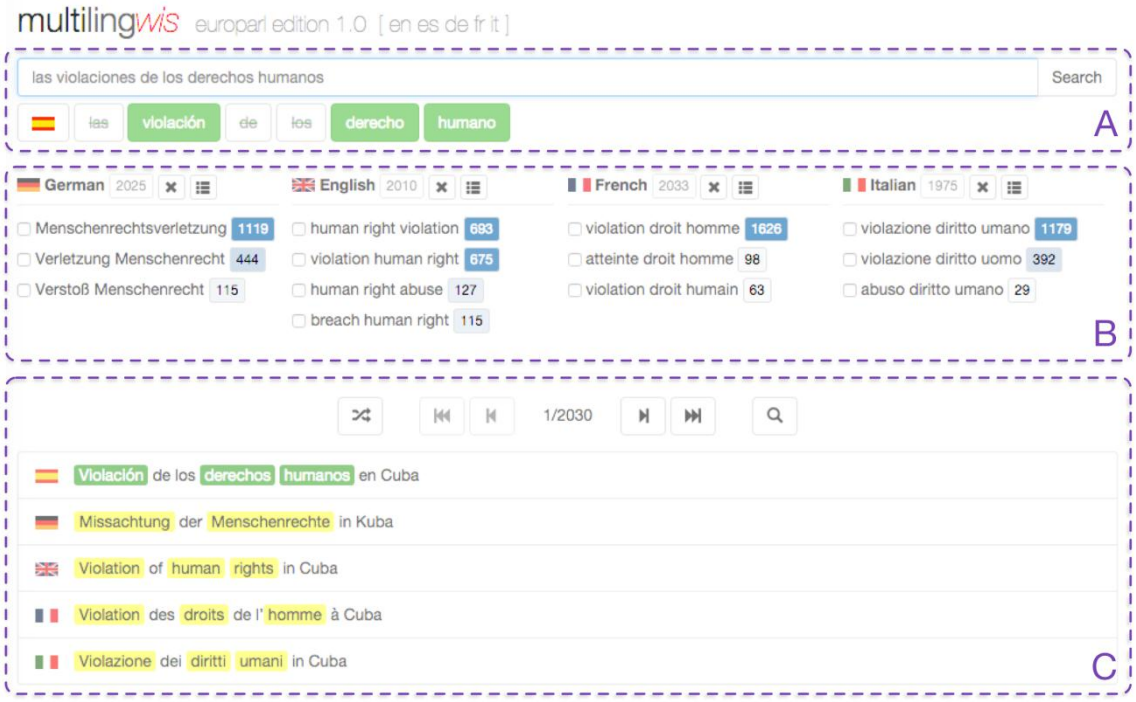
Abstract

We describe a web-based application for searching translations of multi-word units in large, openly available multiparallel corpora. This web application offers a unique resource for multilingual terminologists and translators. The first edition of the tool covers the debates of the European Parliament in five languages: English, French, German, Italian, and Spanish. Our search tool provides a simple and intuitive user interface, which optimally supports content-oriented queries while relieving the user from specifying complicated search expressions in a complex query language. We describe the necessary automatic preprocessing steps of the linguistic data, the retrieval component, and the techniques needed for offering a zero configuration search.

1. INTRODUCTION

Large collections of multiparallel texts, i.e. multilingual documents with aligned paragraphs or sentences across all languages, are openly available. For instance, debates from the European parliament (Koehn 2005), administrative and legislative texts from the EU (Steinberger et al. 2012; Hajlaoui et al. 2014), official documents of the UN in 6 languages (Eisele & Chen 2010) as well as translated movie subtitles (Tiedemann 2012). These corpora are highly useful and valuable for translators, terminologists, and contrastive corpus linguists if they can be exploited effectively.

Multilingwis[142] is a web-based search tool for multiparallel word-aligned corpora that allows its users to find translations[143] of multi-word units efficiently and easily in any of the available languages. The tool is optimized for quick ad-hoc searches and explorations of translation variants and supports content-oriented access to translated multi-word units across multiple languages (typically complex noun phrases, but clearly searches for single



words are also supported). Our goal is to relieve the user from specifying complex search expressions in a corpus query language, and we put substantial effort in providing a zero-configuration query interface that just works as expected. As shown in Figure 1, a user can enter "las violaciones de los derechos humanos" and the system automatically recognizes the language and reduces the input to a sequence of the following lemmatized content words "violación derecho humano", which then will be searched in the corpus. While the order of the search terms must match the order in the sentences of the search language, there is, of course, no order restriction in the corresponding parallel sentences, e.g. we find the following lemmatized English translation variants "human right violation" and "violation human right", which are most popular. As the search is restricted to content words, any parts of speech other than nouns, verbs, adjectives, and adverbs are ignored. In the example sentences presented to the user, naturally, the inflected form including the function words will be shown, e.g. "human rights violations" or "violations of human rights".

Multilingwis is one outcome of an interdisciplinary research project[144] in corpus linguistics and computational linguistics which aims at an automatically computed rich and multi-

[142] https://pub.cl.uzh.ch/purl/multilingwis

[143] Throughout this paper we use the term *translation* to refer to words that express the same content in parallel texts. Finding a translation via a search system as *Multilingwis* does not imply that the search hits are direct translations or that the search words were written in their original language.

[144] http://www.cl.uzh.ch/research/parallelcorpora/sparcling_en.html

layered annotation of multiparallel corpora, including the automatic alignment of text units (e.g. speech turns in debates), sentences, and words. In addition to these levels of alignment, we aim at the sub-sentential alignment of noun groups. In order to fully exploit such complex annotated linguistic structures, a full-fledged linguistic query language is needed with an accordingly steep learning curve for the user. In contrast to a complex query language, *Multilingwis* offers an easy-to-use access to our annotated and aligned data.

1.1. Related Work

Figure 13. User interface with interaction zones A, B, and C marked up

There are a number of web-based search systems available for finding translations of words in parallel corpora. The main benefit of such tools lies the fact that the user immediately sees real-world usage examples of translation pairs in the context of sentences, thus the user is able to judge whether the translation is adequate for a given domain or register. *Linguee*[145] currently supports bilingual searches for 25 languages. Systems such as the highly multilingual translation sharing platform *TAUS DATA*[146], the so-called 'bilingual concordancer' *TradooIT*[147], or *Bilingwis*[148] offer similar functionality for bilingual searches. See Volk et al. (2014) for a more detailed discussion about their usability, the covered languages, and the amount of integrated parallel data.

We know of only two systems which offer searches in multiparallel corpora. The *OPUS-Corpus Query* system[149] (Tiedemann 2012) allows to query several large multiparallel corpora of the OPUS collection using the efficient query infrastructure of the *Corpus Workbench* (CWB) (Evert & Hardie 2011). However, the presentation of the search results is not good because it just shows the parallel sentences and does not include any highlighting of the word-aligned translations of the user's search words. The user must therefore read through the parallel sentences and spot the potential translations by himself. *ParaSol*[150] (von Waldenfels 2011) started as a specialized parallel corpus collection for many Slavic languages with 1 to 4 millions of tokens per language. Additionally, it contains now texts in Romance, Germanic, Baltic, and other European languages. However, due to licencing issues of the text material it is restricted to academic research purposes. *ParaSol* also uses the CWB as its linguistic query engine and also lacks a highlighting of the word translations in the parallel sentences as in the OPUS-Corpus query system.

Our goal is to provide a user-friendly experience of multilingual translation spotting, especially, highlighting the aligned translations in the example sentences and providing frequency distributions of translation patterns which allow the user to quickly identify the most prominent translations variants in order.

[145] http://www.linguee.de

[146] http://www.tausdata.org

[147] http://www.tradooit.com

[148] http://pub.cl.uzh.ch/purl/bilingwis

[149] http://opus.lingfil.uu.se/bin/opuscqp.pl

[150] http://www.parasolcorpus.org

2. PREPROCESSING OF THE LINGUISTIC DATA

We extracted parallel text units in English, French, German, Italian and Spanish from the *Corrected & Structured Europarl Corpus*[151] (Graën et al. 2014), to each of which we subsequently applied the TreeTagger (Schmid 1994) for tokenization, part-of-speech (PoS) tagging and lemmatization. Tagging was done with the language models available from the TreeTagger's web page[152]. We adapted the TreeTagger's tokenizer (abbreviation lexicons, punctuation) and extended its tagging lexicon (especially the German one) with lemmas and PoS tags for frequent words unknown to the language models.

Table 1 shows the amount of tokens per language and the fraction of distinct word forms (=types) which received a TreeTagger lemma. In total, we count 220 million tokens comprising 1 million distinct word forms out of which 214,585 lemmas have been identified by our adapted TreeTagger pipeline. The differences between languages are substantial: the Spanish language model has a low rate of properly lemmatized words, whereas German has the highest absolute number of lemmatized words but due to its large number of distinct word forms cannot reach the lemmatization rate of English. If a token did not receive a TreeTagger lemma, we default it to the word form.

Language	*Tokens*	*Types*	*TT Lemmas*	*TT Fraction*
English	43m	127,105	73,250	57.6%
French	47m	142,898	83,937	58.7%
German	41m	367,159	174,885	47.6%
Italian	43m	181,478	108,147	59.6%
Spanish	45m	175.817	75,187	42.8%

Table 9. Distribution of tokens, types (=distinct word forms), TreeTagger (TT) lemmas, and the fraction of types which received a TreeTagger lemma

We assigned universal part-of-speech tags to each token using the mapping for language-specific tagsets defined by Petrov et al. (2012), and added a few more mappings for some model-specific tags of the TreeTagger. Universal part-of-speech tags helped us to easily separate content words from function words across all languages.[153] Each language has about 22 million content words in our data set.

For sentence alignment, a refined sentence splitting was necessary because some languages use colons or semicolons where others prefer a full stop. For example, the English sentence "However, we have also been guided by another factor, namely the lack of progress on the question of an energy tax." is separated by a colon in French ("Toujours est-il qu'il existait également un autre facteur: l'absence de progrès en matière d'imposition fiscale.") and a full stop in Spanish ("No obstante, había otra cuestión. La del

[151] Altogether 146,652 speech turns are available in all these five languages in CoStEP, which bases on Europarl release v7 (Koehn 2005) and can be obtained freely from http://www.statmt.org/europarl/.

[152] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/#Linux

[153] As content words, we select adjectives (ADJ), adverbs (ADV), nouns (NOUN), verbs (VERB) (including auxiliary verbs). All other 8 categories are treated as function words (including prepositions). Some fine-tuning of this simple classification could improve the results.

estancamiento en el tema del impuesto energético."). Refined sentence boundaries were identified by language-specific rules based on part-of-speech tags and lemmas. Pairwise bilingual sentence alignments was then carried out by the statistical sentence aligner *hunalign* (Varga et al. 2005). Each language has about 1.7 million sentences, with a total of about 16 million pairwise sentence alignments.

For word alignment, we applied the standard tool *GIZA++* (Och & Ney 2003) for each language pair and each direction, resulting in 20 sets of directed 1:n alignments of content words (only adjectives, adverbs, nouns and verbs were aligned). We symmetrized these sets by constructing the union of alignments (see Tiedemann 2011, p.76), thus favoring recall for our application. A total of 110 million content words was the basis for our word alignment.

The linguistic data obtained (tokens with lemmas and part-of-speech tags, sentence segments with their pairwise alignments and word alignments for content words for each language pair) is stored in a relational database. Database features such as *multi-column indexes*, *materialized views* and *stored procedures* allow for an efficient search and retrieval of the corpus data.

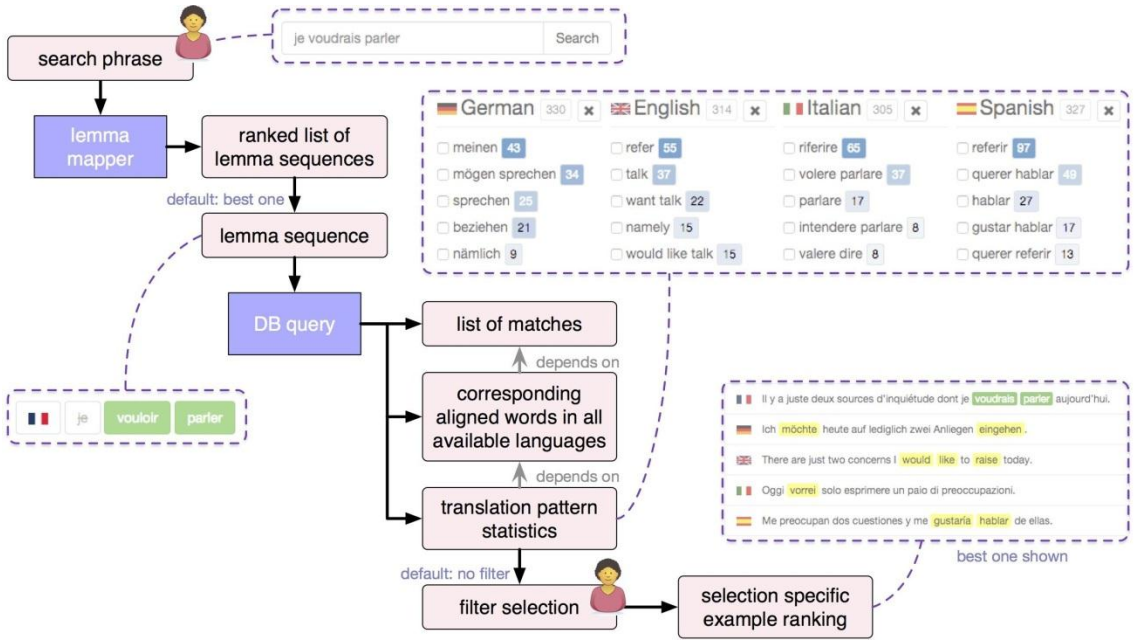


Figure 14. Overall system architecture of *Multilingwis*

3. SYSTEM IMPLEMENTATION

3.1. User Interface

Our user-friendly search interface allows the user to enter a stretch of text in a single input field and to immediately get back a reasonable result set, much alike typical web search interfaces work nowadays. This kind of user-friendliness requires that our application automatically transforms the search string into an appropriate query for the retrieval component. It involves the following steps: (a) tokenization of the input (e.g. separating clitic articles from nouns as in French, "l'auto" splits into "l'" and "auto"), (b) automatic language identification (limited to the 5 languages currently present in our

system), (c) removal of function words (e.g. articles or prepositions, as our system only deals with content words), and (d) lemmatization of inflected content word forms (the lemmas are taken directly from the TreeTagger output). Thus, an input as "las violaciones de los derechos humanos" will be reduced to the following sequence of Spanish content lemmas: "violación derecho humano". We ensure that the system reduces the search strings into the most probable sequence of lemmas of the most probable language. Due to vocabulary overlap between languages, code-switching and proper names, quite a number of word forms can be found in several languages. We also deal with the case that the user enters a sequence of already perfectly lemmatized content words and make sure that each lemma is mapped to itself in the transducer using the frequency counts of its most frequent word form.

If the automatic language identification does not work as expected, the user can select a language-specific search. In this search mode, we also present a drop-down selection for each word form that can be analyzed into more than one lemma. For instance, the inflected word "accepted" can either be a verb with the lemma "accept", or an adjective with the lemma "accepted". We will also develop an advanced search form which supports the search for inflected surface forms and filtering for certain parts of speech, e.g. the lemma "break" can be a noun or an adjective in English, which have different translations.

As can be seen in Figure 1, we present the results of a query in *Multilingwis* in 3 zones. The first zone A shows the user how his input has been interpreted, that is, the sequence of content lemmas searched by the retrieval component. If there is at least one hit, the second zone B displays the distribution of all corresponding translation patterns. At the same time, the third zone C renders an example sentence with the highlighted translations. All example sentences are ordered by a score for good examples (GDEX score), which in our case is currently a value that favors short sentences with a small deviation of sentence length across all languages. More sophisticated GDEX scores as mentioned in the literature (Rychlý et al. 2008) could be computed offline and stored as an attribute for each sentence.

Zone B contains for each language the distribution of all translation variants, in descending order of frequency ("human right violation, human right abuse, breach of human rights"). Each translation variant has a check box attached to it, and the user can enforce the display of an example sentence showing this very translation by checking it. Therefore, zone B offers the user a flexible facility for faceted search refinement. Additionally, each translation variant in zone B provides a hyperlink directed to a new search. Every translation variant can again be queried by a single click, and therefore supports quick explorations across different cross-lingual or monolingual verbalizations of the same concepts.

3.2. Retrieval Component

The first step in the retrieval transforms the query input string into a language-specific sequence of lemmatized content words. This step is implemented by finite-state technology (Lindén et al. 2013), which also allows us to efficiently normalize orthographic variants.

For the language-unspecific search mode, we basically need a transducer that encodes for each language a mapping for each (inflected) word form into its most frequent lemma. Each lemma is annotated by its language code and its frequency class[154] (the higher the

[154] We start from the formula for logarithmic frequency classes which are independent of the corpus size and compare the frequency of a word w against the frequency of the most frequent word w_{max}: $N(w) = \text{floor}(0.5 - \log2(\text{freq}(w)/\text{freq}(w_{max}))$. If a word is in frequency class N, it means that the most frequent word is N times more frequent than w. We transform the numeric value N according to N'(w)=abs(N(w)-N(1/w_{max})-1) in

better) in order to guess the language of the user query string. For instance, the string "die ganze EU" (*the whole EU*) results in the following decorated lemmas for German, "die/de/19 ganz/de/11 EU/de/14", and for English, "die/en/7 EU/en/15", thus preferring German by simply summing up the frequency classes. Function words such as articles are important for language identification, but their lemmas are ignored for the search. According to our experience, such word-based language identification works efficiently, and we see no need to use an external language identifier based on character n-grams (e.g. Lui & Baldwin (2012)), especially, given the fact that such language identifiers typically need at least 30-60 characters for a precise language prediction.

For the language-specific search mode, we built a similar transducer for each language that encodes a mapping for each word form into all admissible lemmas. Each lemma is decorated by its frequency class in order to present ambiguous lemmatizations in descending order of frequency. For instance, *accepted* occurs 4,535 times with the lemma *accept* and only 97 times with the lemma *accepted*, therefore, the default lemmatization of *accepted* is *accept*.

We cannot expect that our users type the search words exactly as we store them in our database. Therefore, we allow a range of spelling variants, for instance, accented or special characters such as "ß" in German or "ç" in French are optionally mapped to their ASCII representation "ss" resp. "c". Furthermore, we handle spelling variants concerning hyphens. If a user searches for "proeuropäischen" ('pro-European'), he is offered both lemmas, "pro-europäisch" and "proeuropäisch".

Once the language-specific sequence of lemmatized content words has been derived from the user input, a database function takes over the tasks to (a) search for a matching sequence of tokens at intervals of at most 4 tokens where the interjacent tokens can only be function words, (b) look up word alignments for each token of each matching sequence[155], and (c) build a statistics of translation patterns on top of it. Figure 2 depicts these steps in the context of the user interface.

Step (a), the search for matching token sequences, starts with a lookup of the first search lemma in a database index based on lemmas and token positions. For every following search lemma, the result is subsequently intersected with another lookup in that index which is limited to the next 4 positions of the previous token found. The tokens in between the matching ones are subsequently filtered for not containing any content word, i.e. their universal part-of-speech tag not being a verb, noun, adjective or adverb.

In step (b), the token sequences are intersected with a database index on the symmetrical word alignments, such that the result set comprises a list of matching tokens in the source language, a list of corresponding tokens in each target language for which we have word alignments, and a sequence of lemmas (=translation variant).

Step (c) ranks these translation variants according to their frequency for each language. The ranking is displayed in zone B.

The whole data set of tokens in the source and target languages together with the respective translation variants are kept until the user performs a new search (either by entering a new query or by clicking on one of the translation variant buttons). Whenever a translation variant is selected or deselected, the sentences shown in zone C get updated

---

order to implement language identification as a maximization of the sum of all N'(w) of a language; for unseen words we set N'=0.

[155] We call the corresponding sequence of lemmas a translation variant.

with the supposedly best example that matches the intersection of any checked translation variant between all languages. If none is chosen, which is the default configuration, all translation variants of the particular language are considered. If there is no example translation for the current selection, the user is advised accordingly.

4. DISCUSSION AND FUTURE WORK

Every preprocessing step for our corpus data can be improved further: Our corpus still contains some misaligned text units. Subsequent sentence and word alignment cannot work for these cases since alignment depend on correct alignment on the text level. In sentence pairs where we align non-corresponding text, our statistical word alignment tool GIZA++ will nonetheless return the most probable alignments, which results in a long tail of incorrect translation variants that occur only once. Therefore, we currently work on the detection of sentences which are not parallel.

Given the fact that we align related languages where translated words often have a similar shape on the level of characters (so-called *cognates*), a more informed approach than the one used by GIZA++ could produce better results (see Sojka et al. (2012)).

For the current version, we already provide domain-adapted external lexicons for the TreeTagger with PoS tags and lemmas, however, there is still room for improved lemmatization. A lot of the more administrative and technical terms in Europarl are not covered by the current TreeTagger models. Another open issue are ambiguous lemmas in the TreeTagger output, for instance, the Italian word "sono" is analyzed into "essere|sonare", which represents the two admissible alternative lemmatizations. We currently store the unmodified TreeTagger lemmas in our database. However, this distorts the translation statistics for the verb "essere". We should therefore either try to disambiguate the ambiguous lemmas (e.g. by preferring the globally more frequent lemma), or we should implement a proper Boolean search for such cases.

Another improvement concerning lemmatization is related to German verbs with separable prefixes, e.g. "ansprechen" (*to address, to speak about*). If such verbs are used as finite forms in main clauses, the finite verb and its prefix are in different topological fields. For instance, "Wir **sprechen** die wichtigen Probleme nicht **an**" (*we do not address the important problems*). In order to provide a proper overview of the translations of "ansprechen", we should attach the prefix "an" to the verb lemma "sprechen" in such cases. This can be done quite reliably and is already implemented by the aforementioned system *Bilingwis* (Volk et al. 2011).

A further question concerning lemmatization is the treatment of words with numbers, e.g. "62jährig" (*62 years old*) in German. Currently, the user has to enter the exact number in order to find translations, which is a bit cumbersome. What a typical user probably would like to see are translation patterns of "DDjährig" where "DD" could be any sequence of digits.

Our formula for the GDEX score currently only considers the consistent shortness of sentences across languages. Although frequent translation patterns will be shown more often than rare ones for obvious reasons, we plan to integrate the frequency of translation patterns into the ranking of the examples.

A different *Multilingwis* edition, for instance, one based on the United Nations corpus with Arabic, Chinese, English, French, Russian, and Spanish would connect less related

languages in a single view. A *Multilingwis* edition with movie subtitles could even be interesting for language learners, however, the text quality (OCR errors, spelling errors) will need some attention.

We did our best to provide an intuitive and practical user interface. In order to gain a better understanding whether our design decisions were adequate, we need to perform usability tests with users interested in multilingual texts and observe via eye tracking devices how they actually interact with our web interface while performing some tasks with *Multilingwis*.

Acknowledgements

This research was supported by the Swiss National Science Foundation under grant 105215_146781/1 through the project "SPARCLING – Large-scale Annotation and Alignment of Parallel Corpora for the Investigation of Linguistic Variation".

References

EISELE, A. & CHEN, Y., 2010. MultiUN: A Multilingual Corpus from United Nation Documents. In *Proc LREC 2010*. pp. 2868–2872.

EVERT, S. & HARDIE, A., 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of Corpus Linguistics 2011*.

GRAËN, J., Batinic, D. & Volk, M., 2014. Cleaning the Europarl Corpus for Linguistic Applications. In *Konvens 2014*. Stiftung Universität Hildesheim.

HAJLAOUI, N. et al., 2014. DCEP - Digital Corpus of the European Parliament. In *Proc LREC 2014*. pp. 3164–3171.

KOEHN, P., 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit*. pp. 79–86.

LINDÉN, K. et al., 2013. Using HFST for Creating Computational Linguistic Applications. In *Computational Linguistics*. Springer Berlin Heidelberg, pp. 3–25.

LUI, M. & BALDWIN, T., 2012. Langid.Py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 25–30.

OCH, F.J. & NEY, H., 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational linguistics*, 29(1), pp.19–51.

PETROV, S., Das, D. & McDonald, R., 2012. A Universal Part-of-Speech Tagset. In *Proc LREC 2012*. pp. 2089–2096.

RYCHLÝ, P. et al., 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proceedings of the XIII EURALEX International Congress*. pp. 425–432.

SCHMID, H., 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. pp. 44–49.

SOJKA, P. et al., 2012. Text, Speech and Dialogue. In Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 370–377.

STEINBERGER, R. et al., 2012. DGT-TM: A freely available Translation Memory in 22 languages. In *Proc LREC 2012*. pp. 454–459.

TIEDEMANN, J., 2011. *Bitext Alignment*, Morgan & Claypool Publishers.

TIEDEMANN, J., 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proc of LREC 2012*. pp. 2214–2218.

VARGA, D. et al., 2005. Parallel corpora for medium density languages. *Proc RANLP*, pp.590–596.

VOLK, M. et al., 2011. Word-aligned parallel text : a new resource for contrastive language studies. In *Supporting Digital Humanities, Conference 2011*.

VOLK, M., GRAËN, J. & CALLEGARO, E., 2014. Innovations in Parallel Corpus Search Tools. In *Proc LREC 2014*. pp. 3172–3178.

VON WALDENFELS, R., 2011. Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In M. Daniela & R. Garabik, eds. *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011*. Bratislava, pp. 156–162.

COMPARATIVE IDIOMS IN CROATIAN MWU APPROACH

Kristina Kocijan
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb

Sara Librenjak
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb

Abstract

This article presents the work aiming to describe comparative idioms in Croatian language for computational processing using NooJ linguistic environment. As a part of a larger project concentrated on annotating and extracting different Croatian idioms as multi-word units (MWUs), this work aims to present automated comparative idiom search in any Croatian text. Using NooJ environment, a user can find any comparative structure in a text and use it for translation, language learning or research purposes.

1. INTRODUCTION

Croatian phraseology is a young discipline whose beginning dates in 1970ies and since that time it was analyzed from different perspectives including semantical, syntactic, etymological, sociolinguistic, stylistic and discourse analysis (Matesic, 1978; Menac, 1978; Mence and Mihalic, 2007), and only recently in computational linguistic (Ljubesic et. al., 2014, Kocijan & Librenjak, 2015).

As the idioms are rooted in the tradition of the language and the society from which they hail from, they need a special treatment in computational linguistics. With our tool, Croatian idioms can be successfully detected in order to be properly matched with their translation in corresponding language, which would eliminate awkward and completely wrong automated translations. Thus, this work seeks to aid not only the successful development of additional language resources for Croatian language (Vuckovic et al., 2010), but a possible assistance in future work relating to machine assisted translation.

In this article, first we introduce the methodology and corpora constructed for the purpose of this work. Then, we explain the technical details of the software which automatically recognizes comparative idioms. Since it was made for the NooJ linguistic environment, we will use the terms pertaining to NooJ software, and thus hope to encourage readers to try and use this environment for their own projects. We will explain the construction of electronic dictionary containing all the necessary comparative idioms in chapter 3, while chapter 4, which explains the types of comparative structures and their treatment in NooJ, takes up the largest portion of this article. For more detailed explanation of NooJ environment see Silberztein (2003). In the last chapter, we will present the evaluation of our tool and comment on its possible uses.

2. METHODOLOGY AND CORPORA

If one wants to make an automated idioms detection tool, one must first collect all the relevant idioms, and construct a digital dictionary. For this purpose, Croatian dictionary of idioms (Menac et al., 2003) was used as a reference, but the list grew in the process of construction, and some additional idioms were added. In our previous work (Kocijan and Librenjak, 2015), we presented an all-around idiom detection tool using the NooJ linguistic environment. This work concentrates on one type of idioms - comparative idioms containing words such as "like" or "as if".

In the second phase, collected idioms were sorted into categories by their syntactic properties. At the same time, we began construction of NooJ files for syntactic processing of texts containing the idioms (called NooJ grammars), which will be discussed in detail in chapter 4.

Simultaneously, three different types of corpora were specifically constructed for the purpose of this work. A smaller corpus of sentences containing only comparisons was made for training the grammars. Then, we constructed two larger corpora for testing. For this research, we specifically made two stylistically different corpora in order to collect statistical data about frequency of comparative idioms in Croatian texts. First test corpus is a general text corpus, collected from the Web, while the second one is a literal works corpus. In the final chapter we will discuss the results in the different corpora and implications about Croatian stylistics which follow from the results.

3. DICTIONARY

There are 533 main entries in the dictionary of comparative idiomatic expressions. Here we define a main entry as the word occurring in the 1st position of an idiom (noun, adjective or a verb) that has a word before *kao* ("like","as"). However, in the null category, i.e. an idiom that starts with *kao*, a main entry is considered the first word that comes after *kao* if it allows the change (*kao heroj - en*^{lit} *as a hero, kao crvena jabuka - en*^{lit} *like a red apple*). The change of the first word is recognized via FLX attribute that is linked to the name of a paradigm, while the change of the second word is recognized via grammar. If no change occurs, entire phrase (without conjunction) is entered as a main entry (*kao ispod cekica - en*^{lit} *as if under the hammer*).

- `heroj,NW+Type=50n+FLX=BRATIC`
- `crven,NW+Type=50an+FLX=PRESPOR+SUFX=jabuka`
- `ispod cekica,NW+Type=500`

To mark the 2nd part of an idiom, we have used the attributes SUFX (suffix), SUFXX (suffix X), SUFXA (suffix A), SUFXB (suffix B) and SUFXC (suffix C) in the following manner: SUFX holds a single word, SUFXX holds an optional expression (one or more words) that may be omitted (*slobodan kao ptica [na grani] - en*^{lit}*free as a bird [on a branch]*), SUFXA and SUFXB split multiple word expression so that we can accommodate more possibilities for the main entry (saka u oko - *en*^{lit} *a punch to the eye* and saka u glavu - *en*$^{lit-}$ *a punch in the head*) or to divide the expression into the part that may and may not change (siromasan kao *crkveni* mis - *en*^{lit} *poor as a church mouse, mlad kao rosa u podne - en*^{lit} *young as a noon dew*) and SUFXC holds the conjunction for coordination (*razlikovati se kao nebo i zemlja - en*$^{lit-}$ *be different as heaven and earth*).

- `slobodan,NW+Type=5an+SUFX=ptica+SUFX=pticica+SUFXX=na grani+FLX=DIVAN`
- `saka,NW+Type=50np+SUFXA=u+SUFXB=oko+SUFXB=glavu`

```
-     siromasan,NW+Type=5aan+SUFXA=crkveni+SUFXB=m
is+FLX
=DIVAN
-     mlad,NW+Type=5anp+SUFXA=rosa+S
UFXB=u
podne+FLX=MLAD
-     razlikovati,NW+pov+Type=5vn+FLX=RAZLIKOVATI+
SUFXA=
nebo+SUFXC=i+SUFXB=zemlja
```

It is possible that one main entry has one or more SUFX, SUFXA and SUFXB attributes. So for the dictionary entry:

```
-     crven,NW+Type=5an+SUFX=paprika+SUFX=rak+SUFX=krv+
SUFX=
mak+SUFX=paradajz+FLX=PRESPOR
```

an adjective 'red' (hr: *crven**)*** has five SUFX attributes, meaning that it is found in five different idiomatic structures of the same subclass. Thus, all valid expressions *crven kao paprika* (en: red as a pepper), *crven kao rak* (en: red as a lobster), *crven kao krv* (en: red as blood), *crven kao mak* (en: red as a poppy), *crven kao paradajz* (en: red as a tomato) are recognized. This variability in form is usually due to synonymy of possible SUFX parts ***(****dati znak* | *mig* | *signal****).*** Although this is not always the case, the meaning still remains the same ***(****tko* | *vrag bi ga znao****)*** (Menac, 1978). However, there are occurrences in our dictionary that have SUFX parts which are quite opposites like

- osjecati se kao riba na suhom - en[ht] *to feel as a fish on dry land* -> en. feel very bad
- osjecati se kao riba u *vodi* - en[ht] *to feel as a fish in water* -> en. feel very good

Any SUFXA may be matched with any SUFXB if found inside the same main entry description. If SUFXA and SUFXB values must not appear together, they have to be entered as a new main entry (bjezati kao stakori s broda koji tone, bjezati kao vrag od tamjana, bjezati kao davo od tamjana)

```
-     bjezati,NW+Type=5vn+FLX=BOJATI+SUFXA=stakor+
SUFXB=
s broda koji tone
-     bjezati,NW+Type=5vn+FLX=BOJATI+SUFXA=vrag+SU
FXA=da
vo+SUFXB=od tamjana
```

Thus, although there are 533 main entries in the dictionary, it actually holds 858 different comparative idioms. Considering all the valid possibilities due to the flective property of nouns, verbs, adjectives and pronouns found in CI and the somewhat free order of its constituents, we are able to recognize many more occurrences by building a syntactic grammar in NooJ. The grammar uses the value of an attribute 'Type' from the dictionary entries to define which words can be found in particular expression. We will explain this attribute in more detail in the following section.

The distribution of comparative idiomatic expressions in the dictionary is given in Figure 1 with the following legend:

- the top row holds the names of CI Types;

Figure 1. Distribution of comparative idioms in NooJ dictionary

4. TYPES OF COMPARATIVE IDIOMATIC EXPRESSIONS

Croatian language is very rich in idiomatic expressions and comparative idioms make one out of five different types of such constructions (Kocijan et al., 2015) that slightly differ from MatesiC (1978). We have observed their occurrences in the text and where thus able to differentiate different subtypes inside this category.

Although all of the comparative expressions appear with conjunction *kao* | *ko* (en. as | like), some of them also appear in corpus with preposition *poput* (en. such as). However, this is only true for idioms that in second position (i.e. after conjunction kao) have either a single noun in nominative (*bijel kao snijeg* -> en. white as a snow), an adjective and a noun both in nominative (razbjezali se kao rakova djeca -> en[ht] to spread around as crab's children, siromasan kao crkveni mis -> en[ht] poor as a church's mouse), a noun in nominative followed by a noun in genitive (*raste kao gruda snijega* -> en[ht] to grow as a snow ball) or a noun in nominative followed by a prepositional phrase (*drhtati kao list na vjetru* -> en[ht] to shiver as a lief on the wind; *kao guska u magli* -> en[ht] as a geese in the fog). This is important since the preposition *poput* does not take nominative construction but genitive one. We have solved this through our grammar since the dictionary allows for only one paradigm to be defined and only for the main entry. Thus, although some expressions were classified at first as fixed or null subtype (Kocijan et al., 2015), we decided to move them into new categories that will allow for the genitive case to be recognized as well.

The second characteristic taken into account was to divide comparative idioms regarding the possibility to change case, number, gender or tense of either only first part of the expression or both first and second part. This required a detailed investigation of all the dictionary entries in order to build the grammars that will recognize all and only valid constructions. Matesic (1978) states that only the verb, adjective and noun positioned before *kao* may change in tense (verbs), gender (verbs and adjectives), number (all three) and case (adjectives and nouns). In the case that a change occurs on any of the words that come after *kao,* the structure ceases to be an idiom. So while zivjet cemo kao bubreg u loju (en: we will live as kidney in suet) is an idiom, the composition zivjet cemo kao bubrezi u loju (en: we will live as kidneys in suet) is not yet a fully

recognized idiom. Still, it can be considered an idiomatic neologism, and is treated as such in our work. This type of language creativity is quite rare according to the corpus, but we predicted it and they are recognized by grammars in case they appear.

There are 18 possible values (50a, 50v, 500, 5fix, 5afix, 5vfix, 5avp, 5va, 5an, 5aan, 5ann, 5anp, 5vn, 5van, 50an, 50n, 50n2, 50np) for the attribute 'Type' coded in the following manner:

El 1st position - 5 : denotes the comparative type of idioms;

El 2nd position - 0 : denotes the null subclass | v : denotes the verbal subclass | a : denotes the adjectival subclass | fix : denotes no change in any part of CI;

El 3rd position - fix | 0[1] : denotes no change in 2nd part of CI | a | an | ann | n | nn | np | nnp | n2 | vp : first letter of word category[2] found in 2nd part of CI in the order of the appearance.

First we will look deeper into the comparative idioms (CI) that do not allow *poput* and second into those that allow it.

4.1. CI not allowing poput

There are 165 dictionary entries that do not allow poput. Of that number, 53 have null 1st position (36 fixed 2nd position and 17 change 2nd position occupied by an adjective or a verb), 4 have a fixed 1st position (and fixed 2nd position) and remaining 108 have changeable 1st position (85 fixed 2nd position and 23 change 2nd position occupied by an adjective) with 92 verbs and 14 adjectives in the 1st position.

The fixed 2nd position is occupied either with a prepositional phrase (bjezati kao *odbuge - en[lit] run like from a plague*), or an adjective (doci kao *narucen - en[lit] come as ordered*), or Dative noun and Nominative noun (*pristajati kao kravi sedlo - en[lit] fit as a saddle fits a* cow), or Nominative noun and Genitive noun (*pun kao* sipak kostica *- en[lit] full like a grenadine is full of seed*).

We recognize eight subtypes in this group of CI coded as 500, 50a, 50v, 5fix, 5afix, 5vfix, 5va and 5avp. Inside the grammar, we have grouped them together if they show similar patterns in their usage.

4.1.1. Type 500

Type 500 CI are all **null CI**, with no 1st position and no changes in the 2nd position. This category has mainly prepositional phrases after conjunction *kao | ko* and since there are no changes, entire phrase is entered in the dictionary as a main entry (*kao* na iglama, kao od sale, kao u raju):

```
-na iglama,NW+Type=500
-od sale,NW+Type=500
-u raju,NW+Type=500
```

[1]Where there is no change of 2nd part of CI, there was no need to mark the word category so we used word 'fix' or '0' instead to mark this section. However, the distribution of word categories found can be seen in Figure 1.

[2]Word categories found are: a - adjective, n - noun, p - prepositional phrase, n2 - coordination of 2 nouns, v - verb.

4.1.2. *Types 50a and 50v*

Types 50a and 50v are similar in a way that they are both **null CI**s but, contrary to the type 500, they have a first word in the 2nd part (adjective or verb) that changes. We were able to accommodate for this change via dictionary and the attribute FLX.

- `lud,NW+Type=50a+FLX=MLAD`
- `pasti,NW+Type=50v+FLX=SJESTI+SUFXA=s+SUFXB=Marsa+SUFXB=neba`

The remaining parts are recognized via grammar (see Figure 1.) so that entire expression is marked as a phraseme, i.e. as <PHR+FRAZEM=kao *da je pao s* Marsa+Type=50v> where PHR (code for phraseme) denotes that the string is a phraseme**,** +FRAZEM holds the recognized string and +Type holds the type of the recognized string which is inherited from the type of the main CI word placed inside the variable F**.**

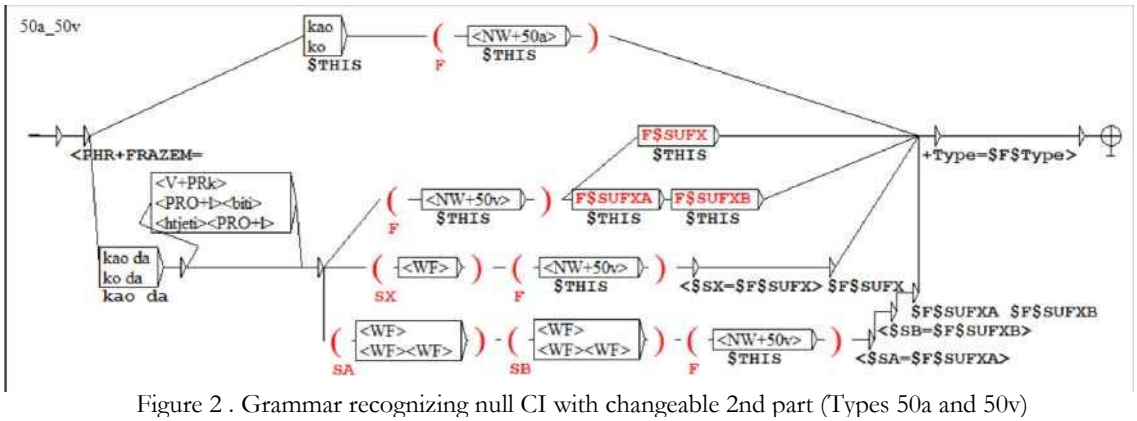


Figure 2 . Grammar recognizing null CI with changeable 2nd part (Types 50a and 50v)

4.1.3. *Type 5fix*

Although there are not many CI that start with a noun (**nominal CI**), we have decided to put them in a separate category. We marked it as '5fix' since neither section, i.e. one prior to and one following conjunctions 'kao' or 'ko', changes its form or position inside the expression. Thus the dictionary entries for this category do not require any paradigm to be connected to them (no FLX attribute).

Dictionary entries for subtype 3 ***(****mrak kao u rogu* - *en**III**.* dark as if you're inside a *horn***,** tisina kao u crkvi - *en**III** silence like in the church***,** tisina kao u grobu - *en**III** silence as if you were in a grave*):

- `mrak,NW+Type=5fix+SUFXA=u+SUFXB=rogu`
- `tisina,NW+Type=5fix+SUFXA=u+SUFXB=crkvi+SUFXB=grobu`

4.1.4. *Types 5va and 5avp*

The category 5va belongs to **verbal CI** while 4avp belongs to **adjectival CI**. They both have changeable parts in the 1st and 2nd part of an expression and do not allow *poput***.** The change in the 1st part is recognized via dictionary as in the previous two categories:

- `osjecati,NW+Type=5va+pov+FLX=SJATI+SUFX=preporoden`
- `gol,NW+Type=5avp+FLX=CRN+SUFXA=od majke+SUFXB=roden`

The change in 2nd part is recognized via grammar (see Figure 3.) either by using the variable $S or variables $SA and $SB. We can check that what is inside these variables agrees with the SUFX or SUFXA and SUFXB of the main entry placed inside the variable $F (for example: $SA=$F$SUFXA checks that whatever is in the variable $SA is equal to the SUFXA of the word found in variable $F).

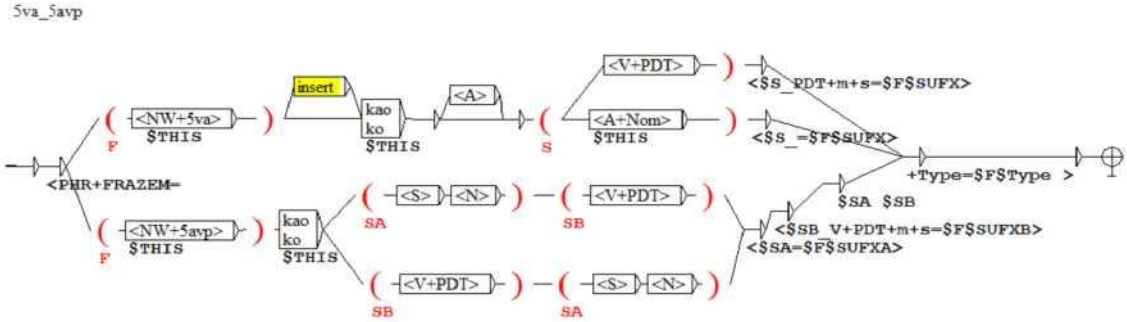


Figure 3. Grammar recognizing subtypes 5va and 5avp

Now, for the two last examples, we can recognize the following forms as well:

- *osjecao* se kao preporoden - en. he felt as if reborn
- osjecala se kao preporodena - en. she felt as if reborn
- gol je kao od majke roden - en[lil]. he is naked as born by mother
- gola je kao od majke rodena - en[lit]. she is naked as born by mother

4.2. CI allowing poput

There are 368 dictionary entries that allow *poput*. Of that number, 48 have null 1st position (all change 2nd position occupied by a single noun or an adjective+noun or noun+prepositional phrase or coordination of two nouns) and remaining 320 have changeable 1st position (all change 2nd position) with 207 verbs and 113 adjectives in the 1st position.

In this category, regardless of the 1st position (null, adjective or verb), the 2nd position may be a single noun, a noun and a prepositional phrase, an adjective and a noun or coordination of two nouns both of which change in number and case (nominative or genitive).

There are no fixed 1st or fixed 2nd positions in this category with 10 subtypes coded 50n, 50an, 50np, 50n2, 5an, 5ann, 5ann, 5anp, 5vn and 5van. They are also grouped together regarding their similarities.

4.2.1. Types 50n, 50np and 50an

These are all subtypes of null CIs with changeable 2nd section that allow *poput*. Type 50n has only one word (noun), while the types 50np and 50an have more than one word segments. Their dictionary entries look like the following examples (*kao zmaj - en[llt] like a dragon, kao grom iz vedra neba - en[llt] like a thunder from a clear sky, kao otvorena knjiga - like an open book*):

- `zmaj,NW+Type=50n+FLX=KRALJ`
- `grom,NW+Type=50np+FLX=BAT+SUFXA=iz+SUFXB=vedra neba`
- `otvoren,NW+Type=50an+SUFX=knjiga+FLX=PRESPOR`

Here, as well, the change of the first word is recognized via the attribute FLX which was enough for the types 50n and 50np. The type 50an however, has a noun that also may change in case and number. In addition, in this category, the main word that usually comes before *kao* | *ko* | *poput* may also appear after it. We solved both these

problems via grammar (see Figure 4) where we allow for any noun (in nominative or genitive case) inside the variable $S, and then check if it matches the value of the SUFX attribute found in the dictionary.

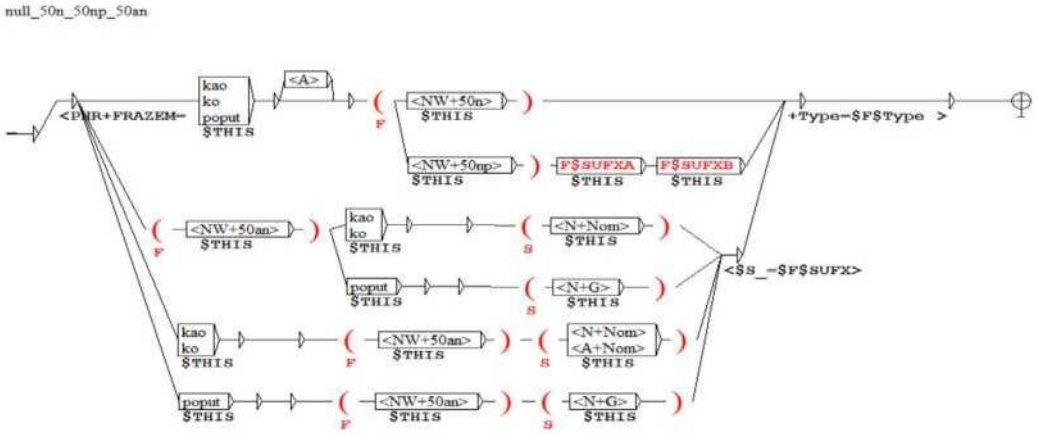


Figure 4. Grammar recognizing subypes 50n, 50np and 50an

4.2.2. *Types 5an, 5aan, 5ann and 5anp*

Types 5an, 5aan, 5ann and 5anp are all subtypes of and adjectival CI which means that they all have an adjective in the first position. In all of these cases, both the adjective in the first position and the word in the second position (adjective or a noun) change.

- `dug,NW+Type=5an+SUFX=vjecnost+FLX=DUG`
  - `ubog,NW+Type=5aan+SUFXA=crkveni+SUFXB=mis+FLX=PRES`
  `POR`
  - `jak,NW+Type=5ann+FLX=JAK+SUFXA=kraljevic+SUFXB=Mar`
  `ko`
  - `slobodan,NW+Type=5anp+SUFXA=ptica+SUFXB=na grani+FLX=DIVAN`

All the valid possibilities are defined with the grammar that recognizes:

- dug kao vjecnost, duga kao vjecnost, dug poput vjecnosti
- ubog kao crkveni mis, uboge kao crkveni mis, ubogi poput crkvenog misa
- jak kao kraljevic Marko, jaki poput kraljevica Marka
- *slobodna kao ptica nag rani, slobodni kao ptice na grani*

4.2.3. *Type 5vn*

Idioms with Type 5vn are verbal CI**s** that have a verb in 1st position and one noun in the 2nd position that may change. It is also valid that the verb moves to the last position or that the idiom is split between the verb and *kao | ko | poput.*

Thus, for the dictionary entry (kretati se kao kornjaca):

- `kretati,NW+pov+Type=5vn+FLX=KRETATI+SUFX=kornjaca`

the grammar recognizes the following examples:

- kretao se kao kornjaca, kretala se poput kornjace, kao kornjaca se *kretao*...

4.2.4. *Type 5van*

Comparative idioms of Type 5van are also all verbal CI**s** that have a verb in the 1st position and an adjective + noun in the 2nd position and all three words may change. To check if the right adjective and noun are used (even when they change case, number

an/or gender), both an adjective and a noun had to be nominalized and as such placed as SUFXA and SUFXB values.

- `planuti,NW+Type=5van+FLX=BLJESNUTI+SUFXA=zivi+SUFXB=vatra`
- `razici,NW+Type=5van+FLX=DOCI+pov+SUFXA=rakov+SUFXB=djeca`

It is also possible that the verb moves from the first to the last position of the expression or that, while in the first position, is interrupted with a noun phrase, prepositional phrase, reflexive pronoun or an auxiliary verb (defined by the node 'insert' in Figure 5).

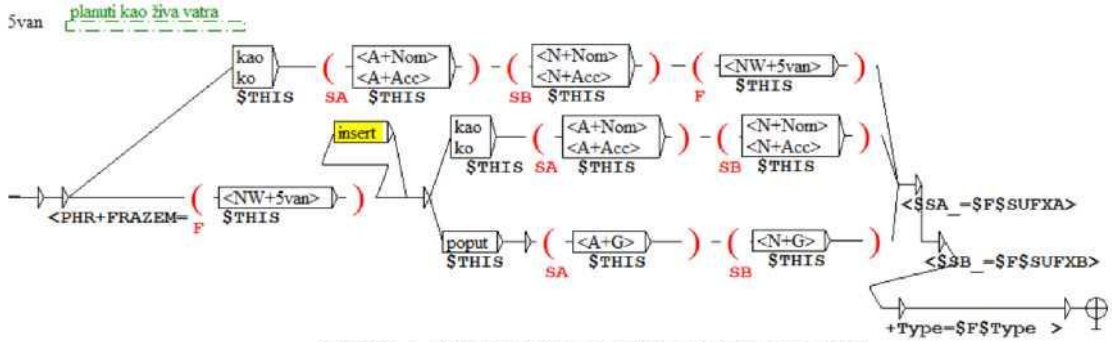


Figure 5. Grammar recognizing subtype 5van CIs

The grammar recognizes also:

- planuti kao ziva vatra, planula je poput zive vatre, kao ziva vatra planuse
- razisli su se kao rakova djeca, razisle su se poput rakove djece

5. RESULTS AND CONCLUSIONS

After training phase on a controlled corpus containing only the sentences with comparative idioms, we found both precision and recall to be 100%. Testing corpora (general-style corpus and corpus of literal works) gave us results of respectively 100% and 100% for precision, while we had 98% and 100% for recall. Table 1 shows all the results, as well as the f-measure.

	Kilo-words (Kw)	Number of structures found	Precision	Recall	F-measure
Training corpus	58 223 w	312	100%	100%	100%
Corpus 1 (web)	2247 Kw	103	100%	98%	98,9899%
Corpus 2 (books)	774 Kw	208	100%	100%	100%

Table 1. Results from the corpora

As it can be seen from the table above, these comparisons are not frequent in the corpora at all. The fact that these comparisons are generally less frequent in oral communication then in texts is supported by linguists (Fink Arnovski et al., 2006), but we were also wondering about the differences by style and purpose of the text. Menac (1978) lists several styles of phrasemes from stylistically neutral to vulgar, among which is a literary style characteristic for written forms of expression with 4 subtypes (literary and artistic, journalistic, scientific, business and administrative). Our examples of corpora belong to the first two subtypes and we have observed differences both in number and type of recognized phrasemes in these two corpora. Namely, in the web

generated general texts corpus, there was a frequency of 0.000045. On the contrary, in specialized literally texts corpus their frequency was 0.00026 or approximately six times more frequent.

We can conclude that this work covers almost all comparative idioms in Croatian language and successfully recognizes them, and can be applied for purposes of computer assisted translation, language learning, computational understanding of Croatian language and many other purposes as a language resource.

References

FINK ARNOVSKI, Z., 2006. Visejezicni rjecnik poredbenih frazema, *Hrvatsko-slavenski rjecnik poredbenih frazema,* Knjigra, Zagreb, p.439.

KOCIJAN, K. AND LIBRENJAK, S., 2015. The Quest for Croatian Idioms as Multi Word Units. To appear in J. Monti, R. Mitkov, G. Corpas Pastor and V. Seretan (eds.) In *Multiword Units in Machine TRanslation and Translation Technology,* John Benjamins Publishing. [in print]

LJUBESIC, N., DOBROVOLJC, K., KREK, S., PERSURIC ANTONIC, M.& FISER, D., 2014. hrMWELex - A MWE lexicon of Croatian extracted from a parsed gigacorpus. In *Language technologies: Proceedings of the 17th International Multiconference Information Society IS2014*. Ljubljana, Slovenia. pp. 25-31.

MATESIC, J., 1978. O poredbenom frazemu u hrvatskom jeziku. In *Filologija 8*. Zagreb, pp. 211-217.

MATESIC, J., 1982. Frazeoloski rjecnik hrvatskoga ili srpskog jezika. Zagreb: Skolska knjiga.

MENAC, A., 1978. Neka pitanja u vezi s klasifikacijom frazelogije. In *Filologija 8,* Zagreb, pp. 219-226.

MENAC, A., FINK-ARSOVSKI, Z. AND VENTURIN, R., 2003. Hrvatski frazeoloski rjecnik. Zagreb: Naklada Ljevak.

MENAC-MIHALIC, M., 2007. Hrvatski Dijalektni Frazemi S Antroponimom Kao Sastavnicom. In *Folia Onomastica Croatica,* no. 12/13, pp. 361-85.

SILBERZTEIN, M. 2003. *NooJManual.* Available at www.nooj4nlp.net [Accessed July 2015]

VUCKOVIC, K., TADIC, M., BEKAVAC, B. 2012. Croatian Language Resources for NooJ. In: *CIT. Journal of computing and information technology 18*. pp. 295-301.

IDENTIFICATION AND CLASSIFICATION OF PHRASEMES IN AN L2 LEARNER CORPUS OF ITALIAN

Christine Konecny
University of Innsbruck
Christine.Konecny@uibk.ac.at

Andrea Abel
European Academy of Bolzano - South Tyrol
Andrea.Abel@eurac.edu

Erica Autelli
University of Innsbruck
Erica.Autelli@uibk.ac.at

Lorenzo Zanasi
European Academy of Bolzano - South Tyrol
Lorenzo.Zanasi@eurac.edu

Keywords: learner corpora, formulaic sequences, referential phrasemes, communicative phrasemes, structural phrasemes

Abstract

The present article deals with the description of the research approach and methodology adopted within the project LEKO, in which, among others, a learner corpus of Italian as an L2 (texts written by South Tyrolean students whose L1 is German) has been analyzed with regard to the – correct or erroneous – use of phraseological units. Our aims of studying and describing the phrasemes to be found in the corpus are mainly (1) to find out which phrasemes are actually used by the students and how they can be classified, (2) to detect recurrent mistakes as well as possible error causes, and (3) to develop suitable and innovative phraseodidactic material which is based on the results of the learner corpora analysis and thus adapted to the needs of South-Tyrolean L2 learners of Italian. In this paper, we mainly concentrate on the first of the three mentioned aspects, explaining the methodological background of our corpus analysis and illustrating how we proceeded in assigning the discovered phrasemes to specific phraseological categories and subcategories. The single phraseme types and phenomena will be also illustrated by examples taken from the LEKO corpus.

1. INTRODUCTION

In spite of the increased application of corpus-based methods in phraseological research in the past years (cf. e.g. Heid 2005; Heid/Weller 2010; Steyer 2013), the initiating interest for phraseological aspects in learner corpora research (cf. Paquot/Granger 2012) and the constantly growing number of phraseodidactic studies (cf. Kühn 1987; 1992; Lorenz-Bourjot/Lüger 2001; Hallsteinsdóttir 2011; Hallsteinsdóttir et al. 2011; Gonzáles Rey 2012; 2013a; 2014; Konecny et al. 2013; Sułkowska 2013),[156] suggestions of

[156] According to Gonzáles Rey (2013b: 7), the term *phraseodidactics* was used for the first time in German by Kühn (1987) (both in the adjectival version *phraseodidaktisch* and in the form of the

appropriate criteria for identifying, classifying and analyzing phrasemes in learner corpora seem to be still underrepresented. Studies of such kind could be useful not only for revealing the actual use of phrasemes at various levels of the *Common European Framework of Reference for Languages* (CEFR 2001) and for detecting recurrent mistakes and error causes, but also for developing suitable didactic material in order to achieve a certain target level in the use of phraseological units at different CEFR levels.

Within the LEKO[157] project, carried out in cooperation between the Department of Romance Languages of the University of Innsbruck and the Institute for Specialized Communication and Multilingualism of the European Academy of Bolzano/Bozen (EURAC),[158] we aim to describe the use of phrasemes on the part of German-speaking L2 learners of Italian, by combining both quantitative and qualitative methodological approaches. To this end, we have been analyzing a subset of the learner corpus KOLIPSI, which is available at EURAC and consists of Italian and German L2 productions that have already been assigned to CEFR levels and contain also a variety of metadata such as school type, gender and L1 of the learners (cf. Wisniewski/Abel 2012: 16f.). The authors of the learner texts are students of educational institutes located in South Tyrol, a province of Northern Italy characterized by a strong multilingualism.[159]

In the following, we will illustrate how the LEKO corpus is composed and how we proceeded in analyzing it with regard to phraseological units (chapter 2), which phrasemes we found in the learner texts and what are the different phraseological (sub)types into which we classified them (chapter 3).

2. CORPUS DATA AND METHODOLOGICAL APPROACH

The LEKO corpus, i.e. the learner texts analyzed within the LEKO project, constitutes a subcorpus of the above mentioned KOLIPSI corpus; the latter contains altogether about 1,000 texts with Italian as an L2 (~ 200,000 tokens) and about 270 texts with German as an L2 (~ 65,000 tokens). The LEKO subcorpus comprises 290 Italian texts (including 54,651 tokens[160]) written by 145 German-L1 pupils (i.e. two texts per pupil) and covers the CEFR levels A2-C1 (the majority of the texts, however, belongs to the levels B1 or B2). The data collection for the KOLIPSI corpus took place during the school year 2007/08 (cf.

---

noun *Phraseodidaktik*), referring to the didactics of German as a foreign language (in the respective article, Kühn compares the phraseodidactics of German as an L2 metaphorically with the Sleeping Beauty, which should no longer be sleeping). However, as is pointed out e.g. in the studies by Ettinger (2010: 88) and Gónzales Rey (2012: 68ff.; 2013b: 7), approaches of teaching phraseology can already be found before the introduction of the term *phraseodidactics* itself: Indeed, a first attempt of including the didactics of phrasemes in language teaching goes back to a French-German vocabulary book published in 1900 (cf. ibid.) by A. Martin and F. Leray (*Exercices sur les Idiotismes et les proverbes de la conversation allemande, classés suivant le plan des Mots allemands groupés d'après le sens de MM*. Bossert et Beck, Paris: Hachette). Also Ch. Bally, who is commonly regarded as the "father" of phraseology (cf. Konecny 2010: 15ff.), in the second volume of his *Traité de Stylistique française* (1909; Heidelberg: Winter) already proposed around 300 phraseological exercises for learners of French (cf. Gónzales Rey 2012: 70).

[157] The acronym LEKO stands primarily for the German project title "*Lexemkombinationen und typisierte Rede im mehrsprachigen Kontext*" ('Lexical combinations and typified speech in a multilingual context') (cf. Konecny/Autelli 2014).

[158] The LEKO project (cf. www.leko-project.org) is funded by the Autonomous Province of Bolzano/Bozen - South Tyrol (Division for the Promotion of Education, Universities and Research).

[159] There are three official languages, namely Italian, German and Ladin (i.e. the Ladin varieties of Rhaeto-Romance), which are accompanied nowadays by several more minority languages.

[160] The quoted number of tokens does not refer to the number of lexemes only, but contains also punctuation marks such as full stops, commas and colons.

Vettori/Abel 2012: 10) and is based on two standardized tests for written production. As far as the text types and genres as well as the subjects of the texts are concerned, the two tasks the students were given consisted of (1) writing an e-mail to a friend on a certain event at the supermarket (~ narrative text) and (2) writing a letter to a friend on holiday planning (~ argumentative text) (cf. Wisniewski/Abel 2012: 24).

For the annotation of the LEKO corpus we used both automatic and manual approaches: (1) First, the texts were tokenized and lemmatized and POS-tagged with Treetagger (cf. Schmid 1994); then SaltNPepper was used as conversion framework (cf. Zipser/Romary 2010) and ANNIS for data query and visualization (cf. Zeldes et al. 2009). (2) Secondly, the texts were annotated manually through MMAX2 (cf. Müller/Strube 2006; for the approach and workflow cf. also Abel et al. 2014; Glaznieks et al. 2014).

However, before being able to start with manual annotation, it was necessary to define our conception of "phraseme", to determine specific criteria to be applied for identifying and classifying them and to elaborate a detailed annotation scheme which served afterwards as a reference model for the actual annotation. For this purpose, we adopted a combination of deductive and inductive methods, i.e. on the one hand following current concepts present in pertinent studies (cf. e.g. Venier 1996; Burger 2003; Hausmann 2004; Voghera 2004; Stein 2007; Paquot/Granger 2012; Abel et al. 2014), and on the other hand pre-analyzing selected texts of the LEKO corpus in order to get an idea of which phrasemes are actually used by the learners and in order to satisfy the need of an approach that allows for the identification of a broad variety of phraseological phenomena at different linguistic levels.

As to the main categories of phrasemes, our typology is based primarily on the functions the phrasemes fulfil in communication and follows the so-called "basic classification" proposed by Burger (2003: 36ff.), i.e. distinguishing between (1) referential phrasemes, (2) communicative phrasemes and (3) structural phrasemes.[161] However, as far as the subclasses within these three main categories are concerned, our classification differs from others suggested in secondary literature and results from our corpus data; thus, for the subcategories we opted for applying a "mixed classification" in terms of Burger (2003: 50), who points out that for the phraseological analysis of text corpora it often turns out to be useful not to refer to one classification criterion only, but to "mix" several criteria, e.g. pragmatic, semantic, syntactic and structural criteria. One of our main aims of adopting such an approach is to grasp as many phraseological phenomena as possible that are difficult to learn and/or important for learners, especially for German-speaking learners of Italian as an L2.

As our conception of phrasemes is therefore a very broad one, from the terminological point of view it may sometimes seem more accurate to use the more general term "formulaic sequence" (FS), which we use as a form of quasi-synonym of "phraseme" and which, according to Wray (2002: 9), can be defined as "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar".

The above-mentioned manual annotation of our corpus comprises four major steps: (a) identifying a particular phraseme/FS and assigning it to a specific, predefined (sub)category (independently of whether the FS is used correctly or not), (b) annotating possible errors, (c) annotating the morpho-syntactic structure of the phraseme/FS and the syntactic

[161] A similar approach, but less detailed with regard to the distinction of lexical phenomena in learner language was adopted within the MERLIN project (cf. Wisniewksi et al. 2014: 17).

functions of its constituents (applicable only in the case of selected FS types), (d) explaining possible errors (including possible interferences from German and in a few cases also from English). In the following chapter 3, we will deal with the main aspects concerning the first step (a) of our manual annotation, which regards the identification and classification of FS used on the part of the learners.

3. IDENTIFIED PHRASEMES AND THEIR CLASSIFICATION INTO (SUB)CATEGORIES

3.1. Referential phrasemes

According to Burger (2003: 36), referential phrasemes are those which refer to "Objekte, Vorgänge oder Sachverhalte der Wirklichkeit (sei es der 'wirklichen' oder fiktiver Welten)", i.e. to objects, processes or circumstances of the extralinguistic reality, both of the "real" and the fictitious reality. Within the category of referential phrasemes, we further distinguish between non- or semi-idiomatic (3.1.1.) and idiomatic referential phrasemes (3.1.2.), which will be illustrated in the following sections, including also the various subcategories, which amount to four in both cases.

3.1.1. Non- or semi-idiomatic referential phrasemes

Referential phrasemes can be regarded as not or semi-idiomatic if their meaning is compositional, i.e. if it results – totally (in the case of non-idiomatic phrasemes) or at least partially (in the case of semi-idiomatic phrasemes) – from the sum of the meanings of its single constituents. The four subgroups we identified in our corpus and into which we classified the phrasemes of this group are listed in paragraphs (a) to (d):

(a) The first subtype is represented by collocations. We assume that the prototypical case are restricted collocations, which are constrained by lexical restrictions that are "arbitrary and differ markedly from one language to another" (Paquot/Granger 2012: 136), like in the case of *perform* or *do a task* in English (ibid.) or of *die Zähne putzen* ('to brush one's teeth') in German (Burger 2003: 51). In some cases, a collocation can also be semi-idiomatic, namely when the collocational "base" is used literally, the "collocator" instead in an idiomatic meaning (cf. Hausmann 2004: 313ff.), like in *numero verde* ('a (toll)free (phone) number', lit. "a green number") (Konecny/Autelli 2012: s.v. *Glossary*). However, during the pre-analysis of selected corpus texts we discovered that lexical collocations *stricto sensu* occur in them only rarely, for which it seemed useful to adopt a broader understanding of collocations for our analysis, including also sequences such as *andare in vacanza* ('to go on holiday'), *andare al supermercato* ('to go to the supermarket') and *stare sulla spiaggia* ('to be on the beach'). The latter are all examples taken from the LEKO corpus: At first sight, they may seem freely combined, but from an interlingual and contrastive point of view it can be noticed that in these combinations at least the prepositions are idiosyncratically bound to the actual verbs. Thus, for sequences of this kind we suggest to use the term "loose collocations": They contain both content and form words, but while the content words can be combined (more or less) freely, the use of the form word (mostly a preposition) is specific in each case and not predictable or rule-based.[162] We included such cases in our category of collocations for two reasons, namely because the authors of our texts used them very often and in several cases we also detected errors in the use of prepositions (e.g. **guardare nell'internet* instead of

[162] In the case of *stare sulla spiaggia*, however, also the content words cannot all be translated literally into German, where the word for 'stay' is not used, but rather the combination *sich am Strand aufhalten* or *Zeit (den Tag, den Nachmittag…) am Strand verbringen*.

su/in internet 'to look/search on the internet', probably due to interference from Germ. *im Internet nachsehen*), and because such combinations could and should also be taught and learned as fixed units; hence, they are also relevant for phraseodidactics.

(b) The second subgroup we included in the category of non- or semi-idiomatic referential phrasemes are (nominal) compound equivalents, i.e. which consist of a *nomen determinatum* (functioning as head of the whole phrase) and another part which functions as *determinans* of the head noun (this second part may be another noun, an adjective or a prepositional phrase). Some examples to be found in the LEKO corpus are *vita notturna* ('night life'), *albergo a tre stelle* ('three stars hotel') and *pensione completa* ('full board'). Within this category we annotate only (nominal) compound equivalents which function in a similar way to collocations, that means that at least one element of the combination is transparent, used in its literal meaning and interpretable as the "base", and the meaning of the whole expression must be (semi-)deducible from the meanings of the single elements (cf. Hausmann 2004: 317).

(c) The third subcategory are so-called "syntagmatic verbs" (Ital. *verbi sintagmatici*, cf. Venier 1996), i.e. verbs consisting of two or more words which have a compositional meaning and in which one of the words is usually an adverb indicating the (concrete or abstract) direction of an action. Examples from the LEKO corpus are *andare giù* ('to go down'), *buttare via qcs.* ('to throw away'), *pensarci sopra* ('to think about it') and *portarsi dietro qcs.* ('to take something along'). We opted for classifying syntagmatic verbs as a separate subcategory within the non- or semi-idiomatic referential phrasemes (i.e. not including them in the collocations category) for two reasons: firstly, because the adverb contained in them can be regarded as an integral part of the verb itself, thus there is no combination of two concepts (as is normally the case with collocations), and secondly because they are used relatively often (and mostly correctly) by the authors of the LEKO texts and hence are important for our analysis.

(d) The fourth and last subcategory are adverbial expressions consisting of more than one word that have a (more or less) transparent meaning and thus function as non- or semi-idiomatic referential phrasemes. Semantically, these expressions may have, among others, temporal, local or modal functions or serve as sentence adverbs. The single words of the expressions are always used in the same way and order. Examples for this category to be found in the LEKO corpus are *all'improvviso* ('suddenly'), *per fortuna* ('fortunately') and *di nuovo* ('once again').We think that it is useful to classify adverbial expressions of this kind as an own category especially in order to distinguish them from the more "classical" forms of non-/semi-idiomatic referential phrasemes. Moreover, as the inductive pre-analyses of the corpus data had shown, adverbial expressions seem to be used rather frequently. It has to be considered, however, that, in contrast to the subclasses (a) and (b), there is normally no combination of two concepts, but only one extralinguistic concept (as is also the case with syntagmatic verbs).

As a very special case and a type of exception, we decided to also include in the fourth subgroup the pronominal expression *tutti e due* (*i/le/gli*) ('both'). Although this combination is rule-based and not phraseological in the strict sense of the term, the reason why we decided to include it is that very often it is used in the wrong way (e.g. **tutti due le parti* instead of *tutte e due le parti*), probably because of the interference from its German equivalent *alle zwei/alle beide(n) (…)*; therefore, the combination can be regarded as relevant for phraseodidactic material and should also be included.

3.1.2. Idiomatic referential phrasemes

Referential phrasemes are defined as (semantically) idiomatic if their overall meaning is non-compositional and does not correspond to the sum of the meanings of the single constituents (cf. e.g. Burger 2003: 31f.; Konecny 2010: 93ff.).

(a) The first group identified within idiomatic referential phrasemes are idioms in the narrow sense of the term. Although it showed that in our corpus they are not used very often (especially at lower levels) and are mostly left out by the learners, it seemed useful to accommodate for idioms as an extra category, since they are probably the most prototypical phraseme category. On the levels B2 and C1 we found at least some idiomatic expressions, like *rendersi conto di* ('become aware of') and *essere alle prese con qcs.* ('to struggle with something'). The rare occurrence of idioms can be seen in accordance with the hypothesis that, in order to communicate successfully, idioms are more important for receptive/passive competences than for active language use (cf. e.g. Hausmann 2004: 313).

(b) The second subcategory are compound equivalents which have an idiomatic meaning, as in the case of *anima gemella* ('soulmate') and *luna di miele* ('honeymoon') (examples from Voghera 2004: 58; 62). However, in our corpus texts there were almost no occurrences for this type, since the compound equivalents used by the learners are normally non-idiomatic.

(c) The third subclass is represented by idiomatic syntagmatic verbs, cf. the examples of *passar su* (in the meaning of 'to not care about something') and *buttar giù* (in the meaning of 'to depress somebody' or 'to put down something in writing'), cited by Venier (1996: 150). As to our corpus, these FS are used rarely and are limited to few examples which are used more often and seem to be frequent in spoken language, like e.g. *mandare giù qcs.* ('to swallow something', in the figurative sense 'to bear something') and *portarsi dietro qcs.* ('to take something along with oneself').

(d) Also adverbial expressions can be included in the category of non-idiomatic referential phrasemes, but only if their meaning is non-compositional, like it is the case of *di punto in bianco* ('out of the blue') and *in fin dei conti* ('in the end', 'finally') (examples from Voghera 1996: 67). However, hardly any examples of this kind occur in the LEKO corpus; regardless we provided for this subcategory in order to have a (more or less) parallel subclassification to that of non- or semi-idiomatic phrasemes (3.1.1.).

3.2. Communicative phrasemes

Communicative phrasemes have the function to fulfil specific tasks for the creation, the realization or performance and the conclusion of communicative activities. They belong to pragmatics, for they depend on the context and on recurring communicative activities or routines. This category of FS is quite heterogeneous; thus it is difficult to grasp and annotation may therefore also be subject to heterogeneity. Following Stein (2007: 226ff.), we further distinguish between two subcategories:

(a) The first subgroup are communicative phrasemes which are bound to specific situations in that they fulfil social functions and refer to a specific pattern of action. Such phrasemes serve to fulfil specific speech acts in a conventional way, as is the case of formulas of greeting, farewell and addressing, apologies, formulas of approval or affirmation, wishes, etc. The LEKO corpus contains numerous FS of this kind, like e.g.

Come stai? ('How are you?'), *Che ne pensi tu?* ('What do you think about it?') and *Tutto bene?* ('Is everythink ok?').[163]

(b) Other communicative phrasemes instead are *not* bound to specific situations in that they fulfil primarily text organizing and interaction organizing functions, serving e.g. to build metacommunicative and text internal references, to guide attention, to ensure understanding, to comment utterances, etc., such as *secondo me* ('in my opinion'), *a dir la verità* ('to tell the truth') and *Per me è lo stesso* ('For me, it doesn't matter.'), which are all examples taken from the LEKO corpus.

3.3. Structural phrasemes

Structural phrasemes, which represent the smallest group within the "basic classification" according to Burger (2003: 37), are constructions that "only" have the function of creating relations within a text (whereas referential phrasemes refer to specific entities or facts (actions, events, states) of the extralinguistic reality and communicative phrasemes have specific functions regarding communicative activities). We further subdivided this type of FS into:

(a) "*Poliremathiche congiunzionali*" according to Voghera (2004: 68), i.e. structural phrasemes in the form of polylexical or correlative conjunctions or connectors, like e.g. *nonostante che* ('although') and *fermo restando che* ('without prejudice to the fact that') (cf. ibid.). Some examples from the LEKO corpus are *o … o* ('either…or'), *visto che* ('in consideration of', 'since') and *dall'altra parte* ('on the other hand').

(b) "*Poliremathiche preposizionali*" according to Voghera (2004: 67f.), i.e. structural phrasemes in the form of polylexical prepositions like *a carico di* ('account of'), *assieme a* ('together with'), *riguardo a* ('as to...', 'regarding') (cf. ibid.). As to the examples discovered in the LEKO corpus, we mostly found instances of polylexical local prepositions such as *in mezzo a* ('in the middle of' or 'among'), *nei pressi di* ('near/close to') and *nel cuore di* ('in the heart of').

3.4. Own dimension: The category "non-existing"

In order to record several possible kinds of error causes, we have provided for a separate category named "non-existing" for those cases in which phrasemes existing in German (or English) were translated literally into non-existing Italian expressions. Such expressions often give problems in comprehension and could sometimes not even be understandable for a native speaker of Italian (especially without any knowledge of German). Within this category, a further distinction has been made between (a) cases in which the expression is totally unusual in Italian and cannot be accepted by a native speaker (e.g. **queste vacanze divantano il martello* 'this holiday will be great', lit. "this holiday becomes the hammer", used due to interference from Germ. *dieser Urlaub wird der Hammer*; cf. Konecny et al. in prep.) and (b) cases in which the expression is not standard in Italian but could be accepted or at least does not hinder understanding (e.g. **fare surfing* instead of *fare surf*).[164]

[163] Although formulaic sequences (FS) are normally multiword units, this subcategory represents an exception because there may be cases in which single words fulfill the same function as polylexical communicative phrasemes, like in the cases of *Ciao!* ('Hello!' or 'Buy!') and *Cavolo!* ('Man!', 'Damn!').

[164] We are aware of the fact that in the case of this category we are already interpreting data with regard to errors and possible interferences, which is normally not the case within the first step (a) of our analysis (cf. chapter 2.), i.e. the phraseme identification and classification. However, it seemed

4. CONCLUSIONS AND OUTLOOK

The theoretical and methodological approach presented in this paper, which consists of combining deductive and inductive methods as well as quantitative and qualitative data, has proven for us to be a good way to examine and describe the use of phraseological units on the part of L2 learners of Italian. However, the learner corpora analyses have not been concluded yet; our next steps will consist of the quantitative and qualitative evaluation of the data collected during the annotation process as well as in error analysis and explanation; doing this we will also try to discover if the results will confirm some expected trends and advanced hypotheses.

Another essential element to do in the future will be to connect the results of the analyses with phraseodidactics (possibly also with regard to different CEFR levels), i.e. to create, proceeding ideally by semantic areas, concrete phraseodidactic material containing phrasemes which learners should know at various levels. An important aspect in this regard seems to be that error analysis in FS cannot be completely separated from the analysis of mistakes connected with the semantics of single words (which should thus be accounted for also in phraseodidactic exercises). This is because the use of FS is always connected, among other things, with divergent polysemy distributions and combinatory profiles of words in different languages, as will be shown and illustrated with examples from the LEKO corpus in Konecny et al. (in prep.).

In this article we dealt exclusively with the preparatory work concerning the annotation of our learner corpus. The further steps, i.e. the annotation process itself as well as the analysis of the correct or erroneous use of phraseological units, will instead be explained in more detail in Abel et al. (in prep.) and Konecny et al. (in prep.).

References

ABEL, A., KONECNY, C. AND AUTELLI, E., in prep. Annotation and Error Analyses of Formulaic Sequences in an L2 Learner Corpus of Italian. In: *Proceedings of the 3rd Learner Corpora Research Conference, LCR 2015, Nijmegen, Netherlands, September 11-13, 2015*. Leiden et al.: Brill/Rodopi.

ABEL, A., WISNIEWSKI, K., NICOLAS, L., BOYD, A., HANA, J. AND MEURERS, D., 2014. A Trilingual Learner Corpus Illustrating European Reference Levels. *Ricognizioni. Rivista di lingue, letterature e culture moderne*, 1(2), pp.111-126. [online] Available at: <http://www.ojs.unito.it/index.php/ricognizioni/issue/view/64> [Accessed 24 October 2015].

BURGER, H., 2003. *Phraseologie. Eine Einführung am Beispiel des Deutschen*. 2nd ed. Berlin: Schmidt.

CEFR, 2001. = COUNCIL OF EUROPE, 2001. *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: CUP.

ETTINGER, S., 2010. Phraseologie und Wortschatzerwerb. Anmerkungen zu A. Martin und F. Leray: *Les idiotismes et les proverbs de la conversation allemande*. Paris 1900. *FLuL*, 39, pp.88-102.

GLAZNIEKS, A., NICOLAS, L., STEMLE, E., ABEL, A. AND LYDING, V., 2014. Establishing a Standardised Procedure for Building Learner Corpora – a Response to Demands and Suggestions of Users. *APPLES. Journal of Applied Language Studies*, 8(3), pp.5-20. [online] Available at: <http://apples.jyu.fi/issue/view/15> [Accessed 27 October 2015].

GONZÁLES REY, M.I. ed., 2012. De la didáctica de la fraseología a la fraseodidáctica. *Paremia*, 21, pp.67-84.

important to annotate such constructions within this step in order not to "lose" them for error annotation afterwards.

GONZÁLES REY, M.I. ed., 2013a. *Phraseodidactic Studies on German as a Foreign Language. / Phraseodidaktische Studien zu Deutsch als Fremdsprache*. Hamburg: Dr. Kovač.

GONZÁLES REY, M.I., 2013b. Presentation: Phraseodidactics, an applied field of Phraseology. In: I. Gonzáles Rey, ed. *Phraseodidactic Studies on German as a Foreign Language*. Hamburg: Dr. Kovač. pp.7-10.

GONZÁLES REY, M.I. ed., 2014. *Outils et méthodes d'apprentissage en phraséodidactique*. Fernelmont, Belgique: EME Editions.

HALLSTEINSDÓTTIR, E., 2011. Aktuelle Forschungsfragen der deutschsprachigen Phraseodidaktik. *Linguistik online*, 47, pp.3-31. [online] Available at: <http://www.linguistik-online.de/47_11/hallsteinsdottir.pdf> [Accessed 1 July 2015].

HALLSTEINSDÓTTIR, E.; WINZER-KIONTKE, B. AND LASKOWSKI, M. eds., 2011. *Phraseodidaktik. / Phraseodidactics* (= *Linguistik online*, 47, 3/2011). [online] Available at: <http://www.linguistik-online.de/47_11/> [Accessed 1 July 2015].

HAUSMANN, F.-J., 2004. Was sind eigentlich Kollokationen? In: K. Steyer, ed. *Wortverbindungen – mehr oder weniger fest*. Berlin; New York: de Gruyter. pp. 309-334.

HEID, U., 2005. Corpusbasierte Gewinnung von Daten zur Interaktion von Lexik und Grammatik: Kollokation – Distribution – Valenz. In: F. Lenz and S.J. Schierholz, eds. *Corpuslinguistik in Lexik und Grammatik*. Tübingen: Stauffenburg. pp.97-122.

HEID, U. AND WELLER, M., 2010. Corpus-derived data on German multiword expressions for lexicography. In: R. Vatvedt Fjeld and J.M. Torjusen, eds. *Proceedings of the 15th Euralex International Congress, Oslo, 7-11 August 2012*. Oslo: Department of Linguistics and Scandinavian Studies of the University of Oslo. pp.331-340.

KONECNY, C., 2010. *Kollokationen. Versuch einer semantisch-begrifflichen Annäherung und Klassifizierung anhand italienischer Beispiele*. München: Martin Meidenbauer.

KONECNY, C. AND AUTELLI, E., 2012. *Italienische Kollokationen. Wortverbindungen der italienischen und deutschen Sprache im Vergleich. Ein Forschungsprojekt*. [online] Available at: <http://www.kollokation.at> [Accessed 24 October 2015].

KONECNY, C., HALLSTEINSDÓTTIR, E. AND KACJAN, B. eds., 2013. *Phraseologie im Sprachunterricht und in der Sprachendidaktik. / Phraseology in language teaching and in language didactics*. Maribor: Mednarodna založba Oddelka za slovanske jezike in književnosti, Filozofska fakulteta.

KONECNY, C. AND AUTELLI, E., 2014. *LEKO – Lexemkombinationen und typisierte Rede im mehrsprachigen Kontext*. [online] Available at: <http://www.leko-project.org> [Accessed 24 October 2015].

KONECNY, C., AUTELLI, E., ZANASI, L. AND ABEL, A., in prep. **Queste vacanze divantano il martello!* Interferenzen beim Gebrauch formelhafter Sequenzen im Italienischen seitens deutschsprachiger Südtiroler/innen und Möglichkeiten ihrer Klassifizierung. In: L. Zybatow et al., eds. *Proceedings des 50. Linguistischen Kolloquiums, Innsbruck, 03.-05.09.2015*. Frankfurt a.M. et al.: Lang.

KÜHN, P., 1987. Deutsch als Fremdsprache im phraseodidaktischen Dornröschenschlaf. Vorschläge für eine Neukonzeption phraseodidaktischer Hilfsmittel. *FLuL*, 16, pp.62-79.

KÜHN, P., 1992. Phraseodidaktik. Entwicklungen, Probleme und Überlegungen für den Muttersprachenunterricht und den Unterricht DaF. *FLuL*, 21, pp.169-189.

LORENZ-BOURJOT, M. AND LÜGER, H.-H. eds., 2001. *Phraseologie und Phraseodidaktik*. Wien: Praesens.

MÜLLER, C. AND STRUBE, M., 2006. Multi-Level Annotation of Linguistic Data with MMAX2. In: S. Braun, K. Kohn and J. Mukherjee, eds. *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt a.M. et al.: Lang. pp.197-214.

PAQUOT, M. AND GRANGER, S., 2012. Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, 32, pp.130-149.

SCHMID, H., 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK*. [online] Available at: <ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf> [Accessed 27 June 2015].

STEIN, S., 2007. Mündlichkeit und Schriftlichkeit aus phraseologischer Perspektive. In: H. Burger, D. Dobrovol'skij, P. Kühn and N.R. Norrick, eds. *Phraseology. An International Handbook of Contemporary Research*. Vol. 1. Berlin; New York: de Gruyter. pp.220-236.

STEYER, K., 2013. *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr.

SULKOWSKA, M., 2013. *De la phraséologie à la phraséodidactique. Études théoriques et pratiques*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

VENIER, F., 1996. I verbi sintagmatici. In: P. Blumenthal, G. Rovere and C. Schwarze, eds. *Lexikalische Analyse romanischer Sprachen*. Tübingen: Niemeyer. pp.149-156.

VETTORI, C. AND ABEL, A., 2012. Introduzione. In: A. Abel, C. Vettori and K. Wisniewski, eds. *Gli studenti altoatesini e la seconda lingua: indagine linguistica e psicosociale. / Die Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung*. Vol. 1 – Bd. 1. Bozen-Bolzano: EURAC. pp.9-12.

VOGHERA, M., 2004. Composizione: Polirematiche. In: M. Grossmann and F. Rainer, eds. *La formazione delle parole in italiano*. Tübingen: Niemeyer. pp.56-69.

WISNIEWSKI, K. AND ABEL, A., 2012. Die Sprachkompetenzenerhebung: Theorie, Methoden, Qualitätssicherung. In: A. Abel, C. Vettori and K. Wisniewski, eds. *Gli studenti altoatesini e la seconda lingua: indagine linguistica e psicosociale. / Die Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung*. Vol. 1 – Bd. 1. Bozen-Bolzano: EURAC. pp.13-64.

WISNIEWSKI, K., WOLDT, C., SCHÖNE, K., ABEL, A., BLASCHITZ, V., ŠTINDLOVÁ, B. AND VODIČKOVÁ, K., 2014. The MERLIN annotation scheme for the annotation of German, Italian and Czech learner language (report). [online] Available at: <http://www.merlin-platform.eu/docs/MERLIN-annotation-scheme.pdf> [Accessed 24 October 2015].

ZELDES, A., RITZ, J., LÜDELING, A. AND CHIARCOS, C., 2009. ANNIS: A search tool for multi-layer annotated corpora. In: *Proceedings of Corpus Linguistics, Liverpool, July 20-23*. [online] Available at: <http://ucrel.lancs.ac.uk/publications/cl2009/> [Accessed 26 June 2015].

ZIPSER, F. AND ROMARY, L., 2010. A model oriented approach to the mapping of annotation formats using standards. In: *Workshop on Language Resource and Language Technology Standards, LREC 2010, May 2010, La Valetta, Malta*. [online] Available at: <http://hal.archives-ouvertes.fr/inria-00527799/en/> [Accessed 27 June 2015].

CORPORA, THE WORLD WIDE WEB AND QUESTIONNAIRES AS SOURCES OF INFORMATION ON RECENT PHRASEOLOGICAL BORROWINGS: THE CASE STUDY OF THE POLISH UNIT *WYGLĄDAĆ JAK MILION DOLARÓW*

Joanna Szerszunowicz
Uniwersytet w Białymstoku
joannaszersz@gmail.com

Abstract

As a modern lingua franca, English is a donor of words and fixed expressions incorporated by other languages. One of them is Polish in which many anglicisms have been attested recently. An example of a loan phrase is *wyglądać jak milion dolarów* – a calque of the English idiom *to look like a million dollars*. As the borrowing is recent, the unit is not included in any lexicographic sources. Therefore, it is necessary to determine its status and characteristics. In order to provide a comprehensive monolingual lexicographic description, the information on the properties of the Polish phrase has to be collected. The aim of the paper is to show how various sources can be used in order to create a linguo-cultural presentation of the phrase at issue.

1. INTRODUCTION

Borrowing, one of the major aspects of language changes, is defined as "taking one word or phrase from one language into another, or from one variety of a language into another" (McArthur 1996: 137). Loans, among which anglicisms constitute an important group (cf. Görlach 2001, 2003; Furiasi, Pulcini, Rodríguez Gonzáles 2012; Mańczak-Wohlfeld 2010), are commonly found across languages. However, so far, most attention has been paid to lexical loans, while the phraseological borrowings seem to have been neglected. It is the recent years that have witnessed a noticeable increase in the studies on phraseological loans. The result of the works are two monographs (Veisbergs 2012, Fiedler

2014) and several papers (cf. those in Korhonen, Mieder, Pirainen and Piñel 2010; Fiedler 2011), mostly focussing on linguistic aspects of the phenomenon. Due to their nature, the recent borrowings are also worth discussing in a lexicographic perspective. The present paper aims to discuss this issue, with a focus on linguo-cultural aspects of the description of phraseological borrowings in a monolingual dictionary.

1.1. Aims

The general aim is to show how various sources can be used in order to create a linguo-cultural presentation of a recent loan phrase. The specific aims are threefold: to analyse *Narodowy Korpus Języka Polskiego* (NKJP), the World Wide Web and the results obtained from specially prepared questionnaires as sources of information on the units *wyglądać jak milion dolarów*; to evaluate their usefulnesss; to create a description of the loan phrase at issue.

1.2. Methodology

Three different sources of information on the expressions at issue were consulted to analyze the unit *wyglądać jak milion dolarów*. The first one is Narodowy Korpus Języka Polskiego, which contains 1.5 billion words. In the present study, the consultation paradigm was implemented, which allowed for extracting and analysing the actual behaviour of given phraseological units (Sailer 2007: 1064–1065; Steyer 2004; Veisbergs 2006).

The second source is The World Wide Web. It is consulted for two reasons: first, in general idioms are not so frequent in corpora (cf. Moon 2003), which is also the case with the analysed unit; second, a great variety of texts can be analysed on the Internet, allowing for searching kinds of texts not available in the consulted corpus (cf. Kilgariff 2001; Colson 2007). The WebCorp® tool was used for verification.

The third source was the information obtained from the questionnaires prepared for the research study. It was semi-structured: the respondents were asked to comment on the unit *wyglądać jak milion dolarów*, referring to certain points (meaning, stylistic markedness, typical contexts of use, frequency). Moreover, the respondents were supposed to evaluate the unit in terms of two aspects: their attitude to the unit and the inclusion of the phrase in the dcitionary. In the second part of the questionare, the respondents could comment freely on the unit.

After collecting the material from the three sources, it was analysed with a special focus on the characteristics of the unit in question. Then, the usefulness of the sources was evaluated by comparing given parametres of the description of the expression.

2. SOURCES OF INFORMATION ON THE UNIT *WYGLĄDAĆ JAK MILION DOLARÓW*

Since the unit *wyglądać jak milion dolarów* is a recent borrowing, the information on its meaning, stylistic markedness, collocability and other characteristics has to be collected. Possessing the information will enable creating a linguo-cultural description of the unit at issue, which is in line with the hybrid character shown by an increasing number of lexicographic works.

2.1. Narodowy Korpus Języka Polskiego and the World Wide Web

The first source consulted is *Narodowy Korpus Języka Polskiego*. In fact, as few as 4 results were obtained for the researched phrase. In fact, the unit used in the form *wygląda jak milion dolarów* expressed the meaning 'he earns a lot of money'. All the results come from press texts about Lew Rywin, Polish a film producer and a key figure in the corruption scandal in 2002. The results comply with other studies, which show that phraseological expressions tend to be underrepresented in corpora, especially if only one corpus is analysed (cf. Moon 2003: 97; Rodríguez Martín 2014). However, it can be concluded that the unit may be used to describe persons who have a special talent for earning money.

The search of the World Wide Web provided 50 results, which confirmed the assumption. The most important findings regarded the collocability of the simile in question. The analysis of the extracted examples shows that the unit was used to describe:

- person (both female and male, e.g. *Jennifer Lopez, Ricky Martin*; *kobieta* [a woman], *jego dziewczyna* [his girlfriend], *mężczyzna w smokingu* [a man in a dinner jacket] *każdy kto przechodzi koło mnie* [everybody who passes me]; *każdy* [everyone], wszyscy [all]),

- things (e.g. *ciasto marchewkowe* [carrot cake], *danie* [dish], *garnek* [pan], *detal* [detail], *miasto* [city], *samochód* [car] – usually makes: *Corvetta cabrio, Land Rover, Lexus RC F*),

- animals (*kot* [cat]),

- places (*miasto* [city]).

The above presentation indicates that apart from describing persons, the unit has a wide range of collocates, comprising both animate and inanimate entities. The preliminary search lead to the conclusion that the unit in question has big potential in terms of collocability and stylistc value. The latter is confirmed by its occurrence in a variety of genres, for example, newspaper articles, travelogues, recipes, advertisements etc. The unit at issue tends to be used as a phrase which captures the readers' attention, carries evaluation and enhances the persuasive character of a given text.

2.2. The results of questionnaires

The third source of information is the survey conducted in a group of 117 Polish philology students at the University of Bialystok in Poland. The participants were native speakers of Polish, aged 20-24. The questionnaire was composed of two parts: in the first part, respondents had to describe the meaning of the phraseological unit *wyglądać jak milion dolarów*, its stylistic markedness, collocability, use, frequency, origins as well as their attitude to the phrase and its possible inclusion in a monolingual dictionary; in the second, they could comment freely on the unit.

As for the meaning, more than half defined it as 'to look attractive' (62%) and 'to be well-dressed' (51%). Many respondents (37%) stressed positive change in somebody's appearance: 'to look better than usual'. They (36%) also emphasized the costs incurred to look good and specified the elements of good looks: 'to have expensive clothes and

jewellery as well as good hairdo and makeup'. Almost one fifth (17%) defined the meaning of the units in a general way as 'to look good' (17%).

In terms of stylistic markedness, the vast majority of them (78%) labeled the unit as informal. As for collocability**,** most respondents stated that the unit tends to be used to describe women (81%), with only 11% extending its use for men. Moreover, almost one third (27%) included celebrities. Many fewer listed expensive cars (6%) and places, espeically those perceived as luxurious holiday resorts (3%)

While discussing the use of the unit, some respondents (14%) perceived the phrase as a pragmatic idiom: as a compliment paid to somebody before going out, which emphasized the persons's looking better than usual. The following areas of use were given by the partcicpants of the survey: Internet genres (71%), press, especially tabloids and fashion magazines (49%), television programs (12%), informal discourse (43%). In fact, the results comply with those obtained from the World Wide Web search.

The respondents were also asked to evaluate the frequency of the unit. The assessment was subjective and it reflected their individual perception of this factor. More than half of them (57%) stated that the unit is used often or quite often (37%). As few as 6% thought that it is used rarely (6%). The results for very often (2 respondents) and never (1 person) were so low that they were not included.

The persons who participated in the survey were also asked to evaluate the unit in terms of its usefulness, attractiveness etc. The question aimed at determining how the phrase is evaluated by native Polish speakers and whether they consider it to enrich the language and fulfill their communicative needs. Their attitude was predominantly positive (61%), however, almost one fourth (23%) expressed a negative assessment of the unit, with 16% neutral answers. Those who accepted the phrase as a new Polish phraseologial unit stressed that it differs from the Polish synonyms and they liked its imagery. The opponents expressed their general reluctance to English borrowings and they did not like comparing a person to money.

Practically all respondents (97%) were aware of the fact that the unit is a borrowing from the English language. The high score can be the result of several factors, the most important of which is the presence of the component *dolary* 'dollars', associated with the USA, and the fact that the vast majority of students learn English. The exposure to a literal translation of the phrase in press and films can also have an influence on Polish users' perception of the phrase.

Finally, they had to take a stand on the possible inclusion of the unit *wyglądać jak milion dolarów* in monolingual dictionaries. A vast majority (72%) was of the opinion that the unit should be included. Almost one fifth of the survey participants (19%) did not give a definite answer. The negative answer was given by 9% of the respondents.

The free writing section offered a possibility to comment on the unit. Many answers contained the explanations of the responses from the first part of the survey, as illustrated by the statements presented below. According to the respondents, the phrase *wyglądać jak milion dolarów*:

- reflects American dream, American perception of life, values etc. (17%),
- has a different meaning from that of the Polish equivalent units (12%),
- should be used, because it sounds positive (9%),
- compares a good-looking person to money, which is surprining (8%),
- sounds good (8%),
- is often modified (7%),
- is difficult to define: more of evaluation than semantic contents (6%),
- sounds American (5%),
- enriches Polish phraseology (4%),
- sounds funny if used about a man (3%),
- should not be used, since there are enough Polish phraseological units expressing the same meaning (3%).

Apart from the comments elaborating on the responses from the first part of the questionnaire, there were some comments regarding the presence of the unit in texts of culture, such as a a song (2%).

The analysis of the responses given in this part gives an insight into the Polish language users' perception of this unit. It also shows which subjective perception of the unit's parametres differ from the results obtained from other sources. For instance, the frequency rate in the corpus analysis is much lower than the one given by the respondents. The analysis of the material excerpted from the World Wide Web shows that it is possible to use the phrase about men, with only 11% of the participants acknowledging such a use and 3% finding it unacceptable. Therefore, using all the sources provides many useful information and clues for lexicographers dealing with the monolingual description of recent borrowings.

3. A PROPOSAL OF AN ENTRY FOR *WYGLĄDAĆ JAK MILION DOLARÓW*

The phraseological unit *wyglądać jak milion dolarów* belongs to the Polish expressions which have come in use fairly recently. As already mentioned, it has not been included in Polish monolingual lexicographic works so far. It means that there is no possibility of referring to existent entries and critically reviewing them, which can be done for many other units. Therefore, the analysis of the cross-linguistic equivalents of the unit and the information from the three kinds of sources is necessary for the proposal of lexicographic description of the unit at issue.

3.1. Systematic equivalents

According to Dobrovol'skij (2011: 7-8), in a cross-linguistic perspective four main kinds of systematic equivalents can be distinguished: absolute, partial, parallel, zero. The Polish potential equivalents of the English unit *to look like a million dollars* represent two kinds: (quasi-)absolute (cf. Fiedler 2007: 118) and parallel. The Polish quasi-absolute equivalent *wyglądać jak milion dolarów* is marked with novelty, whereas the English unit is not. Therefore, it meets the criteria of this type: the semantic and formal structure as well as imagery are identical in both languages, with subtle differences in the stylistic value.

The parallel equivalents of the unit in question are numerous and they differ in many aspects. The imagery is not identical, since the expressions come from different domains, which is culturally conditioned. In Polish phraseology good-looking people, both men and women, are compared to:

• supernatural god-like creatures

wyglądać jak anioł [to look like an angel; usually about a woman: she looks subtle or she is dressed in white] (SP: 189),

wyglądać jak bóstwo [to look like a goddess; about a woman: she takes care of herself, is elegantly dressed and looks beautiful] (SP: 189),

• persons of high social standing

wyglądać jak panisko [to look like a lord; [he is well dressed and self-confident] (SP: 189),

wyglądać jak królewna [to look like a young queen; about a young woman: she looks beautiful] (SP: 189),

• characters from fairy tales

wyglądać jak księżniczka z bajki [to look like a duchess from a fairy tale; about a young woman: she looks beautiful] (SP: 189),

• the stereotyped perception of visual representations of good-looking persons or attractive things by mentioning the text, production etc. where they can be seen or read about

wyglądać jak z obrazka [to look like (one) from a picture; sb's appearance is very nice, but a little common] (SP: 190),

wyglądać jak z żurnala [to look like one from a fashion magazine; sb is dressed according to the lastest fashion] (SP: 198),

wygląda coś jak z bajki [to look like (something) from a fairy tale; about something which seems to be wonderful and too good to be real' (SP: 188),

wygląda coś jak z filmu [to look like (something) from a film; about something which resembles the film image of life] (SP: 188).

As it can be seen from the above presentation of the Polish equivalents, they tend to differ from the English expression. The differences can be observed in various aspects, for example, in the semantics: the meaning of the unit *wyglądać jak panisko* contains the additional component 'self-confident', while the phrase *wyglądać jak z obrazka* comprises the element 'a little common'. Moreover, the collocability of some units is not identical: some of the Polish similes can be used only with reference to a woman (*wyglądać jak anioł, wyglądać jak królewna*), while others – to a man (*wyglądać jak panisko*).

3.2. An entry for *wyglądać jak milion dolarów*

Taking into consideration the linguo-cultural character of borrowing, the lexicographic description should reflect its nature (cf. Szerszunowicz 2011). The proposed microstructure is composed of the following elements: the canonical form of the unit, stylistic label, variant form, explanation of the meaning, collocates, typical areas of use, examples prepared by the lexicographers, origin, antonyms, synonyms, authentic examples / links to authentic examples, visual material illustrating the canonical use of the unit and its creative modifications. The new technological solutions create the possibilities of the inclusion of more rich and varied illustrative material, which is of great importance, especially in terms of the inclusion of cultural component (cf. ACD).

The proposed lexicographic description allows for the inclusion of linguo-cultural information, which provides a comprehensive picture of the unit and its communicative potential. The entry for the unit *wyglądać jak milion dolarów*, which contains the all the elements, is presented below.

wyglądać jak milion dolarów *pot.*

wyglądać jak milion dolców

doskonale wyglądać (o osobie lub rzeczy); nosić drogie ubrania i biżuterię, być zadbanym; wyglądać lepiej niż zazwyczaj [EXPLANATION]

wyrażenie często używane z następującymi nazwami osób i przedmiotów:
kobieta, dziewczyna i określenia synonimiczne (również imię i nazwisko), samochód, towary luksusowe (ubranie, jacht, meble itp.) [COLLOCATES]

często używane w języku mediów: Internetu, telewizji, prasy, zwłaszcza w prasie brukowej i w magazynach poświęconych modzie [TYPICAL AREAS OF USE]

Po wizycie u fryzjera i kosmetyczki Iwona wyglądała jak milion dolarów.
Nowy model Mazdy wygląda jak milion dolarów. [EXAMPLES]

Pochodzenie: Frazeologizm ten został zapożyczony stosunkowo niedawno z języka angielskiego (*to look like a million dollars*). W polszczyźnie porównania odnoszące się do korzystnego wyglądu motywowane były innymi wyobrażeniami, na przykład wiele z nich to jednostki w których występują nazwy istot nadprzyrodzonych. W języku polskim używane są również jednostki odwołujące się do stereotypowych wyobrażeń (wygląd modelki w żurnalu mody). W angielszczyźnie frazeologizmy z domeny finansów mogą opisywać człowieka: jego wygląd (*to look like a million*

dollars) czy samopoczucie (*to feel like two cents*, dosł. czuć się jak dwa centy 'czuć się marnie'). Różnica w sposobie wyrażania znaczenia 'dobrze wyglądać' ma uwarunkowania kulturowe: pośrednio nawiązuje do amerykańskiego marzenia i przyjętego w kulturze amerykańskiej systemu wartości. [LINGUO-CULTURAL INFORMATION INCLUDING ORIGIN]

Synonimy: *wyglądać jak anioł, wyglądać jak bóstwo, wyglądać jak panisko, wyglądać jak księżniczka z bajki, wyglądać jak królewna, wyglądać jak z obrazka* [SYNONYMS]

Antonimy: *wyglądać jak siódme dziecko stróża, wyglądać jak siedem nieszczęść, wyglądać jak śmierć na chorągwi* [ANTONYMS]

Przykładowe użycia [EXAMPLES OF ACTUAL USE]

http://teleshow.wp.pl/img,16874909,gid,16874891,title,Jennifer-Lopez-wyglada-jak-milion-dolarow,tpl,6,galeria.html?ticaid=1151de

http://www.radiozet.pl/Radio/Programy/ZET-za-kolkiem/ZET-za-kolkiem-testuje-Land-Rovera-Discovery-Sport-00007181

http://www.czytajniepytaj.pl/lifestyle/kuchnia/ciasto-marchewkowe-ktore-wyglada-jak-milion-dolarow-2503.html

The examples made by the authors of the dictionary help to illustrate the typical uses of the unit, while the inclusion of the authentic occurrences makes it possible to show real use, often more complex as a result of modifications of the phrase.

It can be concluded that the analysis of the three sources on the information on the phrasological borrowings are useful in terms of the analysis of the unit. The data obtained from the corpus and the World Wide Web provide examples of the actual use of the phraseological unit in question, while the analysis of the results of the survey shows how the respondent, native speakers of Polish, perceive this expression.

4. CONCLUSIONS

Since borrowing in a natural process occurring in language contacts, loan words and phrases belong to the phenomena, which should be analysed in a linguo-cultural perspective. In the contemporary world, the borrowings from English are very common and they are by no means limited to lexical items. English phraseological borrowings are also common due to many reasons, like cultural changes and media influence.

As important vehicles of communication, the recent loan phrases have to be included in monolingual dictionaries. Their inclusion involves defining the meaning, determining the norms regarding the use of a given unit and presenting the cultural information on the described phrase. Therefore, in order to provide a lexicographic description of recent borrowings, a multiaspectual analysis of the units should be conducted.

The inclusion of multiple sources of information contributes to creating a comprehensive description of the units at issue. The first source in the corpus providing a

substantial amount of varied language material. However, as many phraseological units are underrepresented in corpora, the second source is the World Wide Web, which may bring more results for a given phrase. The third source of information is the survey conducted in a group of native speakers of a given language: the proposed questionnaire includes the key information covered by a typical entry, extended with a question regarding the inclusion of the unit in a dictionary, and a free writing section, which allows for obtaining additional information.

The research study done for the Polish unit *wyglądać jak milion dolarów* shows that the sources brought different kind and amount of information. The corpus did not provide much data due to a small number of occurrences, while the World Wide Web was a source of useful information on the typical contexts of use and the collocability of the analysed unit. The questionnaires gave an insight into the Polish language users' perception of the unit. To sum up, it should be emphasized that consulting these three sources of information is commendable, since the information obtained from each of them differs and complements one another.

References

ACD: *Australian Cultural Dictionary*. [online] Available at http://www.culturaldictionary.org/ [Accessed 2 June 2015].

COLSON, J.-P., 2007. The World Wide Web as a corpus for set phrases. In: H. Burger, D. DOBROVOL'SKIJ, P. KÜHN, and N.R. NORRICK, eds. 2007. *Phraseologie. Phraseology. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of Contemporary Research*, Vol. 1. Berlin and New York: Walter de Gruyter. pp. 1071-1077.

DOBROVOL'SKIJ, D., 2011. Cross-Linguistic Equivalence of Idioms: Does It Really Exist?. In: Pamies and D. Dobrovol'skij, eds., 2011. *Linguo-Cultural Competence and Phraseological Motivation*. Baltmannsweiler: Schneider Verlag. pp. 7-24.

FIEDLER, S., 2007. *English Phraseology. A Coursebook*. Tübingen: Narr Verlag.

FIEDLER, S., 2011. The sky is the limit – the influence of English on German phraseology. In: J. Szerszunowicz, B. Nowowiejski, K. Yagi,, and T. anzaki, eds., 2011. *Research on Phraseology in Europe and Asia: Focal Issues of Phraseological Studies*. Białystok: Wydawnictwo Uniwersytetu w Białymstoku. pp.269-290.

FIEDLER, S., 2014. Gläserne Decke *und* Elefant im Raum — *Phraseologische Anglizismen im Deutschen*. Berlin: Logos Verlag.

FURIASI, C., PULCINI, V., AND RODRÍGUES GONZÁLES, F., eds., 2012. *The Anglicization of European Lexis*. Amsterdam and Philadelphia: John Benjamins.

GÖRLACH, M., ed., 2001. *A Dictionary of European Anglicisms*. Oxford: Oxofrd University Press.

GÖRLACH, M., 2003. *English Words Abroad*. Amsterdam: John Benjamins.

KILGARIFF, A., 2001. Web as Corpus. In: G. Sampson, and D. McCarthy, eds., 2004. *Corpus Linguistics: Readings in a Widening Discipline*. London & New York: Continuum. pp. 471-473.

KORHONEN J., MIEDER W., PIIRAINEN E., PIÑEL R., eds., 2010. *Phraseologie global – areal – regional. Akten der Konferenz EUROPHRAS 2008 vom 13.-16.8.2008 in Helsinki*. Tübingen: Narr Verlag.

MAŃCZAK-WOHLFELD E., ed., 2010. *Słownik zapożyczeń angielskich w polszczyźnie* [Dictionary of English borrowings in the Polish language]. Warszawa: Wydawnictwo Naukowe PWN.

MCARTHUR T., ed., 2003. *An Oxford Companion to the English Language*. Oxford: Oxford University Press.

MOON, R., 2003 [1998]. *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford: Oxford University Press.

NKJP: *Narodowy Korpus Języka Polskiego*. [online] Available at: http://www.nkjp.uni.lodz.pl [Accessed 28 April 2015].

RODRÍGUEZ MÁRTIN G.A.., 2014. Canonical modified phraseologial units: Analysis of the Paradox. *Yearbook of Phraseology*, 5, pp. 3-23.

SAILER, M., 2007. Corpus linguistic approaches with German corpora. In: H. Burger, D. Dobrovol'skij, P. Kühn, and N.R. Norrick, eds., 2007. *Phraseologie. Phraseology. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of Contemporary Research*, Vol. 1. Berlin and New York: Walter de Gruyter. pp. 1060-1071.

SP: BAŃKO, M., 2004. *Słownik porównań* [A dictionary of similes]. Warszawa: Wydawnictwo Naukowe PWN.]

STEYER, K., 2004. Kookurenz. Korpusmethodik, linguistisches Modell, lexikographische Perspektiven. In: K. Steyer, Hrsg., 2004. *Wortverbindungen – mehr oder weniger fest*. Berlin and New York: Walter de Gruyter. pp. 87-116.

SZERSZUNOWICZ, J., 2011. The cultural component in bilingual dictionaries of phraseological units. In: K. Akasu and S. Uchida, eds., 2011. *ASIALEX2011 Proceedings. LEXICOGRAPHY: Theoretical and Practical Perspectives. Papers submitted to the Seventh ASIALEX Biennial International Conference Kyoto Terrsa, Kyoto, Japan, August 22-24, 2011*. Tokyo: ASIALEX, pp. 628-637.

VEISBERGS, A., 2006. Kookurenz. Korpusmethodik, linguistisches Modell, lexikographische Perspektiven. In: K. Steyer, Hrsg., 2006. *Wortverbindungen – mehr oder weniger fest*. Berlin and New York: Walter de Gruyter. pp. 87-116.

VEISBERGS, A., 2012. *Phraseological Borrowings*. Logos: Berlin.

ESTUDIO FRASEOLÓGICO BASADO EN EL CORPUS CORBICON

Arsenio Andrades
CES Felipe II (UCM)
arsenioa@ucm.es

Resumen

El objeto de este artículo es presentar una muestra del estudio fraseológico basado en la compilación de un corpus bilingüe inglés-español de textos jurídicos (CORBICON) con la finalidad de identificar y clasificar las expresiones fraseológicas más comunes. Los estudios de fraseología del lenguaje general distinguen diferentes tipos de estructuras fraseológicas (colocaciones, locuciones, binomios y expresiones formulaicas) pero son escasas las investigaciones dedicadas a poner de manifiesto las características fraseológicas específicas de los textos jurídicos. La propuesta de una taxonomía fraseológica centrada en el ámbito jurídico constituye una valiosa herramienta de reflexión para plantear estrategias aplicables a la traducción, la búsqueda de equivalencias, la redacción de textos jurídicos y la enseñanza de lenguajes de especialidad. En este sentido el corpus nos proporciona información que no se encuentra en los diccionarios ya que, como ocurre con la fraseología del lenguaje general, su presencia sigue siendo exigua en repertorios bilingües, glosarios terminológicos o recursos lexicográficos.

1. INTRODUCCIÓN

A pesar de que la fraseología es uno de los ámbitos de la lingüística que más dificultades plantea desde el punto de vista de la traducción se le ha prestado poca atención hasta fecha muy reciente (Roberts, 1998: 61). Esta joven disciplina, que estudia las estructuras lingüísticas prefabricadas, ha experimentado un desarrollo espectacular en los últimos años gracias a los avances de la informática que han hecho posible la compilación de enormes corpus digitales. Así, la masiva difusión de internet que da acceso a todo tipo de textos electrónicos y la creación de herramientas informáticas que permiten analizar exhaustivamente cualquier fenómeno lingüístico han contribuido a multiplicar de manera exponencial el uso de la lingüística de corpus para todo tipo de investigaciones relacionados con la lengua, la lingüística y la traducción.

El presente estudio pretende mostrar unos primeros datos orientativos sobre los principales rasgos fraseológicos de los contratos de derecho civil basándose en la lingüística de corpus desde una perspectiva contrastiva con vistas al establecimiento de equivalencias en una fase posterior.

2. EL CORPUS CORBICON

El presente estudio se basa en un corpus (CORBICON) específicamente compilado sobre la base de una serie de criterios de diseño (tamaño, formato, representatividad, homogeneidad, número de palabras, codificación) con el fin de analizar sus características fraseológicas. Corpas (2001: 160) subraya la importancia del "análisis detallado de corpus multilingües como un paso necesario para la identificación de aquellos patrones recurrentes (sintáctico, semántico y colocacional) que subyacen a la selección de un determinado equivalente de traducción".

Se trata de un corpus especializado y homogéneo puesto que está compuesto por formularios legales de la rama del derecho civil. El corpus es bilingüe y comparable, es decir, está formado por dos subcorpus de textos originales en inglés y español que suman un millón de palabras. Los documentos digitales proceden en su mayor parte de las numerosas páginas web especializadas en la comercialización y difusión de documentos e información jurídicos (Andrades, 2013c), así como de algunos CD que acompañan los libros de modelos de contratos jurídicos.

2.1. Explotación del corpus

El análisis del corpus se basa en la identificación de elementos significativos desde un punto de vista estadístico y su cotejo con las distintas clasificaciones fraseológicas mencionadas en el presente artículo. Para ello recurrimos al programa de concordancias WordSmith Tools (versión 5.0) con el que vamos a generar, en primer lugar, una lista de palabras ordenadas por frecuencia de aparición tanto del subcorpus inglés como del español. A partir de las palabras más frecuentes obtenidas, y mediante la función Concord, realizamos búsquedas de agrupaciones de palabras o *cluster* en ambos subcorpus. Esta función nos permite observar el conjunto de ocurrencias de un término o grupo de palabras en su contexto de producción natural y nos da a conocer sus combinaciones o colocaciones más frecuentes.

El enfoque metodológico que adoptamos consiste en una combinación de las aproximaciones *bottom-up* y *top-down* (Granger & Paquot, 2008), esto es, se parte, en primer lugar, del análisis del CORBICON mediante el programa de concordancias WordSmith Tools, que arroja una serie de datos que nos permite observar muestras reales de los textos y extraer datos estadísticos relevantes. Posteriormente estos datos se contrastan con los patrones fraseológicos seleccionados sobre la base de las clasificaciones fraseológicas preseleccionadas. En definitiva, sendos enfoques el deductivo (*bottom-up*) y el inductivo (*top-down*) se complementan puesto que la frecuencia de ocurrencia en el corpus condiciona la detección de los candidatos a UFE y la identificación como UFE depende de su grado de correspondencia con las taxonomías fraseológicas mencionadas anteriormente.

3. FRASEOLOGÍA ESPECIALIZADA

Al tratarse de un contexto lingüístico especializado, es preciso distinguir entre fraseología general y fraseología jurídica. Lorente (2002: 176-178) caracteriza las unidades fraseológicas especializadas (UFE) como combinaciones de palabras con un cierto grado de fijación, que contienen como mínimo un término, presentan una mayor tendencia a la denotación, se localizan prioritariamente en textos escritos y se identifican con una temática concreta o una comunidad académica específica. Aunque el término UFE engloba distintas clases de combinaciones de palabras pertenecientes al lenguaje especializado, estas tienen en común, en mayor o menor grado, una serie de propiedades como son la presencia de una unidad terminológica, la polilexicalidad, la frecuencia de aparición en textos

especializados, la institucionalización, la estabilidad sintáctica y semántica, la idiomaticidad y la variación.

El interés por la fraseología especializada surge como consecuencia de las necesidades de traductores, redactores técnicos o periodistas que han de reproducir en sus traducciones o en sus documentos la fraseología propia del discurso especializado. Existe cada vez una mayor conciencia sobre el hecho de que el conocimiento no solo se transmite por medio de la terminología sino que la fraseología vehicula una importante carga semántica y formal que conviene conocer. Aguado de Cea (2007a: 53) subraya que el dominio de estos frasemas o unidades fraseológicas, también denominadas *solidaridades léxicas* y *colocaciones*, es el factor que confiere naturalidad al discurso especializado. Su empleo adecuado podrá ser un indicador del dominio del lenguaje de especialidad y del grado de competencia comunicativa del usuario. Tabares y Batista (2012) insisten asimismo en la importancia de que el traductor conozca la fraseología propia de los ordenamientos jurídicos tanto de la lengua de partida como de llegada. En este sentido, Cabré (1998: 52) señala que la "fraseología especializada, que no suele figurar en las obras terminológicas clásicas, constituye un elemento clave de consulta para asegurar la adecuación, precisión y naturalidad de una traducción".

3.1. Taxonomías fraseológicas

No existe unanimidad con respecto a las taxonomías fraseológicas. Las unidades fraseológicas (UF) se ordenan en diversas clasificaciones que varían en función de los criterios aplicados por los distintos estudiosos. Entre las principales taxonomías que nos interesan se encuentran las de Corpas (1997), Benson et al. (1986), Gläser (1994-95), Ruiz Gurillo (1998) y García-Page (2008). A la hora de seleccionar los tipos de expresiones fraseológicas que vamos a estudiar en nuestro corpus tendremos en cuenta dos factores: por un lado, las clasificaciones fraseológicas existentes y, por otro, el tipo de UFE con mayor frecuencia de aparición en el CORBICON.

En lengua española, una de las clasificaciones más completas y rigurosas es la propuesta por Corpas (1997) en la que distingue tres esferas: las colocaciones, las locuciones y los enunciados fraseológicos. Además, según Montero (2002: 140) es posible establecer una clasificación común al sistema fraseológico de la lengua inglesa y española por el alto grado de semejanza existente en cuanto a los aspectos formales, semánticos y pragmáticos. Así pues, dado el elevado nivel de paralelismo entre las dos lenguas del corpus nos basaremos en la taxonomía fraseológica de Corpas con el fin de comprobar qué rasgos de dicha clasificación son aplicables al CORBICON.

No obstante, las propiedades intrínsecas del lenguaje jurídico requieren algunos ajustes que consideramos necesarios para poner de relieve las características fraseológicas más significativas de este lenguaje especializado. En este sentido en nuestra propuesta taxonómica agregamos una categoría dedicada a los binomios y reducimos los distintos tipos de locuciones únicamente a las locuciones prepositivas. A continuación se presentan los principales tipos fraseológicos identificados en el corpus y se exponen las razones teóricas que han motivado la siguiente propuesta de clasificación en: binomios, colocaciones, locuciones prepositivas y fórmulas estereotipadas.

3.1.1. Binomios

Una de las características del lenguaje jurídico más destacadas por la mayoría de los estudiosos del discurso jurídico anglosajón (Mellinkoff, 1963; Bhatia, 1994; Tiersma, 1999; Alcaraz, 1994; Borja, 2000) es la conspicua ocurrencia de unas estructuras lingüísticas conformadas por una secuencia de dos palabras de la misma categoría gramatical unidas

por una conjunción coordinante y que mantienen una relación semántica (Gustafsson, 1984). La propia Gustafsson afirma que la presencia de binomios es cinco veces superior en el discurso jurídico que en otros géneros; por ello, a pesar de que Bosque y Demonte (1999: 620) y Corpas (1997) encuadren estas estructuras dentro de la categoría de locuciones, consideramos que la abultada presencia de binomios en el lenguaje jurídico justifica un tratamiento específico y un apartado exclusivo. El origen, la presencia y el mantenimiento de estas estructuras binómicas se deben a una serie de factores históricos, lingüísticos, jurídicos, estilísticos, literarios y económicos que han incidido de manera decisiva en algún punto del proceso de formación del lenguaje jurídico anglosajón y le han conferido las características que le conocemos en la actualidad (Andrades, 2013b).

Estas expresiones binómicas reciben distintas denominaciones dependiendo de los estudiosos: expresiones binómicas, dobletes, redundancia expresiva, binomios irreversibles, binomios coordinativos, dobletes sinonímicos (Andrades, 2013a: 77-78), y su estructura suele consistir en dos palabras unidas por una conjunción coordinante copulativa (*y/e,* en español; *and,* en inglés) o disyuntiva (*o/u*, en español; *or*, en inglés).

3.1.2. Colocaciones

Las colocaciones son grupos de palabras que aparecen juntas o en un mismo contexto con cierta frecuencia y suelen estar compuestas por un elemento determinante o *base* y un elemento determinado o *colocado*. Benson et al. (1986) consideran que las colocaciones están a medio camino entre las combinaciones libres y las locuciones o expresiones idiomáticas. Desde el punto de vista de la fijación, se trata de grupos de palabras que aparecen juntas habitualmente y que presentan cierta estabilidad sin llegar al nivel de lexicalización de las locuciones.

La clasificación de las colocaciones léxicas simples de Corpas (1997: 66-76), que se basa parcialmente en la de Benson et al. (1986), es una de las más completas; Koike (2001) propone ligeros cambios que suponen una reordenación de ciertas categorías que no afectan sustancialmente a la propuesta de Corpas, con la excepción de la adición de un tipo de colocación (Verbo + Adjetivo) que hemos considerado pertinente agregar al presente trabajo:

1) Sustantivo (sujeto) + Verbo: *estallar una revuelta*
2) Verbo + Sustantivo (objeto): *dictar una sentencia*;
Verbo + Preposición + Sustantivo: *poner a prueba*
3) Sustantivo + Adjetivo/sustantivo: *soltero empedernido; viaje relámpago*
4) Sustantivo + Preposición + Sustantivo: *racimo de uvas*
5) Verbo + Adverbio: *afirmar rotundamente*
6) Adjetivo + Adverbio: *estrechamente relacionado*
7) Verbo + Adjetivo: *resultar ileso*

3.1.3. Fórmulas estereotipadas

En esta categoría se incluyen las UF que representan unidades más cercanas a fórmulas o frases propias de determinados ámbitos discursivos. En el lenguaje general, las expresiones estereotipas son fórmulas que se utilizan para expresar saludo, agradecimiento, disculpas, algún tipo de cortesía, es decir, son fórmulas protocolorias como, por ejemplo: *¿Qué tal?*, *buenos días*, *¡vaya por Dios!*. Este grupo se corresponde con la tercera esfera propuesta por Corpas (1997) en su clasificación fraseológica ya que se trata de expresiones que se utilizan en contextos profesionales o discursivos muy concretos. A veces, como señala Corpas (1997: 133) resulta difícil distinguir los enunciados fraseológicos de las

locuciones. En este grupo se consideran fraseológicas unidades equivalentes a sintagmas, así como unidades más amplias, a las que se denominan expresiones formulaicas.

En general, las fórmulas no suelen tener una gran importancia presencial en el discurso especializado (Aguado, 2007b: 185), pero el ámbito jurídico administrativo es una excepción. Roberts (1994-1995) observa que en determinados discursos especializados como el administrativo es frecuente el empleo de expresiones fijas semejantes a fórmulas, a las que denomina expresiones formulaicas (*formulaic expressions*) como, por ejemplo, *veuillez agréer l'expression de mes sentiments distingués*, o *pour faire suite à*. Monzó y Hoyo (1998) subrayan asimismo la importancia de las fórmulas jurídicas como características de la fraseología.

3.1.4. Locuciones prepositivas

Las locuciones consisten en combinaciones de dos o más palabras con un elevado grado de idiomaticidad, es decir, generalmente su significado no se puede deducir de los elementos que las componen. En cuanto a la distinción conceptual entre las locuciones y las colocaciones, Corpas (1997: 269), en su clasificación, propone el término colocación cuando se trata de una estructura fijada por la norma y locución cuando está fijada por la lengua. Ruiz Gurillo (2000) explica que la diferencia entre las colocaciones y las locuciones radica en que las primeras no son idiomáticas y las segundas sí.

Aunque existen tantos tipos de locuciones como categorías gramaticales (locuciones nominales: *cabeza de turco*; locuciones adjetivas: *corriente y moliente*; locuciones verbales; *meter la pata*; locuciones adverbiales: *a la fuerza*; locuciones preposicionales: *a falta de*; locuciones conjuntivas: *de manera que*, etc.), nos vamos a centrar en las locuciones prepositivas que tienen una especial incidencia en el lenguaje jurídico, tal y como señalan Alcaraz (2002b: 26), De Miguel (2000) y Pontrandolfo (2013: 193): *a los efectos de, a instancia de, en materia de, en virtud de, a tenor de, de acuerdo con*, etc.

4. MUESTRAS DE FRASEOLOGÍA JURÍDICA

A continuación presentamos las primeras muestras del análisis del corpus que confirman la presencia fraseológica que intuíamos y su mayor o menor grado de coincidencia con las características de las tipologías taxonómicas mencionadas. No pretendemos ofrecer datos estadísticos sobre la representatividad de las UFE que identificamos en el CORBICON sino que nuestro objetivo consiste en constatar la presencia de este tipo de unidades, especificar el número de ocurrencias en el corpus (véase frec. en las tablas) verificar su adscripción a una tipología fraseológica e ilustrar con ejemplos un primer muestreo que se ampliará y desarrollará en posteriores estudios.

4.1. Binomios

Categoría gramatical	Binomios	Frec.	Ejemplos del corpus **(EN)**
Sustantivo	*terms and conditions*	274	*[…] delivery and acceptance of the aforesaid Property upon the* ***terms and conditions*** *of this lease.*
Verbo	*represent and warrant*	65	*The individuals executing this Lease on Landlord's behalf* ***represent and warrant*** *[…]*
Preposición	*by and between*	162	*It is mutually agreed* ***by and between*** *Owner and Tenant that the respective parties […]*
Adverbio	*now or hereafter*	58	*Any and all other rights of every and any nature* ***now or hereafter*** *existing […]*

Adjetivo	*due and payable*	129	*[…] any rent or other amount herein provided within ten (10) days after the same is* ***due and payable****, […]*

Tabla 1. Muestra de binomios del subcorpus inglés

Categoría gramatical	Binomios	Frec.	Ejemplos del corpus **(ES)**
Sustantivo	*cargas y gravámenes*	236	*El establecimiento se encuentra libre de* ***cargas y gravámenes****, y al corriente del […]*
Verbo	*otorgan y firman*	80	*Y en prueba de conformidad* ***otorgan y firman*** *el presente contrato en el lugar y fecha […]*
Preposición	-------	----	----------
Adverbio	*mutua y recíprocamente*	66	*Las partes se reconocen* ***mutua y recíprocamente*** *capacidad legal para el […]*
Adjetivo	*libres y espontáneas*	124	*Las Partes, de sus* ***libres y espontáneas*** *voluntades, manifiestan tener y […]*

Tabla 2. Muestra de binomios del subcorpus español

4.2. Colocaciones

Tipo de colocación	Colocación	Frec.	Ejemplos del corpus **(EN)**
Sust. (sujeto) + Verbo	*tenant agrees*	347	*The* ***Tenant agrees*** *that no more than 2 residents are allowed to […]*
Verbo + Sust. (objeto)	*surrender the premises*	53	*Tenant shall* ***surrender the Premises*** *in as good a state and condition […]*
V + prep. + Sust.	*remain in possession*	41	*If Tenant* ***remains in possession*** *of the Premises with the consent […]*
Adj. + Sust.	*due execution*	54	*Upon the* ***due execution*** *of this Agreement, Tenant shall deposit […]*
Sust. + Prep. + Sust.	*right of entry*	41	*The* ***right of entry*** *shall likewise exist for the purpose of removing […]*
Verb. + Adv.	*terminate immediately*	66	*[…] Agreement shall, at Landlord's option,* ***terminate immediately*** *[…]*
Adj. + Adv.	*immediately due*	60	*[…] balance and earned interest on this Note* ***immediately due****.*
Verb. + Adj.	*held liable*	17	*[…] Tenant may be* ***held liable*** *for the balance of the unpaid rent […]*

Tabla 3. Muestra de colocaciones del subcorpus inglés

Tipo de colocación	Colocación	Frec.	Ejemplos del corpus **(ES)**
Sust. (sujeto) + Verbo	*el contrato se rige*	110	*El presente contrato* ***se regirá*** *por lo previsto en la Ley […]*
Verbo + Sust. (objeto)	*otorgar el contrato*	151	*Que habiendo convenido* ***otorgar el contrato*** *de arrendamiento […]*
V + prep. + Sust.	*ser de cuenta*	412	***Será de cuenta*** *de la parte compradora […]*
Adj. + Sust.	*persona jurídica*	140	*[…] entidad/sociedad/* ***persona jurídica****) […]*
Sust. + Prep. + Sust.	*derecho de adquisición*	95	*[…] arrendatario dispondrá del* ***derecho de adquisición*** *[…]*
Verb. + Adv.	*se reconocen mutuamente*	126	*[…] partes tienen y* ***se reconocen mutuamente*** *plena capacidad […]*
Adj. + Adv.	*inmediatamente anterior*	92	*[…] recibo del pago del Precio* ***inmediatamente anterior*** *al […]*

Verb. + Adj.	*se halla representado*	125	***Se halla representada** en este acto por Don/Doña […]*

Tabla 4. Muestra de colocaciones del subcorpus español

4.3. Fórmulas estereotipadas

Fórmula estereotipada	Frec.	Ejemplos del corpus **(ES)**
conforme a derecho	59	*[…] saneamiento de la cosa subarrendada **conforme a Derecho**.*
en prueba de conformidad/en prueba de su conformidad	145	***Y en prueba de conformidad** con cuanto antecede, ambas partes […]*

Tabla 5. Muestra de fórmulas estereotipadas del subcorpus español

Fórmula estereotipada	Frec.	Ejemplos del corpus **(ES)**
by statute	19	*[…] or materially fails to comply with any duties imposed on Tenant **by statute**, within seven (7) […]*
In witness whereof	86	***IN WITNESS WHEREOF**, the parties have hereunto set their hands the day and year first […]*

Tabla 6. Muestra de fórmulas estereotipadas del subcorpus inglés

4.4. Locuciones prepositivas

Locuciones prepositivas	Frec.	Ejemplos del corpus **(ES)**
en (el) caso de	396	***En caso de** abonar directamente al arrendador la renta […]*
en virtud de	290	*[…] gravan la titularidad de la finca vendida **en virtud de** este contrato.*
Locuciones prepositivas	Frec.	Ejemplos del corpus **(EN)**
pursuant to	256	*However, this amount is subject to increase **pursuant to** the results of the energy audit, which audit shall be […]*
in the event of	158	***In the event of** any dispute subject to this provision, either party may initiate a request for mediation […]*

Tabla 7. Muestra de locuciones prepositivas del corpus

5. CONCLUSIONES

En primer lugar cabe señalar que tanto la taxonomía propuesta como los datos expuestos son los primeros resultados de la fase inicial de un estudio y, por ende, son meramente orientativos. Solo sirven para indicar una tendencia que habrá de verificarse en posteriores análisis más exhaustivos y rigurosos.

En todo caso, la explotación del corpus nos permite presentar un panorama general de los principales tipos fraseológicos existentes en el CORBICON y corrobora cierta correlación entre la fraseología general y la jurídica. La identificación de un número importante de ejemplos de cada uno de los tipos fraseológicos mencionados confirma asimismo la pertinencia de nuestra propuesta taxonómica que, naturalmente, requiere desarrollarse y contrastarse con datos cuantitativos y cualitativos de mayor envergadura.

El hecho de que se repitan patrones fraseológicos similares en el par de lenguas del CORBICON abre la posibilidad a que se puedan establecer equivalencias entre determinadas UFE de ambos subcorpus. Este tipo de investigaciones contribuye a dar a conocer las combinaciones léxicas más frecuentes desde una perspectiva contrastiva y a

concienciar sobre la importancia de adquirir una competencia fraseológica en el ámbito jurídico. Los datos obtenidos pueden aportar información práctica para traductores especializados, redactores técnicos, lexicógrafos o profesores de lenguas con fines específicos que trabajan en el campo jurídico. Como afirma Corpas (2001), el reconocimiento e identificación de la fraseología es el primer paso imprescindible para hacer una valoración de su carga semántica y proceder posteriormente al establecimiento de correspondencias interlingüísticas entre los idiomas en cuestión.

En definitiva, el presente estudio pretende simplemente abrir una vía de investigación en el ámbito de la fraseología jurídica con la intención de que en posteriores estudios se complete y se concrete la propuesta taxonómica, se analicen las propiedades de las UFE con precisión y se pueda verificar si los primeros resultados obtenidos son extrapolables a otros géneros del campo jurídico. En caso de confirmarse los resultados obtenidos en esta fase preliminar se podría elaborar una clasificación específica de UFE aplicable al lenguaje jurídico. Este estudio constituye un punto de partida desde el que poder ampliar y fomentar la investigación de las UFE en el ámbito jurídico y plantear una perspectiva enfocada a la búsqueda de equivalencias en el par de lenguas inglés-español.

Bibliografía

AGUADO DE CEA, G., 2007a. La fraseología en las lenguas de especialidad. In: E. Alcaraz, J. M. Martínez and F. Yus Ramos, eds. *Las lenguas profesionales y académicas*. Barcelona: Ariel. pp.53-65.

AGUADO DE CEA, G., 2007b. A multiperspective approach to specialized phraseology: Internet as a reference corpus for phraseology. In: S. Posteguillo, M. J. Esteve and M. L. Gea-Valor, eds. *The Texture of Internet: Netlinguistics in Progress*. Cambridge: Cambridge Scholars Publishing. pp.182-207.

ALCARAZ, E., 1994. *El inglés jurídico. Textos y documentos*. Barcelona: Ariel.

ALCARAZ, E., 2000. *El inglés profesional y académico*. Madrid: Alianza Editorial.

ALCARAZ, E., CAMPOS, M. A. AND MIGUÉLEZ, C., 2001. *El inglés jurídico norteamericano*. Barcelona: Ariel.

ALCARAZ, E. AND HUGHES, B., 2002a. *Diccionario Bilingüe de Términos Jurídicos: inglés-español, español-inglés*. Barcelona: Ariel.

ALCARAZ, E. AND HUGHES, B., 2002b. *El español jurídico*. Barcelona: Ariel.

ANDRADES, A., 2013a. *Estudio contrastivo de las unidades fraseológicas especializadas (UFE) en un corpus comparable bilingüe de documentos jurídicos vinculantes en lengua inglesa y española*. Tesis doctoral inédita. Madrid: Universidad Complutense de Madrid.

ANDRADES, A., 2013b. La importancia de los binomios en la traducción jurídica. In: E. Ortega Arjonilla, dir. *Translating Culture/Traduire la Culture/Traducir la cultura*. Vol. 3. Granada: Comares. pp.401-413.

ANDRADES, A., 2013c. Internet como fuente para la compilación de corpus jurídicos. *Enlaces*, [e-journal] 15. Available through: Revista del CES Felipe II < http://www.cesfelipesegundo.com/revista/INTRO.HTML> [Accessed 12 June 2015].

BENSON, M., BENSON, E. AND ILSON, R., 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Amsterdam: John Benjamins.

BHATIA, V., 1994. Cognitive structuring in legislative provisions. In: J. Gibbons, ed. *Language and the Law*. London: Longman. pp.136-155.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. AND FINEGAN, E., 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

BLACK, H. C., 1891/1991. *Black's Law Dictionary*. St. Paul, Minn.: West Publishing.

BORJA, A., 2000. *El texto jurídico inglés y su traducción al español*. Barcelona: Ariel.

BORJA, A. 2004. La investigación en traducción jurídica. In: M. A. García Peinado y E. Ortega, eds. *Panorama actual de la investigación en traducción e interpretación*. Granada: Atrio. pp.415-426.

BOWKER, L., 2002. *Computer-Aided Translation Technology: A Practical Introduction*. Otawa: University of Otawa Press.

CABRÉ, M. T., 1998. Las fuentes terminológicas para la traducción. In: P. Fernández Nistal y J. M. Bravo Gozalo, *La traducción: orientaciones lingüísticas y culturales*. Valladolid: Universidad de Valladolid.

CARVALHO, L., 2008. Translating contracts and agreements: a *corpus* linguistic perspective. In: S. E. O. Tagnin y O. A. Vale. *Avanços da Linguística de Corpus no Brasil*. São Paulo: Humanitas.

CORPAS, G., 1997. *Manual de fraseología española*. Madrid: Gredos.

CORPAS, G., 2001. La traducción de las unidades fraseológicas: técnicas y estrategias. In: I. de la Cruz, C. Santamaría, C. Tejedor y C. Valero, eds. *La Lingüística aplicada a finales del S. XX: ensayos y propuestas*. Alcalá: Universidad de Alcalá. pp.779-786.

CORPAS, G. ed., 2003. *Diez años de investigación en fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos*. Frankfurt/Madrid: Vervuert/Lingüística Iberoamericana.

CORPAS, G., 2008. *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt: Peter Lang.

GARCÍA-PAGE, M., 2008. *Introducción a la fraseología española*. Barcelona: Anthropos.

GLÄSER, R., 1994-95. Relations between phraseology and terminology with special reference to English. *Alfa*, 7 (8). p.41.

GRANGER, S., 2009. Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Journal of Shanghai Jiaotong University*, [online] Available at: <http://sites.uclouvain.be/cecl/archives/Granger_Crosslinguistic_research.pdf> [Accessed 10 June 2015]

GRANGER, S. AND PAQUOT, M., 2008. Disentangling the phraseological web. In: S. Granger and F. Meunier, eds. *Phraseology: an Interdisciplinary Perspective*. Amsterdam: John Benjamins. pp.27-49.

GUSTAFSSON, M., 1975. *Binomial expressions in present-day English: a syntactic and semantic study*. Turku: Turun yliopisto

GUSTAFSSON, M., 1984. The syntactic features of binomial expressions in legal English, *Text*, 4 (1-3), p.123.

KOIKE, K., 2001. *Colocaciones léxicas en el español actual: estudio formal y léxico-semántico*. Alcalá de Henares: Universidad de Alcalá.

LORENTE, M., 2002. Terminología y fraseología especializada: del léxico a la sintaxis. In: G. Guerrero y L. F. Pérez Lagos, eds. *Panorama actual de la terminología*. Granada: Comares. pp.159-180.

MELLINKOFF, D., 1963. *The Language of the Law*. Boston: Little, Brown & Co.

MIGUEL (DE), E., 2000. El texto jurídico-administrativo: análisis de una orden ministerial. *Revista de Lengua y Literatura Españolas*, [e-journal] 2: 6-31. Available through: < http://www.ucm.es/info/circulo/no4/demiguel.htm> [Accessed 12 June 2015].

MONTERO MARTÍNEZ, S., 2002. *Estructuración conceptual y formalización terminológica de frasemas en el subdominio de la oncología*. Tesis doctoral inédita. Valladolid: Universidad de Valladolid.

MONZÓ, E. AND HOYO, E., 1998. La traducció dels textos jurídics al DOGV, *Fòrum de Recerca*, 3. [online] Available at: <http://sic.uji.es/publ/edicions/jfi3> [Accessed 10 June 2015]

PONTRANDOLFO, G., 2013. La fraseología como estilema del lenguaje judicial: el caso de las locuciones prepositivas desde una perspectiva contrastiva. In: L. Chierichetti and G. Garofalo, eds. *Discurso profesional y lingüística de corpus. Perspectivas de investigación*. Bergamo: CELSB. pp.187-215.

READ, J. AND NATION, P., 2004. Measurement of formulaic sequences. In: N. Schmitt, ed. *Formulaic Sequences*. Amsterdam: John Benjamins. pp.23-35.

ROBERTS, R., 1998. Phraseology and Translation. In: P. Fernández Nistal and J. M. Bravo Gozalo, eds. *La traducción: Orientaciones Lingüísticas y Culturales*. Valladolid: SAE. pp.61-78.

ROBERTS, R., 1994-1995. Identifying the Phraseology of Languages for Special Purposes (LSPs), *Alfa: Actes de langue française et de linguistique*, vol. 7(8). p. 61

RUIZ GURILLO, L., 1998. Una clasificación no discreta de las unidades fraseológicas del español. In: G. Wotjak, ed. *Estudios de fraseología y fraseografía del español actual*. Frankfurt/Madrid: Vervuert/Iberoamericana. pp.13-37.

RUIZ GURILLO, L., 2000. Cómo integrar la fraseología en los diccionarios. In: G. Corpas, ed. *Las lenguas de Europa: Estudios de Fraseología, Fraseografía y Traducción*. Granada: Comares. pp.261-274.

SCOTT, M., 2012. Wordsmith Tools 6.0. [computer program] Available at: <http://www.lexically.net/wordsmith/index.html> [Accessed 10 June 2015].

SINCLAIR, J., 2005. Corpus and Text - Basic Principles. In: M. Wynne, ed. *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, [online] Available at: <http://ota.ahds.ac.uk/documents/creating/dlc/chapter1.htm> [Accessed 10 June 2015].

TABARES, E. AND BATISTA, J. J., 2005. Notas sobre fraseología jurídica comparada español-alemán. In: R. Almela Pérez, E. Ramón Trives and G. Wotjak, eds. *Fraseología contrastiva: con ejemplos tomados del alemán, español, francés e italiano*. Murcia: Universidad de Murcia. pp.345-356.

TABARES, E. AND BATISTA, J. J., 2012. La competencia terminofraseológica del traductor jurídico. *Redit*, 8-1. p.13.

TEUBERT, W., 2005. My Version of Corpus Linguistics. *International Journal of Corpus Linguistics*, 10, p.1.

TIERSMA, P., 1999. *Legal Language*. Chicago/London: The University of Chicago Press.

ZANETTIN, F., 1998. Bilingual Comparable Corpora and the Training of Translators. *Meta*, XLIII, 4, [online] Available at: [http://www.erudit.org/erudit/meta/v43n04/] [Accessed 10 June 2015]

THE ADAPTATION OF ANGLICISMS – PHRASEOLOGICAL UNITS IN CROATIAN ECONOMIC TERMINOLOGY

Ivo Fabijanić
English Department
University of Zadar
ivo.fabijanic@unizd.hr

Lidija Štrmelj
English Department
University of Zadar
lstrmelj@unizd.hr

Abstract

In this research we were interested in finding practical solutions for a) the classification of anglicisms – phraseological units in Croatian on the material of economic terms, within the following groups of PUs: *phraseological nominations*, *phrasal verbs* and *irreversible binominals*, and b) determining the degrees of their adaptation in the Croatian language. The research provides data about various features of PUs (e.g. their structural features, semantic stability, idiomaticity, etc.), on the basis of which they were classified into the appropriate class of the aforementioned types of PUs. In the second part of the research, after the typological part had been done, the analysis of eighty-one PUs was conducted, which, according to our preliminary studies, showed a high probability of typological classification of adapted anglicisms in a way similar to that, by which nominal syntagms were analysed. Our study has shown that the principal degrees of adaptation of anglicisms – phraseological units on the levels of orthography and semantics, might be defined within the following frame of degrees: 1) the zero degree of orthographic and semantic adaptation, 2) the degree of orthographic and semantic compromise adaptation, and 3) the degree of complete orthographic and semantic adaptation of PUs.

1. INTRODUCTION

In previous studies on contact between English and Croatian in economic and medical terminology (Fabijanić 2008; Fabijanić, Malenica 2013), several new trends in the adaptation of anglicisms were noticed. Firstly, the contact has become more profound at a general, lexical level, and secondly, the same trend of their adaptation is evident at a narrower (more specialized), phraseological level. At the former, the anglicisms were analysed as polylexemic units, adapted to Croatian in a variety of ways, showing different degrees of their adaptation (Fabijanić 2008; Fabijanić 2011). According to the degrees of *transmorphemization* and the types of primary and secondary adaptation they were classified into, those adapted by *zero transmorphemization* represent the adaptation of nominal syntagms with or without any changes in morphemic structure and in the order of the components

(EN *free enterprise* > HR *free enterprise*); those adapted by the *compromise transmorphemization* contain the components with English zero free morphemes and/or English bound morphemes (EN *front runner* > HR *front runner*; EN *global sourcing* > HR *global sourcing*), or a combination of components with a bound morpheme from Croatian, the receiving language, and a bound morpheme from English, the giving language (EN *internet banking* > HR *internet bankarstvo*), and finally, to those adapted by the *complete transmorphemization*, whose order of components has been changed completely (EN *trend analysis* > HR *analiza trenda*). As far as the medical terminology is concerned, in Fabijanić, Malenica 2013 English medical abbreviations and their adaptation to Croatian were analysed, and it was concluded that, at the orthographic level (*transgraphemization*), abbreviations can be classified according to the same three-degree principle of adaptation: original English abbreviations with their orthographic and pronouncing forms were adapted by *zero orthographic change/zero transgraphemization*; those with English orthography and partially applied Croatian pronunciation, were adapted by *compromise transgraphemization*. The abbreviations with English orthography and pronounced as ordinary Croatian abbreviations were adapted by *complete transgraphemization*.

In this research, referring to the latter phraseological level, we are interested in finding practical solutions for: a) the classification of anglicisms - phraseological units (PUs) in Croatian, on the material of economic terms, and b) determining the degrees of their adaptation in the Croatian language, at the orthographic and semantic levels. Due to the fact that the PUs we are interested in, can be grouped within the confines of different specific types, for the sake of simplicity we used the classification proposed by Fiedler, who classified them into two major types: a) conventional types of PUs (as found in the majority of systemic descriptions) and b) special types of PUs (2007: 39). Consequently, the article can be described as consisting of two parts: the first part deals with the typological features of the PUs analysed and the second part deals with the analysis of their adaptation as anglicisms in the Croatian language.

2. THE METHODOLOGY

As to the methodology of research, we used the one devised by Filipović (1986, 1990) for the semantic adaptation of phraseological units, the work of Fabijanić (2011) for the adaptation of PU's at the morphological level, and the work of Fabijanić, Malenica (2013) for the analysis of adaptation at the orthographic level. We are aware of the fact that some recent research (Ajduković 2012, 2014) have proposed different and more systematic approaches in the adaptational analysis of PUs (the term used is *phraseologism* and the cover term for its adaptation is *contactophraseme*; Ajduković 2012: 141), but due to the fact that they refer to a wider scope of linguistic levels, i.e. to the adaptation of PUs at the levels of: orthography, prosody, phonology, morphology/grammar, rection, syntax, style, paradigmatic and conceptual levels, and that they refer to the adaptation of russianisms into the Slavonic languages, requiring an applied and considerably wider research of PUs, which would not fit the confines of this study and to the proposed length of the contributions set by the Europhras Scientific Committee, we decided to use a narrower approach by the auhors mentioned at the beginning of the paragraph.

3. THE STUDY SAMPLE AND THE SOURCES

The corpus is comprised of eighty-one English economic terms[165]. They originate from a variety of economic fields and subfields, like banking, finance, accountancy, investment, management, marketing, insurance, etc. The list of lexicographic sources for the English economic terms is given at the end of the article (under the title *Lexicographic sources*). Here we bring them up in alphabetical order – firstly, anglicisms, replicas in Croatian, and secondly, the English models[166]: *at or better*, HR *baby bond/baby bonds/bebi obveznica* (< EN *baby bond*), HR *back office* (< EN *back office*), *baloon payment*, HR *Baltic exchange* (< EN *the Baltic exchange*), HR *banana republika* (< EN *banana republic*), HR *bankovni kliring* (< EN *bank clearing*), *bear market*, *big bath*, HR *big board* (< EN *the big board*), HR *blue chip/-s* (< EN *blue chip*), HR *blue sky laws* (< EN *blue sky/blue-sky law/-s*), *bottom line*, HR *bottom up* (< EN *bottom-up approach*), *brain drain*, HR *brend imidž/imidž brenda* (< EN *brand image*), *bull market*, *bulldog bond*, *bulldog market*, HR *buy in* (< EN *buy-in*), HR *buy out* (< E *buyout/buy-out*), HR *call in pay* (< EN *call-in pay*), HR *cash flow* (< EN *cash flow*), *cash & carry*, HR *čarter party* (< EN *charter party*), *closed company*, *closed shop*, *closing day*, HR *kreditna linija* (< EN *credit line*), HR *direct/direkt costing* (< EN *direct costing*), HR *disaster recovery centar* (< EN *disaster recovery*), *dollar cost averaging*, HR *drop-off charges* (< EN *drop-off charge*), *evergreen credit*, *fiat money*, HR *financijska piramida* (< EN *financial pyramid*), HR *financijski supermarket* (< *financial supermarket*), *flat tax*, *fleet manager*, *float money*, *floor broker*, HR *free lance/-er* (< EN *freelance/-er*), *front office*, *going concern*, *going private*, *going public*, *grace period*, HR *greenfield investicija/ulaganje* (< EN *green field/greenfield investment*), HR *head hunting tvrtka* (< EN *headhunting firm*), *hot bunking*, *hot money*, *insider dealing*, *insider trading*, HR *jumbo/džambo zajam* (< EN *jumbo loan*), HR *junk bonds/obveznice* (< EN *junk bond*), *lead manager*, *long position*, *mail order*, *market maker*, HR *menedžerska revolucija* (< EN *managerial revolution*), *missing the market*, HR *moralni hazard* (< EN *moral hazard*), *mystery shopping*, HR *New Deal/new deal* (< EN *New Deal*), HR *odd dates* (< EN *odd date*), *open interest*, *open order*, *open shop*, *out of the money*, *predatory pricing*, *price cutting*, HR *put opcija* (< EN *put option*), *red herring*, *red tape*, *senior security*, *short position*, *short sale*, *sinking fund*, *Wall Street*, *wholesale bank*, *yankee bond*.

As to the types of aforementioned English PUs, they can be classified into three types (according to Fiedler 2007): *phraseological nominations*, *phrasal verbs*, and *irreverisble binominals*. They were sorted out into the corresponding types: one of them belongs to the group of phrasal verbs, i.e. *buy out*, and two belong to the group of irreversible binominals, i.e. *at or better* and *cash and carry*. The rest belong to the group of phraseological nominations.

With regard to the grammatical (morphological) structure of the PUs, they were classified into six categories, according to their inherent parts of speech. The majority of PUs (75 or 94% of PUs) have the structure of noun phrases, i.e. they are either composed of **N** + **N**, e.g. *bull market*, or of **Adj** + **N**, e.g. *financial pyramid*, *red herring*. Rarely will the noun phrases be consisted of a **Ger** + **N** or **N** + **Ger**, i.e. *closing day* and *bank clearing*. Prepositional phrases or adjuncts modifying verbs are represented by one example (1%) only, i.e. *out of the money*. Other structural types include: **V** + **Obj/N** (3 PUs or 3%), e.g. *missing the market*; **V** + **Adv** (1 PUs or 1%), e.g. *buy out*; and **Prep** + **Conj** + **Adj/Comp** (1 PU or 1%), e.g. *at or better*.

According to the information provided in the dictionaries the terms were elicited from, labels describing their sense restrictions tell us that for most of them their main characteristic is their formal, official use, whereas only one was labelled as informal i.e. *red*

[165] Not all of them, according to the lexicographic sources, are phraseological units in English, e.g. *buy-in*, *buy-out/buyout*, *freelance/-r*.

[166] Where there are no indications of origin, i.e. no symbols EN and HR, models and replicas are represented by identical forms.

tape. Some of them were labelled as jargonistic expressions, e.g. *banana republic*, *the big board*, *missing the market*, *out of the money*, and *wholesale bank*.

4. THE ANALYSIS

Previous studies on anglicisms (Filipović 1986, 1990) and anglicisms – nominal syntagms (Fabijanić 2008, 2011), have confirmed the possibility of their analyses and classifications according to three adaptational degrees on morphological levels, within the frames of primary and secondary adaptation: zero, compromise and complete transmorphemization. This study also showed that the principle degrees of adaptation, with some respective changes, can be applied to the analysis of anglicisms – PUs in Croatian economic terminology.

Firstly, we dealt with their orthographic adaptation. In Fabijanić, Malenica 2013, it was explained through the three degrees of zero, compromise and complete trangraphemization of abbreviations. The same degrees were succesfully applied to the corpus of PUs. Zero trangraphemization is applied to the units which have retained their original English orthography and pronunication, e.g. *closing day*, *open shop*, *Wall Street*, and it refers to the 84% of the sample. Compromise transgraphemization describes the adaptation of PUs with a combination of a component with the English orthography and (at least) one fully adapted component with the Croatian pronunciation, e.g. *čarter party*, *financijski supermarket*, *moralni hazard*[167] (10%). Complete transgraphemization describes those PUs (6%) which have retained English models but whose pronunciation has been fully adapted to Croatian pronunciation, e.g. *brend imidž/imidž brenda*, *bankovni kliring*, *kreditna linija*, *financijska piramida*[168].

As to the morphological structure of PUs - economic terms in Croatian, they can be classified into the classes having *bipartite*, *tripartite* and *quadripartite structure*. The most common is the bipartite structure class (approx. 90%), e.g. *big bath*, *bottom line*, *brend imidž*, which is followed by approximately 9% of those having tripartite structure, e.g. *blue sky laws*, and only one (approx. 1%) with the quadripartite structure, e.g. *out of the money*. The analysed PUs consist mostly of two autosemantic components whose main characteristic is stability and fixety. Most bipartite types have the structure of two autosemantic words. Rarely does a synsemantic word appear in their structure, e.g. a preposition *up* in *bottom up*. The rate of synsemantic words (the examples of conjunctions, articles and prepositions were also registered in our corpus) in tripartite and quadripartite types is a bit higher than in bipartite types, due to their multiple-component structure.

Regarding the elements of their morphological structure, each of the above types (cf. § 3) is represented by several subtypes. In bipartite structural type the most numerous ones are those consisting of **Adj + N** (30 PUs, e.g. *big bath*, *blue chip/-s*, *hot money*, *long position*, *red herring*) and **N + N** (28 examples, e.g. *baloon payment*, *bulldog market*, *brain drain*, *float money*, *floor broker*). The former type usually consists of both English elements, but ocassionally it may consist of an aglicism with a Croatian adjective-forming suffix (*-ski/-ska*, *-ni*)

[167] Some of them have adapted their orthography in a variety of ways but still fall into the type of compromise transgraphemization. Thus, in *junk obveznice* there is a calqued component *obveznice* (< *bonds*), in *head hunting tvrtka* there is a calqued component *tvrtka* (< *company*), in *direkt costing* the first component was adapted according to the Croatian pronunciation and the second was not. In *disaster recovery centar* (< *disaster recovery*), a third component *centar*, otherwise not found in the model, was added, with its Croatian pronunciation.

[168] A hybrid *džambo zajam* (< *jumbo loan*), composed of fully adapted adjectival anglicism *džambo* and a calqued nominal component *zajam*, belongs to the type of PUs with the complete transgraphemization.

combined with a noun, like *financijska piramida*, *menadžerska revolucija*, *moralni hazard*. The latter type can sometimes have the English plural form for its second constituent, e. g. *baby bonds*, *junk bonds*, *drop-off charges*, *odd dates* or *blue chips*. Other bipartite structures include types, like **Ger + N** (*closing day*, *going concern*, *sinking fund*); **N + Ger** (*insider dealing*, *insider trading*, *mystery shopping, price cutting*); **Adj + Ger** (*hot bunking*, *predatory pricing*). In the example of *bankovni kliring*, having the same Adj + Ger structure, the first constituent has been completely adapted to the morphological structure of Croatian adjectives by adding the suffix *-ovni*). Among other types, we also found **Ger + Adj** (*going private*, *going public*); **V + Adv** (*buy in*, *buy out*[169]), and **N + Prep** (*bottom up*[170]). In tripartite structural types the following combinations of constituents were analysed: **N + N + N** (*disaster recovery centar*[171]), **N + N + Ger** (*dollar cost averaging*), **N + Ger + N** (*head hunting tvrtka*[172]), **Prep + Conj + Adj/Comp** (*at or better*), **N + Prep + N** (*call in pay*), **Ger + Art + N** (*missing the market*), **N + Conj + V** (*cash & carry*), and **Adj + N + N/Pl** (*blue sky laws*). With the reference to the quadripartite structural type, only one example was recorded – **Adv + Prep + Art + N** (*out of the money*).

Desemantization of PUs was also confirmed at two levels (Fink-Arsovski 2002: 7). Most numerous one is the level of *partial desemantization* in which a part of PUs' components have lost their primary lexical meaning so that the non-desemanticized ones contribute to their phraseological meaning. Partial desemantization is evident in almost 90% of examples in bipartite and tripartite types, e.g. in *Baltic exchange* (< *the Baltic Exchange*) the meaning of the geographical entry has been lost but the meaning of the second component has not – it still refers to "[…the] agreements for the transport of raw materials between countries, especially by ship." (cf. *Oxford Business English Dictionary for learners of English*). In *predatory pricing*, the act of predation is understood metaphorically within the practice of "[…] setting prices at very low levels with the objective of weakening or eliminating competitors […]. (cf. *Dictionary of Economics*). A tripartite type with partially desemanticized components will be exemplified with *blue sky laws* (< *blue sky law*). The component *law* represents a state law, whereas *blue sky* refers to the size of fraud, which this type of law is supposed to give protection against (cf. *Webster's New World Finance and Investment Dictionary*). In bipartite and quadripartite types we found fourteen examples of *completely desemanticized* PUs. Complete desemantization refers to the full semantic reshaping of all components in PUs. Such are the cases of: *big bath*, *big board*, *blue chip*, *brain drain*, *closed shop*, *hot bunking*, *lead manager*, *long position*, *market maker*, *open shop*, *red herring*, *red tape*, *short position*, and *out of the money*. The PU *red herring*, for example, stands for "a preliminary prospectus issued by a company that is planning a public securities offering […]" (cf. *Webster's New World Finance and Investment Dictionary*). The PU *out of the money* (< *out-of-the-money/out of the money*) refers to "An option that has no value. A call option, which gives the holder the right to buy a security, is out-of-the-money when the price of the underlying security is below the option's strike price […]" (cf. *Webster's New World Finance and Investment Dictionary*).

Regarding the degrees of the overall semantic adaptation, phraseological units under analysis were adapted to Croatian through the *zero*, *compromise* and *complete degree of adaptation*. Contrary to the classification of nominal syntagms' adaptation, whose morphological structure and consequent morphological changes were the primal reason of their appropriate three-degree classification, the analysis of anglicisms – PUs in economic

[169] These examples are not considered to be PUs in English due to their monolexemic, compounded structure – *buy-in* and *buy-out*.

[170] The English *bottom-up approach* has gone through the elision of its second element and of a dash in the compounded first element.

[171] In English it is an expression of the bipartite structure, i.e. *disaster recovery*.

[172] In English it is an expression of the bipartite structure, i.e. *headhunting firm*.

terminology has indicated a single most important factor in classifying different degrees of their adaptation, and this is the adaptation at the semantic level, showing various nuances and contrasts in the adapted meaning/-s.

The zero degree of transsemantization refers to PUs which retained their original structure and the order of structural elements from the giving language into the receiving language, together with the authentic meaning/-s and idiomaticity, e.g. ***flat tax*** (one meaning in English) is '*a system in which tax is paid at the same rate, however much you earn or spend*' (*Oxford Business English Dictionary*), and in Croatian it also has one meaning and stands for: '*paušalna porezna stopa, paušalni porez*'[173]; ***red tape*** (one sense) in English stands for '*official rules that seem more complicated than is necessary and prevent things being done quickly*' (OBED), and in Croatian (one sense) for: '*birokratska rutina, procedura, posebice ona koja je nepotrebna ili izuzetno kompliciranа, birokracija*'; ***predatory pricing*** (one sense) describes '*a situation where a company makes its prices very low, even though this will lose money, so that other companies cannot compete and have to stop selling similar goods or services*' (OBED), and its equivalent (one) sense in Croatian is: '*politika utvrđivanja niskih cijena koju vodi poduzeće s ciljem da s tržišta istisne konkurente*'.

The second degree of transsemantization or the degree of compromise semantic adaptation of PUs is attributed to those which have retained their original structure and morphological elements of the giving language with a possibility of adding morphological structures from the receiving language, and a relatively variable degree of authentic idiomaticity, i.e. anglicisms have undergone the specialization of meaning either in number of meanings or in a specific field of the meaning, e.g. the meaning of ***hot money*** (four senses[174] in *Dictionary of Banking Terms*), i.e. '*money that is moved quickly between countries in order to make profits from changes in interest rates or in the value of currencies*' specialized its meaning in number, and is equivalent in Croatian to '*vrući novac*[175], *novčani kapital koji se naglo, ali privremeno prebacuje u inozemstvo da bi se izbjegle posljedice devalvacije u vlastitoj zemlji ili ostvarile kamate više od domaćih*'. The same goes for ***sinking fund*** (three meanings in *Dictionary of International Economic Terms* and in *Dictionary of Banking Terms*) – '*money that a company keeps and adds to regularly in order to pay debts, pay for equipment, etc. at a fixed date in the future*', in Croatian, one specialized meaning - *'sredstva rezervirana za otkup dionica, povlačenje obveznica ili za zamjenu osnovnih sredstava, amortizacijski fond*'. Specialization within the field of meaning is evident for ***call in pay*** (one sense in OBED), i.e. '*an amount of money paid to workers who are asked to go to work outside their usual hours, even if there is no work for them to do*', which in Croatian specialized its meaning in the field and it means: '*minimalni iznos plaće koji dobiva zaposleni za svoj posao, neovisno o tomu je li posao završen ili nije*.'

The third degree of transsemantization or the complete semantic adaptation is ascribed to the PUs whose original structure and the order of elements have been retained with occasional examples of adapted morphological structures from the receiving language with the appropriate valency, but whose degree of variability/alteration of idiomaticity in replicas has substantially risen. The change is evident in the widening of PUs' meaning, either in number of meanings or in the field, e.g. ***short position*** in English describes *'a situation in which an investor sells or agrees to sell shares, currencies, etc. that he/she does not own yet,*

[173] For the list of lexicographic sources used in the analysis of semantic adaptation of PUs, see the list of the Croatian monolingual, general and specialized dictionaries, as well as the bilingual Croatian-English/English-Croatian dictionaries, general and specialized, and web-sources, in Fabijanić 2008, pp.30 – 32.

[174] For the sake of saving space, we bring only one which was adapted to Croatian. The same approach is used in explication of *sinking fund*.

[175] This is a calque of the anglicisms *hot money*, thus testifying not only its further develpoment and adaptation within the specialized lexicon of economic terminology, but also a new entry within the phrasicon of the Croatian language.

hoping that the price will fall and they will make a profit by buying them later at a lower price', whereas in Croatian it stands for '*bilo koja knjigovodstvena pozicija unutar računa vrijednosnih papira tvrtke koji ima potražni saldo*'. ***Kreditna linija*** (< *credit line*) – '*an amount of credit that a bank, company, etc. makes available to a person or a company for a particular period*', widened its meaning in Croatian, i.e. '*međusobni sporazum o kreditiranju u vezi s poslovima platnog prometa*'. It has also widened its meaning within the semantic field, and it means: '*kredit banke banci u inozemstvu kako bi se na teret kreditne linije financirao izvoz komitenta banke kreditora*' and '*dugoročni kreditni aranžman Međunarodne banke za obnovu i razvoj s bankama iz zemalja članica za financiranje posebnih namjena*'. ***Mail order*** stands for '*a system of buying and selling goods through the mail*', and in Croatian it widened its field of meaning into '*uspostavljanje kontakta između kupca i prodavača neupravnim komunikacijskim sredstvima kao npr., katalozima, telefonskim pozivima, propagandnim letcima ili teletekstom*'.

5. CONCLUSION

It can be concluded that the sample of 81 anglicisms – PUs within the specialised phrasicon of anglicisms in Croatian economic terminology, was successfully classified within the delimitations of different typological features of phrasiological units in both, the borrowing and the receiving languages (English and Croatian). The conducted analyses have proved the presence of the idiomatic features, the corresponding grammatical/morphological features, different degrees of desemantization, and the formality status. Our hypotheses for the appropriate classification and typology of anglicisms – PUs, according to the corresponding degrees of adaptation, have also been confirmed. At the orthographical level, the PUs were successfully classified under three degrees of transgraphemization: zero, compromise and complete, and at the semantic level under the degrees of zero, compromised and complete transsemantization.

References

AJDUKOVIĆ, J., 2012. *Radovi iz lingvističke kontaktologije*. Beograd: Foto-futura.

AJDUKOVIĆ, J., 2014. *Rusko-inoslavenska kontaktološka istraživanja*. Beograd: Foto-futura.

ALLERTON, D.J., NESSELHAUF, N. AND SKANDERA, P. eds., 2004. *Phraseological Units: basic concepts and their application*. Basel: Schwabe.

BANNOCK, G., ET AL., 2003. *The Penguin Dictionary of Economics*. 5th Ed. London: Penguin.

BLACK, J., 2003. *A Dictionary of Economics*. Oxford: OUP.

CAPELA, J. J. AND HARTMAN, S. W., 2004. *Dictionary of International Business Terms*. New York: Barron's.

CLARK, J. ED., 2006. *Dictionary of International Economic Terms*. London: LES5ONS.

DOWNES, J. AND GOODMAN, J. E., 2006. *Dictionary of Finance and Investment Terms*. New York: Barron's.

ETZEL, B. J, 2003. *Webster's New World Finance and Investment Dictionary*. Indianapolis: Wiley Publishing.

FABIJANIĆ, I., 2009. *Anglizmi u ruskoj i hrvatskoj ekonomskoj terminologiji*. doktorska disertacija u rukopisu. Zadar.

FABIJANIĆ, I., 2011. Reinterpretacija transmorfemizacije anglizama – imeničkih sintagma u ruskome i hrvatskom jeziku. *Fluminensia*. 23(1), pp. 67-83.

FABIJANIĆ, I. AND MALENICA, F., 2013. Abbreviations in English Medical Terminology and their Adaptation to Croatian. *JAHR*. 4(7), pp. 71-105.

FIEDLER, S., 2007. *English Phraseology: a Coursebook*. Tuebingen: Narr.

FINK-ARSOVSKI, Ž, 2002. *Poredbena frazeologija: pogled izvana i iznutra*. Zagreb: Filozofski fakultet.

FITCH, T. P., 2006. *Dictionary of Banking Terms*. New York: Barron's.

FIRTH, J. R., 1957/1986. A Synopsis of Linguistic Theory, 1930-55. In: Selected Papers of J.R. Firth 1952-59. Ed. by F.R. Palmer (1968). London: Longman, pp.168-205.

MARSHALL, G., ED., 1998. *A Dictionary of Sociology*. New York: OUP.

MERRIAM WEBSTER'S COLLEGIATE DICTIONARY. 1994.1998. In: *Encyclopaedia Britannica*. CD-ROM.

PARKINSON, D., ED., 2005. *Oxford Business English Dictionary for learners of English.* Oxford: OUP.

The New International Webster's Pocket Business Dictionary of the English Language. Naples: Trident Press International.1998.

(Websites)

http://www.answers.com
http://www.investopedia.com
http://www.dictionary.bnet.com
http://www.duke.edu/~charvey/Classes/wpg/glossary.htm

ITALIAN MULTIWORD ADVERBS: DISTRIBUTIONAL FEATURES AND FUNCTIONAL PROPERTIES. A CORPUS BASED ANALYSIS

Valentina Piunno
Roma Tre University
valentina.piunno@uniroma3.it

Abstract

This research proposes a theoretical and descriptive study of a set of Italian Multiword Adverbs extracted from a corpus; the corpus-based investigation has been driven simultaneously considering the interaction of distributional and functional properties of selected Multiword Adverbs. In particular, the research deals with prepositional phrases covering an adverbial function and showing a particularly high degree of cohesion between their constituents (*Multiword Adverbs-PP*), such as: *di solito* 'usually', *a occhio e croce* 'more or less', *per filo e per segno* 'chapter and verse'. The quantitative corpus analysis reveals that this analytical resource is particularly exploited in Italian. The investigation aims at achieving two main objectives. Firstly, the computational analysis reveals to which extent the distribution of Multiword Adverbs-PP may vary depending on their functional properties. Thirdly, the corpus-based investigation gives evidence of subcategories, according to the distributional properties as well as to the functional values of sets of Multiword Adverbs-PP.

1. MULTIWORD ADVERBS: NATURE, TYPES, AND CONFIGURATIONS

According to typological studies, languages can employ several lexical devices to express the adverbial modification (Van der Auwera 1988, Hengeveld 1992). Furthermore, the category of adverbs is semantically, morphologically, syntactically and functionally heterogeneous (Van der Auwera 1988, Cinque 1999, Givón 2001, Haspelmath 2001). In Italian, as well as in other Romance languages, the adverbial function can be realized by different lexical elements, such as single lexical items (i.e. non-derived adverbs such as *sempre* 'always', derived adverbs such as *lentamente* 'slowly') or adverbials constructions – mainly nominal phrases (i.e. *spalle al muro* 'back to the wall') and prepositional phrases (i.e. *alla bell'e meglio* 'roughly'), but also propositions (i.e. the gerundive, subordinate clauses, etc.) (Lonzi 2001) –[176]. As far as Italian is concerned, the prepositional phrase is the most

176 Among others, Gross 1990, Ramat and Ricca 1994, Alexiadou 1997, Lonzi 2001, Bosque and Demonte 2009, Bosque 2010.

common syntactic configuration for adverbial constructions (Voghera 2004). Furthermore, the distribution of prepositional phrases with an adverbial function is usually described as less free than the one of proper adverbs (Alexiadou 1997). Prepositional phrases functioning as Multiword Adverbs (i.e. *di solito* 'usually', *a occhio e croce* 'more or less') will be the focus of this investigation.

This research proposes a theoretical and descriptive study of Multiword Adverbs-PP, resulting from an on-going corpus-based investigation driven simultaneously considering the interaction of distributional and functional properties.

Multiword Adverbs-PP (hereinafter *MAs-PP*) will be referred to as to complex lexical items composed of two or more words syntactically bound, covering an adverbial function in context. MAs-PP form an heterogeneous class which has to be analysed along a *continuum* of fixedness, cohesion and lexicalisation. Classifications may vary on the basis of different parameters, such as: i) syntactic configurations, ii) function, iii) morphological, syntactic, semantic and lexical restrictions, iv) degree of cohesion and lexical specification.

MAs can be subdivided into several groups, depending on their syntactic configurations. Table 1 exemplifies some of the most common patterns of Italian prepositional phrases functioning as Multiword Adverbs.

Syntactic configuration	Example	Translation
Prep Noun	*per esempio*	'for example'
Prep Adj Noun	*in primo luogo*	'firstly'
Prep Noun Pre Noun	*di volta in volta*	'time by time'
Prep Noun Adj	*in caso contrario*	'otherwise'
Prep Art Adj Noun	*per la prima volta*	'for the first time'
Prep Adv	*per ora*	'for the moment'
Prep Art Noun	*per il momento*	'for the moment'
Prep Noun Conj Noun	*in fretta e furia*	'hastily'
Prep Art Noun Pre Art Noun	*con la grazia di un elefante*	'with the elegance of an elephant'
Prep Art Adv	*al meglio*	'to the best'
Prep Noun Conj Pre Noun	*a torto o a ragione*	'whether right or wrong'

Table 1. Most common configurations for Italian MAs-PP

As far as the functional properties are concerned, Italian MAs-PP – just like proper adverbs – can fulfil the verbal modification (1), the adverbial modification (2), the adjectival modification (3), as well as the sentence modification (4).

(1) *Lucia esce* ***di rado***
'Lucia goes out rarely'

(2) *Marina arriverà* ***al più tardi*** *domani*
'Marina will arrive tomorrow at the latest'

(3) *Ho mangiato un panino* ***a dir poco*** *eccellente*
'I ate a simply great sandwich'

(4) ***A parte gli scherzi****, il film era davvero noioso*
'Joking aside, the movie was really boring'

It is worth noting that the MA-PP is usually located next to the word it modifies; however, its syntactic distribution may vary according to different factors, such as the specific function (different syntactic positions can be associated to each function, cf. ¶2.2) or particular stylistic choices (such as in poetry, for example).

Several kinds of restrictions can operate on MA-PP[177]: *morphological*, *semantic*, *syntactic* and *lexical restrictions*. As many studies have already highlighted[178], MA-PP usually show restricted morphological variability (i.e. constituents tend to be not pluralizable), as well as an extremely limited lexical and syntactic openness to variation (they usually do not allow for paradigmatic variation, lexical insertion, lexical suppression, and inversion of constituents); moreover, their semantics is often not compositional. With respect to semantic and syntactic properties, a low level of semantic compositionality generally corresponds to a high degree of syntactic cohesion and lexicalisation of the multiword unit (Piunno 2013). However, the extent to which restrictions are enforced may vary according to the type of MA (Svensson 2004, Lavieu 2005, Piunno 2013): on the one hand, some sequences can only allow for some variations and refuse other transformations, and on the other hand, similar multiword units do not regularly respond in the same way to restrictions.

On the basis of the fixedness of their constituents, MAs show a lower or a higher degree of syntactic cohesion: the stronger is the cohesion between its constituents, the higher is the fixedness of a MA. Depending on the fixedness degree, MAs can be distinguished into *partially fixed* and *totally fixed* (Piunno 2013, Piunno in press). While totally fixed MAs do not allow for any lexical variations, partially fixed MAs show weaker cohesion and internal lexical variation (Piunno in press). Partially fixed MAs may also show several degrees of lexical specification (Simone *et al.* 2013). Some MAs allow for the substitution of one of their constituents, even if the paradigmatic choice may be limited to particular semantic fields, as in (5):

(5) [*in* $\text{ADJ}_{\text{primo/secondo/...}}$ *luogo*]
Adj={primo, secondo, terzo,...}--->Adj={ordinal adjectives}

In this case the adjective shows a – thus restricted – paradigmatic variability (since the MA only admits ordinal adjectives).

2. OBJECTIVES AND METHODOLOGY

This investigation aims at defining the functional and syntactic properties of the MAs-PP class of Italian. The research has three main objectives. Firstly, the computational analysis can reveal to which extent the distribution of MAs-PP may vary depending on their functional properties. Secondly, it intends to detect the most frequent distributions associated to a specific function. Thirdly, the corpus-based investigation will provide a classification in subcategories, according to the distributional properties as well as to the functional and semantic values of sets of MAs-PP.

The function and the syntactic behaviour of Italian MAs has been determined relying on quantitative data extracted from a lemmatized and PoS-tagged corpus of written Italian (*La Repubblica* corpus[179]).

177 Cf. among others, G. Gross (1996), Mejri (1997), Mel'čuk et al. (1995), Tutin-Grossmann (2002), Lamiroy (2003), Voghera (1994), Simone (2006a, 2006b).

178 Cf. Piunno (2013) for Italian.

179 La Repubblica is a corpus of Italian written language including includes 380.000.000 tokens (Baroni et al., 2004).

Firstly, a set of Italian MAs-PP has been extracted from the corpus; the extraction has been driven using frequency lists based on PoS-tagging. The configurations represented in Table 1 have been selected in a set of different patterns, and only 630 MAs the most frequently occurring in the corpus have been chosen for the analysis. The table below registers the 11 selected syntactic configurations and the number of MAs collected for each pattern.

Syntactic configuration	Number of MAs analysed
Prep Noun	228
Prep Adj Noun	130
Prep Noun Pre Noun	86
Prep Noun Adj	78
Prep Art Adj Noun	30
Prep Adv	23
Prep Art Noun	20
Prep Noun Conj Noun	18
Prep Art Noun Pre Art Noun	9
Prep Art Adv	7
Prep Noun Conj Pre Noun	1
TOTAL	**630**

Table 2. Configurations of selected Italian MA-PP

Secondly, each one of the selected MAs has been tested on the corpus for its function and for the related distributional properties. Tests have been driven applying two different methods: firstly, data have been automatically extracted through combined PoS - patterns - token queries[180], and secondly, each set of results has been manually evaluated.

The testing possibilities for functions are four: i. MAs having a modifying function over adjectives, ii. MAs having a modifying function over adverbs, iii. MAs having a modifying function over verbs, iv. MAs having a modifying function over sentences[181]. Furthermore, the testing possibilities for the adverbial distribution depends on the function covered by the prepositional phrase, as the table below shows:

Function	Distribution		
Adjective Modifier	Before the adjective	After the adjective	
Adverb Modifier	Before the adverb	After the adverb	
Verb Modifier	Before the verb	Between the auxiliary and the verb	After the verb
Sentence Modifier	At the beginning of the sentence	At the end of the sentence	

Table 3. Distribution of Italian MAs-PP depending on their function

180 E.g. "any adjective followed by a MA-PP", or "any adjective preceding a MA-PP".

181 Even if it would be particularly interesting to investigate on semantic classifications of each group, this paper will leave semantic analysis apart, being this type of theoretical features not the focus of this work.

Since in Italian the adjective can be modified by an adverb that is usually preposed or postposed to the adjective, for the adjectival modification the pre-adjectival and post-adjectival positions have been considered. For the adverbial modification, the positions before the adverb and after it have been evaluated. For the verbal modification, three different positions have been considered: i. preverbal position (that is to say, the position before the finite form of the verb), ii. post-auxiliary position (namely, the position between the auxiliary and the past participle form of the verb), and iii. post-verbal position (after the finite form of the verb). For sentence modification, two different positions have been evaluated: i. the sentence initial position followed by a comma intonation, and ii. the final position, preceded by a comma.

Each MA-PP has been evaluated against its functional and distributional properties; then, in order to put in relation the distributional features to the function covered, the observed frequencies have been normalised dividing the actual number of occurrences in a certain position by total occurrences of the MA-PP in a specific function.

3. RESULTS OF THE ANALYSIS

Since the main focus of the analysis is the definition of MAs-PP on the basis of their functional and distributional properties, two types of results have been considered. On the one hand, different MAs have been divided into classes on the basis of their syntactic function (¶3.1). On the other hand, MAs have been investigated considering the most frequently occurring syntactic position connected to a specific function (¶3.2).

3.1 Types of MAs depending on their syntactic function

On the basis of their syntactic function, five classes of MAs have been attested in the corpus, considering the 630 selected sequences[182]:

1) MAs only modifying verbs[183]:

(6) *Annuncia* ***in una sola volta*** *due spettacoli*
announce.3SG in a only time two exhibition.PL
'It announces two exhibitions all at once'

2) MAs modifying verbs and adjectives:

(7) *Sono sbucati* ***sul davanti***
come out.PST on.the front
'They came out in the front'

(8) *Però con l' abito azzurro, libero* ***sul davanti***
But with the dress light blue, free on.the front
'But with the light blue, empty in the front'

3) MAs modifying both verbs and the whole clause:

(9) *Di un professore ricorderà* ***al più*** *che pretendeva troppo*
of a professor remember.FUT.3SG at most that demand.PST.3SG too much
'He will remember at most his professor for demanding too much'.

(10) ***Al più,*** *[...], la facoltà di zapping può passare di padre in figlio.*

182 Since this investigation is driven considering just 630 Italian MAs-PP, it may be not excluded that the number of classes and the MAs classification could vary increasing the number of samples.

183 This group includes MAs-PP behaving as argument or as adjuncts. Even if this topic has gained considerable interest from linguists studying the adverbial phrases, this paper will not take into account the debate about the complex distinction between argument and adjuncts, and the issue related to the intermediate category of argument/adjunct. For further readings, cf. Dik 1980, Van Valin - LaPolla 1997, Lonzi 2001, Dowty 2003, Mereu 2009.

at most the power of zapping can.3SG pass.INF of father in son
'At most, [...], the zapping power can pass from father to son'

4) MAs modifying verbs, adjectives and sentences:

(11) *Le dichiarazioni confermano* ***in un certo modo*** *l' ufficialità di un cambiamento*
the declaration.PL confirm.3PL in a certain way the officiality of a change
'The declarations confirm in a certain way the official change'

(12) *Quando dico "regionale", intendo regionale* ***in un certo modo***
When say.1SG regional, mean.1SG regional in a certain way
'When I say "regional", I mean regional in a certain way'

(13) ***In un certo modo,*** *tutto congiura contro di noi*
in a certain way, everything conspire.3SG against of PRO.1PL
'In a certain way, everything conspires against us'

5) MAs modifying verbs, sentences, adjectives and adverbs:

(14) *Lo vidi* ***per la prima volta*** *a Parigi*
CLI see.PST.1SG for the first time at Paris
'I saw him for the first time in Paris'

(15) ***Per la prima volta****, uno dei capibanda è stato condannato*
For the first time, one of.the ringleader sentence.PASS.PST.3SG
'For the first time, one of the ringleaders has been sentenced'

(16) *La piattaforma,* ***per la prima volta*** *unitaria, prevede…*
The platform, for the first time unitary, consider.3SG
'The platform, unitary for the first time, considers…'

(17) *Suore e femministe* ***per la prima volta*** *insieme*
Nun.PL and feminist.PL for the first time together
'Nuns and feminists for the first time together'

The results of the functional classification are resumed in the following table:

	Adverb Modifier	Adjective Modifier	Sentence Modifier	Verb Modifier
I TYPE				+
II TYPE		+		+
III TYPE			+	+
IV TYPE		+	+	+
V TYPE	+	+	+	+

Table 4. Classes of Italian MAs-PP depending on their function

It is worth noting that the database does not contain examples of MAs *only functioning* as adverb modifiers or adjective modifiers or sentence modifiers. In fact, MAs usually modify a predicative element, functioning as the nucleolus of the sentence. More interestingly, a great number of multiword units, only acting as verb modifiers (Class I) in adverbial function (18), can also modify a noun, as in (19):

(18) *tornano* ***al sicuro*** *anche la madre e la sorella*
come back.3PL to.the safe also the mother and the sister
'His mother and sister come back to safety too'

(19) *Racconta delle due figlie* ***al sicuro*** *a Scutari*

tell.3SG of.the.PL two daughter.PL to.the safe at Scutari
'He tells about his two daughters at safe in Scutari'

In (19) the multiword unit "al sicuro" acquires an adjectival function, while in (18) it clearly has an adverbial function[184]. Furthermore, many MAs of the Class V are mainly employed in the role of sentence modifier, most of them denoting time or frequency (e.g. *per ora* 'for the moment', *di volta in volta* 'time by time').

3.2 Types of MAs depending on distributional features

From a distributional point of view, depending on their functional properties, Italian MAs-PP can cover different syntactic positions.

When the MA is employed as a verb modifier, it can be located along three different positions:

i) preverbal position (before the finite form of the verb):

(20) *La signora* ***a volte*** *parla* *tutto* *d' un fiato*
The woman at time.PL speak.3SG all of a breath
'The woman sometimes speaks without breathing'

(21) *Bahia* ***in fretta e furia*** *cambia* *le regole*
Bahia in haste and rush change.3SG the rule.PL
'Bahia hastily changes the rules'

ii) post-auxiliary position (namely, the position between the auxiliary and the past participle form of the verb):

(22) *Major* *veniva* ***di nuovo*** *interrogato*
Major come.PST.3SG of new question.PPAST
'Major was questioned again'

iii) post-verbal position (after the finite form of the verb):

(23) *Il* *pubblico* *commenta* ***ad alta voce***
The audience comment.3SG at high voice
'The audience comments loudly'

(24) *Il caos* *cresce* ***di giorno*** ***in giorno***
the confusion grow.3SG of day in day
'The confusion grows day by day'

The table below represents the frequency of occurrence of different patterns of MAs in the function of verb modifiers.

184 This is the case of Mixed Modifiers, multiword units having both an adverbial and an adjectival function, analysed in Piunno (in press), Piunno - Ganfi (2014).

Syntactic Configuration	Number of selected MAs	Preverbal position	Post-auxiliary position	Post-verbal position
Pre Noun	228	35422	79	228131
Pre Adj Noun	130	2436	8	18264
Pre Noun Pre Noun	86	589	4	6208
Pre Noun Adj	78	361	0	5145
Pre Adv	23	7066	6	21175
Pre Noun Conj Noun	18	115	0	425
Pre Art Adj Noun	30	5204	0	12971
Pre Art Noun	20	13571	6	25044
Pre Art Noun Pre Art Noun	9	1	0	103
Pre Art Adv	7	238	0	807
Pre Noun Conj Pre Noun	1	18	0	20

Table 5. Absolute frequency of verb modifiers[185]

By comparing the attested frequencies, an important deduction immediately leaps out: the post-verbal position is by far the most common distribution for MAs-PP employed as verb modifiers, while preverbal position is comparatively rarely attested, and post-auxiliary position is generally avoided. This is better explained in Fig.1, where the occurrences are represented as percentages that have been normalized according to the total frequency of occurrence of MAs as verb modifiers.

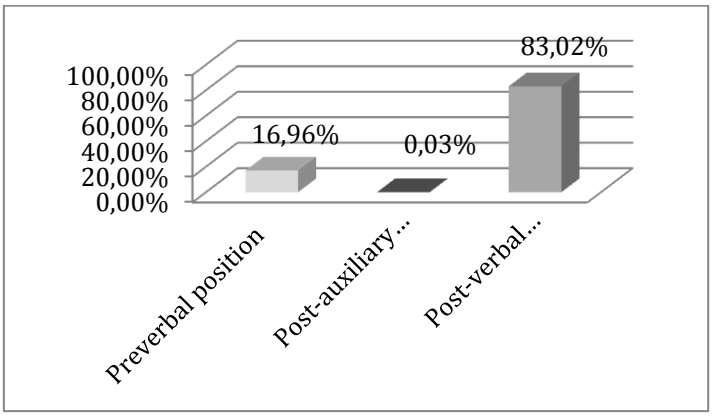


Figure 1. Normalized percentages of occurrence as verb modifiers

When a MA-PP occurs as a sentence modifier, two different positions are possible:

i. the sentence initial position (preceded by a full stop or a semicolon), followed by a comma intonation:

(25) ***In realtà,*** *c' è* *poco* *da dire*
in reality, there be.3SG less to say.INF
'Actually, there is nothing to say'

(26) ..; ***in pratica,*** *però,* *le cose sono* *più* *complesse*
in practice, however, the thing.PL be.3PL more complex.PL
'Basically, however, things are more complex'

185 This test has been driven on 630 MAs-PP.

ii. the sentence final position, with the MA preceded by a comma.

(27) *Ha vinto* *il buon senso,* ***per fortuna***
win.PST.3SG the good sense, for luck
'Luckily, the wisdom has won'

The following table shows the frequency of occurrence of both patterns:

Syntactic Configuration	Number of selected MAs	Sentence initial position	Sentence final position
Pre Noun	228	166289	2277
Pre Adj Noun	130	13377	358
Pre Noun Pre Noun	86	1615	15
Pre Noun Adj	78	2975	90
Pre Adv	23	34641	471
Pre Noun Conj Noun	18	390	1
Pre Art Adj Noun	30	19457	239
Pre Art Noun	20	66430	794
Pre Art Noun Pre Art Noun	9	0	0
Pre Art Adv	7	303	18
Pre Noun Conj Pre Noun	1	67	2

Table 6. Absolute frequency of sentence modifiers[186]

The graph below summarizes the percentages of occurrence of MAs in the function of sentence modifiers.

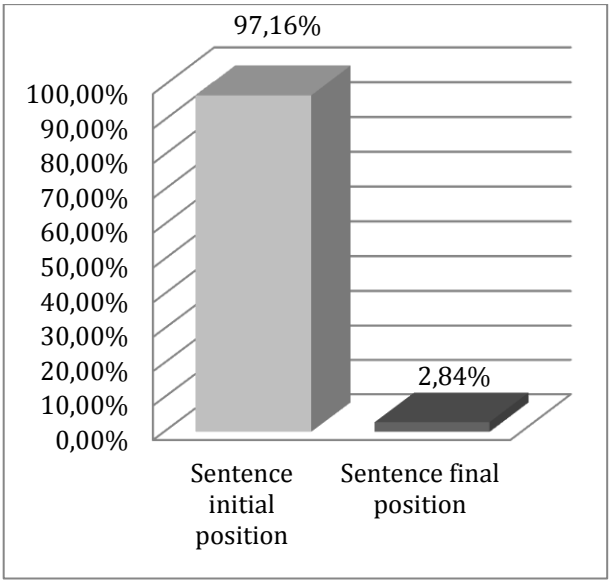


Figure 2. Normalized percentages of occurrence as sentence modifiers

The graph shows that the sentence initial position is extremely frequent, while the final position is admitted, but comparatively less common.

As far as the adjectival modification is concerned, the pre-adjectival and post-adjectival positions have been considered[187].

186 This test has been driven on 630 MAs-PP.

i. pre-adjectival position:

(28) *i disoccupati, giovani e meridionali* ***per la maggior parte***
the unemployed.PL young.PL and southern.PL for the major part
'Unemployed people, the majority young and southern'

ii. post-adjectival position:

(29) *La scaletta , **per ora** parziale, del mercoledi di RaiDue …*
The schedule for now partial of.the Wednesday of RaiDue
'The Wednesday schedule of RaiDue, preliminary for the moment …'

The data analysis – still underway for this test – has been driven considering a restricted number of MAs-PP[188]; however, some distributional tendencies are already traceable. The table below shows the first results of data extraction:

Syntactic Configuration	Number of selected MAs	Pre-adjectival position	Post-adjectival position
Pre Noun	3	630	491
Pre Adj Noun	3	1656	94
Pre Noun Pre Noun	10	182	363
Pre Noun Adj	3	2	155
Pre Adv	2	1226	176
Pre Noun Conj Noun	4	3	30
Pre Art Adj Noun	11	1136	709
Pre Art Noun Pre Art Noun	9	0	1
Pre Art Adv	7	0	56
Pre Noun Conj Pre Noun	1	6	0

Table 7. Absolute frequency of adjective modifiers

The graph below summarizes the distributional features of MAs covering the function of adjective modifier:

187 Due to limits imposed by the tagging of corpus, for this test both proper adjectives and past participle have been considered. However, MAs only selecting past participle have not been taken into account.

188 This test has been driven on 53 MAs-PP.

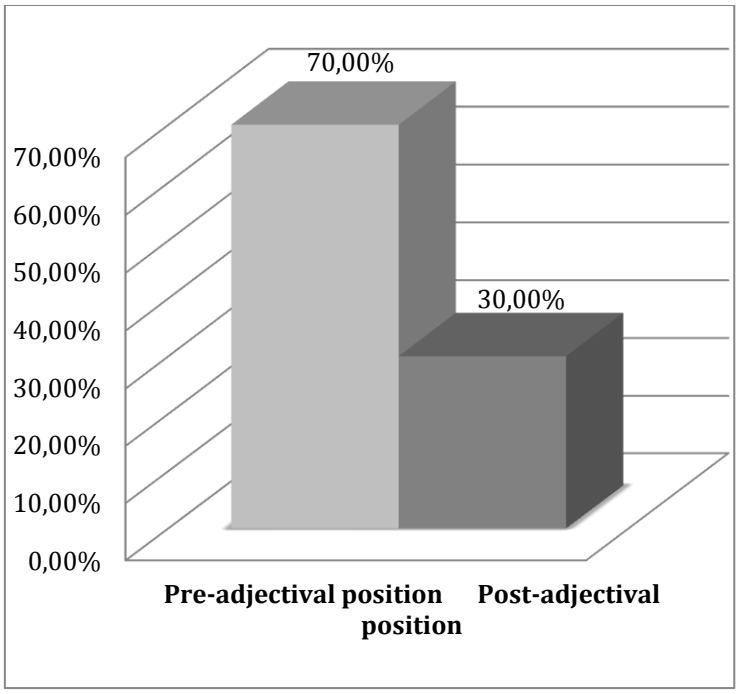


Figure 3. Normalized percentages of occurrence as adjective modifiers

As it is clear from the graph, the pre-adjectival distribution is much more preferred to the post-adjectival one. This could appear quite unusual if referred to prepositional phrases: in fact prepositional phrases use to follow the lexical element they modify. However, this is in line with the distributional properties of the proper adverb, most frequently occurring in the pre-adjectival position. Moreover, MAs-PP tend to select the pre-adjectival position because in the post-adjectival one they usually act as complements of the adjective.

MAs can act as an adverb modifier selecting two different positions:

i. pre-adverbial position:

(30) *dice,* ***per ora*** *sottovoce, che il ministro è un'anatra zoppa*
say.3SG, for now in a low voice, that the ministry be.3SG a duck lame
'He says, for the moment in a low voice, that the ministry is a lame duck'.

ii. post-adverbial position:

(31) *...i cechi e gli slovacchi, insieme* ***per l' ultima volta***
...the.PL Czech.PL and the.PL Slovakian.PL , together for the last time
'The Czech and the Slovakian together for the last time'

The table below shows the absolute frequency of MAs modifying adverbs[189].

[189] Also in this case the data analysis has been driven considering a restricted number of MAs-PP, since this test has been driven on 140 MAs.

Syntactic Configuration	Number of selected MAs	Pre-adverbial position	Post-adverbial position
Pre Noun	3	15	17
Pre Adj Noun	3	4	7
Pre Noun Pre Noun	86	20	77
Pre Noun Adj	3	4	2
Pre Noun Conj Noun	18	0	4
Pre Art Adj Noun	10	682	394
Pre Art Noun Pre Art Noun	9	0	0
Pre Art Adv	7	0	0
Pre Noun Conj Pre Noun	1	0	0

Table 8. Absolute frequency of adverb modifiers

The graph below summarizes the percentages of occurrence of MAs as adverb modifiers.

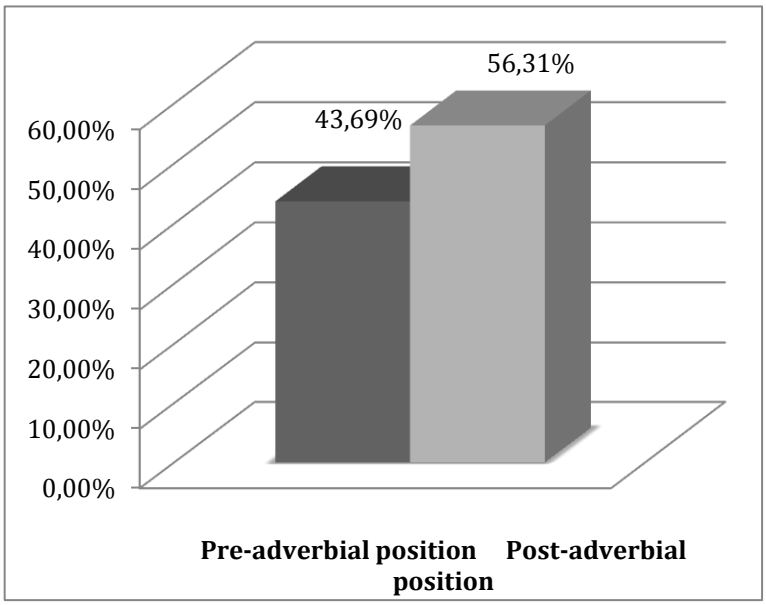


Figure 4. Normalized percentages of occurrence as adverb modifiers

The graph shows that the overall trend of the distribution of MA in adverbial function is almost equal in both positions, with a lower percentage of occurrences in pre-adverbial position.

4. CONCLUSIONS

This analysis has shown that, just as proper adverbs, Italian Multiword Adverbs-PP can be divided into different classes depending on their syntactic function.

The corpus-based investigation has highlighted five different subcategories of Multiword Adverbs, according to their functional values. All selected MAs are verb modifiers (some of them acting both as adverbials and as adjectivals). Most of MAs are also sentence modifiers, but not many of them are adverb and adjective modifiers.

Furthermore, the corpus-based analysis has revealed that the distribution of MAs-PP may vary depending on their functional properties. Thus, some generalizations are possible:

a) when they modify an adverb, they can be located before or after it, without any specific preferences;
b) when they modify an adjective, they tend to be located before it;
c) when they modify a verb, their location is preferably after it; they almost never take the post-auxiliary position.
d) if they cover the function of sentence modification, their syntactic distribution is preferably in starting position.

These distributional preferences could be related to different causes. Some restrictions are due to syntactic dependencies (as seen for adjectival modification in ¶ 3.2). Furthermore, prosodic heaviness factors may play a crucial role on the word order selection of MAs-PP (Piunno 2013). In fact, the syntactic complexity of multiword units may be phonologically too heavy to be managed by the speaker (cf. Voghera - Turco 2006). This is a highly relevant issue, still requiring further analysis.

Abbreviations

1SG	first person singular	FUT	future
1PL	first person plural	INF	infinitive
3SG	third person singular	PASS	passive
3PL	third person plural	PL	plural
ADJ	adjective	PRE	preposition
ADV	adverb	PRO	pronoun
ART	article	PST	past
CLI	clitic	PPAST	past participle
CONJ	conjunction		

References

ALEXIADOU, A., 1997. *Adverb Placement: A case study in antisymmetric syntax*, Amsterdam/Philadelphia: John Benjamins Publishing Company.

BARONI, M., BERNARDINI, S., COMASTRI, F., PICCIONI, L., VOLPI, A., ASTON, G. e MAZZOLENI, M., 2004. Introducing the La Repubblica corpus: a large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In: M. T. Lino *et al.*, eds. *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004, Lisbon, may 26-28). Paris: European Language Resources Association. pp. 1771-1774.

BOSQUE, I., 2010. *Nueva gramática de la lengua española*, Madrid: Real Academia Española.

BOSQUE, I. AND DEMONTE, V. eds., 2009. *Gramática descriptiva de la lengua española*. Madrid: Espasa.

CINQUE, G., 1999. *Adverbs and Functional Heads. A Cross-Linguistic Perspective*, Oxford: Oxford University Press.

DIK**,** S. C., 1980. *Studies in Functional Grammar*, London/New York: Academic Press.

DOWTY, D., 2003. The Dual Analysis of Adjuncts/Complements in Categorial Grammar, In: Ewald, L. *et al.*, eds. *Modifying Adjuncts*, New York: Mouton de Gruyter.

GIVÓN, T., 2001. *Syntax. An introduction*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

GROSS, G., 1996. *Les expressions fig.es en français*. Paris: Ophrys.

GROSS, M., 1990. *Grammaire transformationnelle du français*. 3. Paris: Asstril.

HASPELMATH, M., 2001. Word Classes and Parts of Speech. In: P. Baltes and N. Smelser, eds. *International Encyclopedia of the Social and Behavioral Sciences*. Amsterdam: Pergamon. pp. 16538–16545.

HENGEVELD, K., 1992. *Non-verbal Predication: theory, typology, diachrony*. Berlin: Mouton de Gruyter.

LAMIROY, B., 2003. Les notions linguistiques de figement et de contrainte, *Lingvisticae Investigationes*. 26 (1). pp. 1-14.

LONZI, L., 2001. Il sintagma avverbiale. In: L. Renzi, and G. Salvi, *Grande grammatica italiana di consultazione*. vol. 2. Bologna: Il Mulino. pp. 341-412.

MEJRI, S., 1997. Le figement lexical. Descriptions linguistiques et structuration sémantique, Publications de la Faculté des lettres de la Mannouba.

MEL'CUK, I., CLAS, A., AND POLGUERE, A., 1995. Introduction à la lexicologie explicative et combinatoire, Louvain-la-Neuve: Duculot.

MEREU**,** L., 2009. Gli aggiunti tra sintassi e pragmatica. In: E. Lombardi Vallauri and L. Mereu, eds., *Spazi linguistici. Studi in onore di Raffaele Simone*, Roma: Bulzoni, 93-109.

PIUNNO, V., (in press), Multiword Modifiers in Romance languages. Semantic formats and syntactic templates. *Yearbook of Phraseology,* Berlin: Mouton de Gruyter.

PIUNNO, V., 2013. *Modificatori sintagmatici con funzione aggettivale e avverbiale.* PhD Thesis. Roma: Università Roma Tre.

PIUNNO, V. AND GANFI, V., 2014, Distribution and evolution of Multiword Modifiers across Romance languages. A constructionist perspective, Paper presented at *8th International Conference on Construction Grammar*, University of Osnabrück, 3-6 Settembre 2014.

RAMAT P. AND RICCA D., 1994. Prototypical adverbs: on the scalarity / radiality of the notion of adverb*, Rivista di Linguistica.* 6 (2). pp. 289-326.

SIMONE, R., 2006a. Constructions and categories in verbal and signed languages. In: E. Pizzuto, P. Pietrandrea, and R. Simone, eds. *Verbal and Signed Languages. Comparing Structures, Constructs, and Methodologies*, Berlin-New York: Mouton De Gruyter, , pp. 198-252.

SIMONE, R. (2006b), Nominales sintagmáticos y no-sintagmáticos. In: E. De Miguel, A. Palacios and A. Serradilla, eds. *Estructuras léxicas y estructuras del léxico*, Berlin: Peter Lang, pp. 221-241.

SIMONE, R., MASINI, F., PIUNNO, V. AND CASTAGNOLI S., 2013. Combinazioni di parole in italiano: risorse lessicografiche e proposte di tipologia, Paper presented at *XLVII Congresso Internazionale SLI*, Salerno 26-28 September 2013.

TUTIN, A. AND GROSSMANN F., 2002. Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif, *Revue française de Linguistique appliquée*. 7(1). pp. 7-25.

VAN DER AUWERA, J. (ed.) 1998. *Adverbial Constructions in the Languages of Europe*. Berlin/New York: Mouton de Gruyter.

VAN VALIN, R. D. Jr. AND LA POLLA, R. J. (1997), *Syntax. Structure, meaning and function*, Cambridge: Cambridge University Press.

VOGHERA, M., 1994. Lessemi complessi: percorsi di lessicalizzazione a confronto. *Lingua e Stile*. 29 (2). pp. 185-214.

VOGHERA, M., 2004. Polirematiche. In: M. Grossmann and F. Rainer, eds. *La formazione delle parole in italiano*. Tubingen: Max Niemeyer Verlag. pp. 56-69.

VOGHERA, M. AND TURCO G., 2006. Il peso del parlare e dello scrivere. In: M. Pettorino, A. Giannini, M. Vallone and R. Savy, eds. *La comunicazione parlata*. Napoli: Liguori. I. pp. 727-760.

Corpora

LA REPUBBLICA, Available at:

http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica

CONTRIBUTION OF MULTI-ELEMENT FEATURES IN AUTOMATIC TEXT CLASSIFICATION FOR AUTHORSHIP ATTRIBUTION

Antonio Rico-Sulayes
Universidad de las Américas Puebla
antonio.rico@udlap.mx

Abstract

The use of multi-element units in computational linguistics has a long tradition that is reflected on both automatic text classification and the related forensic linguistic task of authorship attribution. Frequently performed by means of automated processes, the task of authorship attribution is aimed at assigning an anonymous piece of text to a subject within a list of potential authors. In order to achieve this goal, a constant proposal of new classificatory features has characterized the research on this task for a number of decades (Rudman, 1998). More recently, the number of features used in authorship attribution has exploded as all sorts of multi-element units have been introduced in the task, such as character, word and POS n-grams (Rico Sulayes, 2014). Easy to tag by computational tools, these multi-element units can produce long lists of features even in the small corpora –text collections authored by some set of subjects– which are standard in authorship attribution. The present study uses contributions to organized crime-related online forums to randomly create a number of corpora with an increasing number of subjects. Using these corpora to run several hundreds of experiments, two types of different classificatory features are tested: a rather short, previously selected list of multi-element function words and a large list with all word unigrams in a given corpus. The combined set of all these features is fed to a suite of the most common and successful machine learning classifiers in this task. As will be reported here, the best results averaged by some classifier over all corpora are obtained after reducing the list of features by statistical techniques. In the significantly reduced sets of features that render the best performance, there are only a few of the multi-element function words previously selected; however, a further analysis of the features selected from the list of unigrams shows that many of these elements are either part of or they represent themselves multi-element units from a phraseological point of view (Beck and Mel'čuk, 2011; Corpas, 2013; Shanavas, 1996).

1. INTRODUCTION

In forensic linguistics, assigning the authorship of an anonymous text to a subject, included in a set of potential authors, is usually performed through a comparison of the features of such text and the features of a number of documents authored by this set individuals. This task, which represents clearly a text classification problem, is known as authorship attribution. The text assignment just described has two main components. A rather language-related component is the selection of classificatory features. These features

are aimed at representing the authorship of the different subjects in the set. Therefore, the features should be as discriminatory as possible to set apart the different writing styles of the potential authors. The other component of this text assignment is the processing of classificatory features in order to make classification judgments when given an anonymous (or anonymized) text. These classification judgments, when generated by automated means, are essentially based on statistics and mathematical inferences.

For practical reasons, the first component of authorship attribution is feature selection; without features there is no information to process and make inferences about with a statistics-based classifier. Feature selection has been characterized in research literature by a constant increase in the number of features not only proposed but also used in a single experiment to represent the authorship of individuals. By the end of last century, Rudman (1998) found out about 1,000 different features proposed in more than 300 papers published over the three previous decades. At the same time, there has been a trend in research to move from univariate approaches, which looked for a single feature that globally represented the authorship of an individual, to multivariate approaches that combine information from the distributional patterns of various features for the same purpose (Juola, 2008). Also as computational power has become broadly available, the number of features used in a single experiment has increased dramatically over the last decade (Rico-Sulayes, 2012). To some extent this is the result of the wide use of what I have called elsewhere exhaustive feature lists. Exhaustive feature lists include direct representations of all the data in the text. Namely, if types (differentiated word forms) are chosen as a feature, all the words in a collection of texts are looked at and counted in order to feed their figures to a classifier. Typical examples of exhaustive lists are the various sorts of *n*-grams lists. These lists include all sequences with *n* number of elements (such as characters, words, or POS) in a text. Exhaustive lists usually produce large number of features, even with small text collections, like the ones that are characteristic of authorship attribution (Abbasi and Chen, 2005; Gamon, 2004; Koppel, Schler, and Argamon, 2009).

Since very long lists of features usually have the disadvantage of including noisy features, i.e., non-discriminative or redundant features (Manning, Raghavan, and Schütze, 2008), it is common in authorship attribution to reduce these lists by statistical techniques (Abbasi and Chen, 2008; Burrows, 2002; Chaski, 2005; 2007; Gamon, 2004; Grant, 2007; Hirst and Feiguina, 2007; Jockers, Witten, and Criddle, 2008; Koppel, Schler, and Argamon, 2009; Mikros and Argiri, 2007; Orebaugh and Allnutt, 2009; Peng, et al., 2003; Raghavan, Kovashka, and Mooney, 2010; Rico-Sulayes, 2011; Spassova, 2009; Tambouratzis and Vassiliou, 2007; Tearle, Taylor, and Demuth, 2008). These techniques, are aimed at improving the success rate in various classification tasks in computational linguistics. They have also the advantage of making easier to analyze and interpret the much smaller sets of features that produce the best results in a large number of experiments. This is exactly what this article does from a phraseological perspective. In a first analysis of the consistent, reduced set of features that produce the best text classification results in over 750 authorship attribution experiments, just a few predefined multiword units are selected; however, a further analysis of the rest of the most highly discriminatory features shows that many have a compositional quality that makes them also multi-element phrasemes, a number of which are especially suitable for automated tagging.

2. MULTI-ELEMENT FEATURES IN AUTOSHIP ATTRIBUTION

The use of features that are a compound of two or more elements has become widespread not only in the general field of computational linguistics, but also in the specific

context of authorship attribution (Corney, 2003; Gamon, 2004; Grieve, 2007; Hirst and Feiguina, 2007; Juola and Baayen, 2005; Koppel, Schler, and Argamon, 2009; Nazar and Sánchez Pol, 2007; Peng, et al., 2003; Raghavan, Kovashka, and Mooney, 2010; Rico-Sulayes, 2011; Spassova, 2009; Spassova and Turell, 2007). Given this popularity, these multi-element features can be very diverse in nature. On one extreme of the spectrum, there are multi-element features which are rather non-analyzable, i.e., they do not necessarily contain meaning on their own. A non-analyzable multi-element type of feature commonly used in authorship attribution is character n-grams (Corney, 2003; Grieve, 2007; Juola and Baayen, 2005; Peng, et al., 2003). For example, in the list of character bigrams for the word *character* ("ch", "ha", "ar", "ra", "ac", "ct", "te", "er"), none of its two-character sequences has some independent meaning. Even when some character sequence may have some meaning in a specific context (such as the last listed bigram, "er", in words like *greater* or *teller*), the majority of character n-grams obtained from a word do not have a morphological or semantic value.

On the other extreme of the various forms of multi-element features used in authorship attribution, word n-grams (usually just called n-grams) are the result of aggregating adjacent words in a text. In the sentence *The air conditioner is not working*, there are five word bigrams ("the air", "air conditioner", "conditioner is", "is not", "not working"). Word n-grams may correspond to the kind of multi-word units usually targeted in phraseological studies (Bergenholtz and Gouws, 2013; Corpas, in press), such as the word bigram "air conditioner" that is a fixed, idiomatic expression. Another type of multi-element unit commonly used in computational linguistics, and authorship attribution as well (Gamon, 2004; Hirst and Feiguina, 2007; Koppel, Schler, and Argamon, 2009; Spassova, 2009; Spassova and Turell, 2007), are POS n-grams. In this case, a sentence such as *What time is it?* is first tagged with its POS. Using the Brown tag set (Jurafsky and Martin, 2008) for example, the former sentence becomes "WDT NN BEZ PPS ." (with the period representing a tag). This renders the following four POS bigrams: "WDT NN", "NN BEZ", "BEZ PPS", "PPS .". Although there are other possibilities to employ multi-element features, the three just mentioned (character, word and POS n-grams) are the most commonly used in authorship attribution. The following three sections, then, discuss a number of phraseological constructs that will be helpful to interpret the most successful authorship attribution experiments later analyzed in this article.

2.1. Multi-word units in authorship attribution

As mentioned above, word n-grams are the multi-element features that are the closest to the type of multi-word lexical units commonly studied in phraseological research. These n-grams are also popular among authorship attribution studies (e.g., Grieve, 2007; Nazar and Sánchez Pol, 2007; Raghavan, Kovashka, and Mooney, 2010). Since many of the word n-grams in a text are intrinsically linked to its topic, this type of features have been successfully used in computational linguistics tasks for topic detection (Jurafsky and Martin, 2008). This very advantage, however, can be seen as a disadvantage in the context of authorship attribution, where a correlation between a topic and a given author could render a seemingly good classification result that would not be sustained once the author changes its writing topic. The problem of using features that represent parallel classes, such as topic in this case, and whose influence should not be carried over into the classification of relevant classes, such as the authorship of different individuals, has been targeted in a number of ways in authorship attribution. One response to this issue has been to identify and eliminate features that interact between topics and authors (Mikros and Argiri, 2007). Another solution has been an explicit attempt to control the topic that the individuals in the set of potential authors write about (Chaski, 2005). Finally, the most common response by researchers to this issue has been employing features that are generally considered topic

independent, such as function words (Abbasi and Chen, 2005; 2008; Argamon, Šarić, and Stein, 2003; Baayen et al., 2002; Corney, 2003; Gamon, 2004; Juola and Baayen, 2005; Koppel, Schler, and Argamon, 2009; Koppel, Schler, and Messeri, 2008; Mikros and Argiri, 2007; Zheng et al., 2003; 2006). This article combines this last solution with a multi-word element approach and uses a list of Spanish multi-unit function words.

The list of multi-unit function words used here is part of a Spanish lexicon collected during the summer of 2011 for various computational linguistics tasks. This list contains words from two Spanish closed class categories, i.e., whose members is relatively fixed: prepositions (such as *después de* 'after' and *lejos de* 'far from') and conjunctions (such as *después de que* 'after' and *mientras que* 'whereas'). The multi-word prepositions include combinations which can take different forms when they are contracted with articles (for example *de* 'of' and *del* 'of (the)', as in *después de* '*after*' and *después del* 'after (the)'). The list of multi-unit function words employed in the experiments here reported has a total of 131 multi-word sequences, which work together as individual lexical items. The multi-word sequences in this list consist of 68 bigrams, 56 trigrams, and 7 four-grams (i.e., sequences of two, three, and four words, respectively).

2.2. Morphological phrasemes

The former section has introduced the entire list of multi-word lexical units that will be used in the experiments reported in this article. However, as it will be exemplified later in the discussion of results, there are other type of multi-element items, considered also part of phraseological research, which will be helpful to interpret the experiments here conducted. This is the case of phraseological phrasemes. According to Beck and Mel'čuk (2011), although there is not a general acceptance of the concept of phraseologized expressions at the morphological level, recognizing the existence of conventionalized uses of sublexical elements that bear meaning and extending the concept of phraseologization to these elements allows researchers to explain a number of linguistic phenomena, especially in agglutinative languages. This will be particularly true in the interpretation of the highly discriminatory features used in the most successful authorship attribution experiments here reported. In this sense, this articles does something similar to Beck and Mel'čuk (2011), presenting a set of novel and unusual data that is better explained through the use of an original theoretical framework.

2.3. Reduplication in phraseology

Contrary to the construct of morphological phrasemes, the concepts of reduplication and repetition have a longer and more accepted tradition in phraseological studies (e.g., Shanavas, 1996; Teliya, 2002; Levin and Lindquist, 2013). Productive at several different levels (such as in Mexican Spanish multi-word expressions *casi casi* "almost (intensified)", *luego luego* "immediately" or *mero mero* / *mera mera* "the boss"), this concept will be combined with that of morphological phrasemes to explain a number of authorship features that are highly discriminatory in the experiments presented. Also, a number of features for reduplicated punctuation first proposed in Rico-Sulayes (2011) will be explained as phraseological units. Since punctuation marks are standardly split from the words they attach to and are treated as individual lexical items in computational linguistics (Jurafsky and Martin, 2008), the strings of reduplicated punctuation marks (such as "!!", "!!!", "!!!!", an so forth) will be treated here as phraseological units. Expressed in an abstract way as any string of two, three, four, five, six, and seven or more reduplicated punctuation marks, these phraseological features will be shown to be highly efficient in differentiating individuals for their authorship in the context of the data presented in the next section.

3. EXPERIMENTAL CORPORA

The experiments conducted and interpreted in this article have been run on data harvested from what was one of the first online forums devoted to the topic of organized crime in Mexico (Foros Blog del Narco, 2010). A phenomenon that first started in the year 2010 and rapidly became popular (Rico-Sulayes, 2011), this type of forums has a complex history that later resulted in the murder of some of their users (Rico-Sulayes, 2012). The online forum retrieved here was created in April 2010 and became allegedly the most popular forum in that year. Six months after its creation, 41,751 users' contributions were spidered from the forum. Once this data was collected, its content was processed and cleaned to obtain a total of 37,571 contributions posted by 1,026 logged users with a total 2,128,049 word tokens. With this processed information, 39 corpora were built, each with an increasing number of authors, from two to forty. These corpora include contributions from prolific users, who posted a minimum of 40 contributions, with at least 2,000 words of original text. Randomly sampling these contributions, a total of four text samples by each individual were put together to perform a maximum of 160 text classification assignments in the data set with 40 authors.

4. TEXT CLASSIFICATION METHODOLOGY

As mentioned in Section 1, the second component of an authorship attribution approach consists of processing a number of selected features to classify anonymous texts as belonging to an individual in a set of potential authors. In order to perform this classification task, a number of statistics-based and machine learning algorithms have been used in authorship attribution research. In this article, the most widespread classification algorithms in experimental studies (the decision tree C4.5, discriminant analysis (DA), multivariate naïve Bayes (MNB), the Bernoulli model of naïve Bayes (BNB), and support vector machines (SVMs)) are tested in combination with some of the most common techniques to reduce long sets of features and improve classification results (frequency, information gain (IG), and correlation-based feature subset selection (CFS)) (Rico-Sulayes, 2014). Given the combination of five classifiers and four sets of features (three reduced ones and one with all features included), there are twenty possible combinations of a classifier and a feature set. Applying these twenty configurations to the 39 corpora introduced in the last section, a total of 780 authorship attribution experiments have been conducted.

5. CLASSIFICATION RESULTS

The best averaged result over all 39 corpora was obtained with the MNB classifier and the feature set reduced with IG. This combination of a classifier and a feature set obtained an accuracy of 0.947. Probably the most important aspect for the current discussion regarding this very competitive figure in authorship attribution research (Rico-Sulayes, 2012) is that this number was obtained using a very small set of features. As mentioned in Section 1, having small feature sets makes it much easier to analyze their information from a linguistic perspective, including a phraseological point of view. Following this idea, it should be emphasized that all four classifiers obtained their best results when given a reduced set of features. This can be seen in Table 1 below.

Feature set	Classifier				
	C4.5	DA	MNB	BNB	SVMs
No reduction	0.456	0.313	0.743	0.831	0.457
Freq.	0.489	0.466	0.942	0.821	0.829
IG	0.671	0.700	0.947	0.848	0.775
CFS	0.660	0.811	0.940	0.820	0.726

Table 1. Averaged results over 39 corpora by classifier and feature set

Another important result that can be seen in Table 1 is that all but one of the classifiers (SVMs) obtained their highest accuracy when fed the feature set reduced with either IG or CFS. Furthermore, even in the case of SVMs, this classifier accuracy improves dramatically with the sets reduced by these two techniques compared to the set with all features. Observing the effect of these two feature reduction techniques is particularly important because they consistently produced very small feature sets over all 39 corpora. Compared to the full set of features, which has in the case of the largest corpora with 40 authors a total of 13,249 features, the entire collection of feature sets selected by IG and CFS in all corpora have just 144 distinctive features. These features are selected in a fairly consistent way across all experiments. With a rather small set of features that consistently produced the best results over all experiments, it is then much easier to make an analysis of the kind of linguistic information utilized in the most successful authorship attribution experiments.

5.1. Highly discriminant features

An analysis of the linguistic patterns in the 144 features that render the best results shows that a great proportion of them do not represent content specific information, necessarily linked to the topic or some other parallel class. As mentioned in Section 2.1, this kind of information is seen as a disadvantage in authorship attribution, since the benefit of carrying it over to authorship will disappear once an author changes the topic or some other parallel class in question. In the global set of 144 features selected by IG and CFS, there are 9 features that represent conventions in the structure of conversation (such as *bienvenido* 'welcome' and *saludos* 'regards'), 22 features are forms that convey emotional content (such as the forms *ja* and *jaja*, both meaning 'hehe', or the English borrowing *wow*), 13 lexical items represent forms of swearing or euphemisms (such as *cabron* [sic] 'fucker' and *wey* [sic] 'dude'), 5 features are netabbrevs -- lexical items that are developed in communities that use electronic communications -- and function words at the same time (such as *ke* for *que* 'that' and *x* for *por* 'for'), 17 are punctuation-related marks (such as periods, commas, and exclamation points), 23 are function words (including prepositions, determiners, conjunctions, and pronouns), 11 are adverbs (e.g., *más* 'more' and *aquí* and *aqui* [sic] 'here'), and 7 are structural features related to the means of communication (such as the use of a bold font or a font with a special size). Besides this first linguistic classification, there are also a number of these features that represent phraseological multi-element units.

5.1.1. Multi-element features for best results

As mentioned in section 2.1, a list of 131 multi-word sequences that work together as individual function words was included in the experiments. These multi-word sequences included 68 bigrams, 56 trigrams, and 7 four-grams. A first look at the 144 features consistently selected in the most successful experiments does not seem especially productive in term of phraseological features. Only 2 out of the list of 131 multi-unit function words (*por que* 'because of which' and *para que* 'for which') are among the 144

features selected by IG and CFS. However, a further analysis of the rest of these selected features reveals that there are several other features that can be considered as phraseological units. All the features for reduplicated punctuation, for series of two, three, four, five, six, and seven or more repeated and aggregated punctuation marks were selected. These features in particular were encoded in a system that automatically tagged all the corpora and were then an integral part of the search for phraseological, multi-element units. In contrast with these features, there were other forms that also result from a reduplication (of morphemes in this case), but were not intentionally tagged for. For example, a series of forms that reduplicate the forms *ja*, *je*, and *ha* [sic] 'he' (all of them onomatopoeias of laughter) include jaja, jajajajajaja, jeje, *hahahah*, and *hahahaha*. Therefore, a closer look at the features selected by IG and CFS shows that there are many more phraseological units than the few ones picked from the predefined set of n-grams.

6. CONCLUSION

In an attempt to avoid using features that correlated with topic, a parallel class that is usually controlled for in authorship attribution, the methodological approach here outlined used a list of multi-word features that work as function words. Only few items in this rather restricted list were selected by the feature reduction techniques that rendered the best results across all the experiments conducted. However, a more careful analysis of the reduced feature sets, combined with a novel phraseological perspective, shows that there are several more phraseologized multi-element features that contribute to the comparatively very good results (Rico-Sulayes, 2012) obtained in the text classification tasks here presented. Furthermore, many of these multi-element features are especially suited to be detected by more abstract, computational approaches, as it was the case with the engineering of reduplicated punctuation features. The combination of approaches to explain the phraseologization of the features presented in the section above echoes the idea in Beck and Mel'čuk (2011) that a set of novel and unusual data can be better explained and interpreted through the use of an original theoretical framework.

References

ABBASI, A., AND CHEN, H., 2005. Applying Authorship Analysis to Extremist-Group Web Forum Messages. *IEEE Intelligent Systems*, 20(5), pp. 67-75.

ABBASI, A., AND CHEN, H., 2008. Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection. *ACM Transactions on Information Systems*, 26(2), pp. 1-29.

ARGAMON, S., ŠARIĆ, M., AND STEIN, S.S., 2003. Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

BAAYEN, H., VAN HALTEREN, H., NEIJT, A., AND TWEEDIE, F., 2002. An Experiment in Authorship Attribution. *Proceedings of JADT 2002: Sixth International Conference on Textual Data Statistical Analysis*, pp. 29-37.

BECK, D. AND MEL'ČUK, I. A., 2011. Morphological phrasemes and Totonacan verbal morphology. *Linguistics*, 49, pp. 175-228.

BERGENHOLTZ, H, AND GOUWS, R., 2013. A Lexicographical Perspective on the Classification of Multiword Combinations. *International Journal of Lexicography*, 27(1), pp. 1-24.

BURROWS, J., 2002. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3), pp. 267-86.

CHASKI, C.E., 2005. Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence*, 4(1), pp. 1-13.

CHASKI, C.E., 2007. The Keyboard Dilemma and Authorship Identification. In *Advances in Digital Forensics III*, eds P. Craiger and S. Shenoi, Ney York, NY, Springer, pp. 133-146.

CORNEY, M., 2003. *Analysing E-mail Text Authorship for Forensic Purposes*. MA thesis, Queensland University of Technology. Available at:http://eprints.qut.edu.au/16069/1/Malcolm_Corney_Thesis.pdf [Accessed 27 July 2015].

CORPAS PASTOR, G., 2013. Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In *Fraseopragmática*, eds I. Olza and E. Manero, eds Berlin, Frank & Timme. pp. 335-373.

CORPAS PASTOR, G., in press. Collocations in e-bilingual dictionaries: from underlying theoretical assumptions to practical lexicography and translation issues. In *Collocations and other lexical combinations in Spanish. Theoretical and Applied approaches*, eds S. Torner and E. Bernal, Chicago, IL, Ohio State University.

FOROS BLOG DEL NARCO, 2010. Available from: http://www.foro.blogdelnarco.com/. [12 September 2010].

GAMON, M., 2004. Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features. *Proceedings of the 20th International Conference on Computational Linguistics*, vol. 4, pp. 611-617. Stroudsburg, PA: Association for Computational Linguistics.

GRANT, T., 2007. Quantifying Evidence in Forensic Authorship Analysis. *International Journal of Speech, Language and the Law*, 14(1), pp. 1-25.

GRIEVE, J., 2007. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3), pp. 425-442.

HIRST, G., AND FEIGUINA, O., 2007. Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4), pp. 405-417.

JOCKERS, M.L., WITTEN, D.M., AND CRIDDLE C.S., 2008. Reassessing Authorship of the Book of Mormon Using Delta and Nearest Shrunken Centroid Classification. *Literary and Linguistic Computing*, 23(4), pp. 465-491.

JUOLA, P., 2008. *Authoship Attribution*. Now Publishers, Hanover, MA.

JUOLA, P., AND BAAYEN, H., 2005. A Controlled –Corpus Experiment in Authorship Attribution by Cross-Entropy. *Literary and Linguistic Computing*, 20(1), pp. 59-67.

JURAFSKY, D., AND MARTIN, J.H., 2008. *Speech and Language Processing: An Introduction to Language Natural Processing, Computational Linguistics, and Speech Recognition*, 2nd edn, Pearson-Prentice Hall, Upper-Saddle River, NJ.

KOPPEL, M., SCHLER, J., AND ARGAMON, S., 2009. Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, 60(1), pp. 9-26.

KOPPEL, M., SCHLER, J., AND MESSERI, E. (2008). Authorship Attribution in Law Enforcement Scenarios. In *Security Informatics and Terrorism: Patrolling the Web*, eds C.S. Gal, P. Kantor, and B. Saphira, IOS, Amsterdam, pp.111-119.

LEVIN, M., AND LINDQUIST, H., 2013. Like I said again and again and over and over: On the ADV1 and ADV1 construction with adverbs of direction in English. *International Journal of Corpus Linguistics*, 28, pp. 7-34.

MANNING, C.D., RAGHAVAN, P., AND SCHÜTZE, H., 2008. *Introduction to Information Retrieval*. New York, NY: Cambridge.

MEL'ČUK, I. A., 2004. La non-compositionnalité en morphologie linguistique. *Verbum*, 26, pp. 439-458.

MIKROS, G.K., AND ARGIRI, E.K., 2007. Investigating Topic Influence in Authorship Attribution. In *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*, eds B. Stein, M. Koppel, and E. Stamatatos, *PAN 2007*.

NAZAR, R., AND SÁNCHEZ POL, M., 2007. An Extremely Simple Authorship Attribution System, in *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics / Language and the Law 2006*, eds M.T. Turell, J. Cicres, and M.S. Spassova, Documenta Universitaria, Barcelona.

OREBAUGH, A., AND ALLNUTT, J., 2009. Classification of Instant Messaging: Communications for Forensics Analysis. *The International Journal of Forensic Computer Science*, 4(1), pp. 22-28.

PENG, F., SCHUURMANS, D., KESELJ, V., AND WANG, S., 2003. Language Independent Authorship Attribution Using Character Level Language Models. *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, vol. 1, pp. 267-274. Stroudsburg, PA: Association for Computational Linguistics.

RAGHAVAN, S., KOVASHKA, A., AND MOONEY, R., 2010. Authorship Attribution Using Probabilistic Context-Free Grammars. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 38-42.

RICO-SULAYES, A., 2011. Statistical Authorship Attribution of Mexican Drug Trafficking Online Forum Posts. *International Journal of Speech, Language and the Law,* 18(1), pp. 53-74.

RICO-SULAYES, A., 2012. *Quantitative Authorship Attribution of Users of Mexican Drug Deal-ing Related Online Forums*. PhD thesis, Georgetown University.

RICO SULAYES, A., 2014. Técnicas de reducción, algoritmos resistentes al ruido o ambos. Opciones para el manejo de rasgos clasificatorios en la atribución de autoría. *Research in Computing Science*, 80, pp. 43-53.

RUDMAN, J., 1998. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31, pp. 351-365.

SHANAVAS, S.A., 1996. *Structure of a computational lexicon of Malayalam*. Ph. D thesis, Jawaharlal Nehru University. Available at:http://shodhganga.inflibnet.ac.in/handle/10603/35036 [Accessed 28 February 2015].

SPASSOVA, M.S., 2009. *El Potencial Discriminatorio de las Secuencias de Categorías Gramaticales en la Atribución Forense de Autoría de Textos en Español*. PhD thesis, Universitat Pompeu Fabra.

SPASSOVA, M.S., & TURELL, M.T., 2007. The Use of Morphosyntactically Annotated Tag Sequences as Markers of Authorship, in *Proceedings of the 2nd European IAFL Conference on Forensic Linguistics / Language and the Law 2006*, eds M.T. Turell, J. Cicres, and M.S. Spassova, Documenta Universitaria, Barcelona, pp. 229-237.

TAMBOURATZIS, G., & VASSILIOU, M., 2007. Employing Thematic Variables for Enhancing Classification Accuracy within Author Discrimination Experiments. *Literary and Linguistic Computing*, 22(2), pp. 207-224.

TEARLE, M., TAYLOR, K., AND DEMUTH, H., 2008. An Algorithm for Automated Authorship Attribution Using Neural Networks. *Literary and Linguistic Computing*, 23(4), pp. 251-270.

TELIYA, V., BRAGINA, N., OPARINA, E., AND SANDOMIRSKAYA, I., 2002. Phraseology as a Language of Culture: Its Role in the Representation of a Collective Mentality. In Phraseology: Theory, Analysis, and Applications, ed A.P. Cowie, Oxford University, New York, NY, pp. 55-75.

ZHENG, R., LI, J., CHEN, H., AND HUANG, Z., 2006. A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. *Journal of the American Society for Information Science and Technology*, 57(3), pp. 378-393.

ZHENG, R., QIN, Y., HUANG, Z., AND CHEN, H., 2003. Authorship Analysis in Cybercrime Investigation. *Proceedings of the 1st NSF/NIJ Symposium, ISI2003*, Springer-Verlag, Berlin, pp. 59-73.

LA MARCACIÓN DE LAS UNIDADES FRASEOLÓGICAS A PARTIR DEL EXAMEN DE CORPUS[190]

Ana María Ruiz Martínez
Universidad de Alcalá
ana.ruiz@uah.es

Resumen

La falta de sistematicidad que caracteriza a algunos diccionarios a la hora de informar sobre las particularidades del uso de la lengua española la hemos observado de manera específica con el empleo de la marca *lit* (literario), a partir del análisis que hemos llevado a cabo de diferentes clases de unidades fraseológicas extraídas del *Diccionario fraseológico documentado del español actual* (Seco, Andrés y Ramos 2004). El hecho de que este diccionario ofrezca, mediante la citada marca, información sobre el nivel de uso formal y elevado de las unidades fraseológicas (información diafásica), además de otras indicaciones de distinta naturaleza (uso propio de obras literarias o de la lengua escrita), y el hecho de que para un buen número de ejemplos de unidades no exista una correspondencia con las informaciones ofrecidas por otros diccionarios, nos ha llevado a organizar nuestra investigación en torno a dos objetivos principales: 1) revisar los problemas que ocasiona la marca *literario*; 2) y exponer por qué los corpus de lengua deben convertirse en una herramienta de consulta imprescindible para lexicógrafos, lingüistas, traductores, fraseólogos o profesores de idiomas, a la hora de precisar cuáles son las condiciones de uso que presentan algunas unidades fraseológicas.

1. INTRODUCCIÓN

Los diccionarios proporcionan información acerca de las particularidades o restricciones en el uso de las unidades fraseológicas por medio de etiquetas o marcas[191] que suelen acompañar a la definición. Por esta razón, tanto los diccionarios generales como los fraseológicos son un instrumento de consulta obligada para que el traductor, el estudiante o el usuario de una lengua conozca el rasgo formal, informal o vulgar que tiene una determinada unidad fraseológica y que condiciona su funcionamiento en el discurso.

[190] En el texto que hemos preparado ofrecemos un resumen de la comunicación presentada en el Congreso Internacional de la Sociedad Europea de Fraseología (EUROPHRAS 2015), celebrado en Málaga del 29 de junio al 1 de julio de 2015.

[191] Siguiendo a Fajardo (1996-1997: 32), a lo largo de este trabajo hablaremos de *marcas* para referirnos a "las informaciones concretas sobre los muy diversos tipos de particularidades que restringen o condicionan el uso de las unidades léxicas". Para la distinción y revisión de las marcas identificadas en los artículos lexicográficos de los diccionarios del español, pueden consultarse, entre otros, los trabajos de Garriga (1997) o Porto Dapena (2002: 251-265).

Sin embargo, ni en los prólogos[192] de los propios diccionarios ni en los trabajos que se ocupan de su redacción es habitual encontrar una explicación detallada de los criterios que siguen los lexicógrafos para asignar las marcas correspondientes a cada unidad, así como el alcance que tiene cada una de ellas[193], lo que origina que la atribución y caracterización de ciertas marcas lexicográficas sea un asunto no exento de dificultades. Si a esto añadimos que algunos lexicógrafos se limitan a reproducir una marca de acuerdo con una tradición en la que esta no queda definida de una manera clara[194] o recoge diferentes contenidos[195], resulta evidente concluir que la concreción de una teoría que guíe la utilización de las marcas de uso en los diccionarios sigue siendo todavía un aspecto pendiente para la lexicografía en lengua española.

Esta falta de sistematicidad e imprecisión que caracteriza a la práctica lexicográfica en el empleo de algunas marcas de uso la hemos observado de manera específica con la marca *literario* (*lit*), a partir del análisis que hemos llevado a cabo de diferentes clases de unidades fraseológicas extraídas del *Diccionario fraseológico documentado del español actual* (Seco, Andrés y Ramos 2004). El hecho de que este diccionario ofrezca, mediante la citada marca, información sobre el estilo formal vinculado con las unidades fraseológicas así marcadas (información diafásica), además de otras indicaciones de distinta naturaleza (uso propio de obras literarias o de la lengua escrita), y el hecho de que para un buen número de ejemplos de unidades fraseológicas no encontremos una correspondencia con las informaciones ofrecidas por otros diccionarios de la lengua española, nos ha llevado a organizar nuestra investigación en torno a dos objetivos principales:

1) revisar los problemas que ocasiona la utilización de la marca *literario* en virtud de los diferentes valores que puede aglutinar, y a tenor de la falta de sistematicidad con la que los diccionarios presentan esas restricciones en su microestructura y la heterogeneidad de las informaciones que contienen;

2) y exponer por qué los corpus de lengua se han convertido en una herramienta de consulta imprescindible para lexicógrafos, lingüistas, traductores, fraseólogos o profesores de idiomas, a la hora de informar sobre las particularidades en el uso que presentan algunas unidades fraseológicas, dado que el rápido acceso a una ingente cantidad de datos pone a nuestra disposición la posibilidad de realizar un análisis de las situaciones comunicativas, los tipos de textos, los discursos, etc., que facilitan la ocurrencia de una determinada unidad.

[192] Casares (1969: XV) llama la atención sobre la importancia de los prólogos como depositarios de información sobre la labor lexicográfica llevada a cabo en los diccionarios modernos.

[193] En la mayoría de los diccionarios no suele explicarse el significado de las abreviaturas con las que se marcan las unidades, lo que impide la comprensión del sistema de marcación utilizado por el diccionario.

[194] De acuerdo con Fajardo (1996-1997: 43), la marca *fam.* (familiar) es muy abundante en la lexicografía española y, sin embargo, no se encuentra una definición sobre ella en los sistemas de marcación de los diccionarios. Y esto llama todavía más la atención si tenemos en cuenta que la indicación de *familiar* está ya contenida en el *Diccionario de Autoridades*. Esta manera de proceder puede explicarse por el hecho de que las obras lexicográficas "toman siempre como punto de partida las ya existentes", de manera que la tradición lexicográfica desempeña un papel fundamental en cualquier estudio sobre diccionarios (Fernández Sevilla 1974: 159).

[195] Por ejemplo, la marca *vulgar* se utiliza tanto para las voces utilizadas por los estratos socioculturales de escasa instrucción (marca diastrática), como para las que resultan malsonantes (marca diafásica).

2. LA MARCA *LITERARIO* EN LOS DICCIONARIOS GENERALES Y FRASEOLÓGICOS

Tal y como acabamos de adelantar, nuestro interés por la marca *literario* surge a raíz del empleo que de ella hace el *Diccionario fraseológico documentado del español actual*. En el "Glosario de términos lingüísticos utilizados" (pp. XXIX-XXXII) los autores indican el valor con el que utilizan la citada marca mediante la siguiente información:

> **literario** (***lit***) Uso propio, en general, de obras literarias, o de la lengua escrita, o de situaciones formales; en especial, de la expresión voluntariamente elegante o elevada.

Después de leer este texto, tenemos la impresión de que con la marca *literario* no se acota de una manera firme o clara la información que propone el diccionario, puesto que, al hacer referencia a diferentes restricciones, provoca cierta imprecisión en la información ofrecida en las unidades marcadas mediante la abreviatura *lit*[196], de manera que cuando nos encontramos ante las locuciones *a la mano* o *en extremo*, por ejemplo, no sabemos exactamente si la locución en cuestión es propia de un estilo de lengua formal, de obras literarias o de la modalidad escrita.

Esta falta de precisión en el contenido que comunica la aparición de la marca *literario* con algunas unidades fraseológicas, nos ha llevado a realizar un estudio comparativo con la información proporcionada por otros diccionarios de la lengua española para un corpus de 130 locuciones adverbiales que aparecen seguidas de la abreviatura *lit* en el *Diccionario fraseológi*co. A partir del corpus de unidades seleccionadas, hemos revisado si otros diccionarios ofrecen también restricciones sobre su uso y qué marcas emplean para ello. Para llevar a cabo esta labor, hemos tenido en cuenta los prólogos, las listas de abreviaturas y la información que aparece formando parte de la definición de la locución examinada. Y todo ello sin perder de vista que la comparación entre los sistemas de marcación que sigue cada diccionario no está exenta de obstáculos, puesto que este tipo de obras, en función de sus objetivos, pueden emplear marcas diferentes y utilizar sistemas de marcación de distintos grados.

Los diccionarios que hemos considerado para esta fase de la investigación han sido ocho: seis diccionarios generales y dos fraseológicos. Los diccionarios generales son: *Diccionario de la lengua española* (Real Academia Española 2014), *Diccionario esencial Santillana de la lengua española* (Sánchez Cerezo (dir.) 1994), *Diccionario para la enseñanza de la lengua española* (Moreno Fernández (dir.) 1995), *Diccionario Salamanca de la lengua española* (Gutiérrez Cuadrado (dir.) 1996), *Diccionario de uso del español* (Moliner 1966) y el *Gran diccionario de uso del español actual* (Sánchez Pérez (dir.) 2001). En cuanto a los diccionarios fraseológicos, hemos consultado el *Diccionario de locuciones adverbiales para la enseñanza del español* (Penadés 2005) y el *Diccionario fraseológico del español moderno* (Varela y Kubarth 1994).

Tras el análisis de la información ofrecida por los diccionarios, hemos obtenido los siguientes resultados:

A) Los únicos diccionarios que recogen las abreviaturas *lit*., *liter*., LIT O LIT, con su equivalencia (*literario*), o la marca *literario* son: el *Diccionario de uso del español*, el *Diccionario*

[196] Esta situación es un reflejo de los distintos valores con los que la marca ha sido utilizada habitualmente por la lexicografía española. En Fajardo (1996-1997: 36), donde podemos revisar los principales tipos de oposiciones que mantienen las marcas que forman parte de algunos microsistemas, con la etiqueta *literario* el uso lingüístico queda relacionado tanto con la lengua escrita como con los textos literarios: "[…] 3) neutro // hablado / escrito: *lit.* […] 6) neutro // poético / literario / periodístico / administrativo: *Adm., lit., Poét.*"

esencial Santillana de la lengua española, el *Diccionario para la enseñanza de la lengua española*, el *Diccionario Salamanca de la lengua española* y el *Gran diccionario de uso del español actual*. Al examinar si estos diccionarios contenían las locuciones adverbiales seleccionadas, hemos comprobado que:

1.- Son pocas las unidades para las que los diccionarios indican algún tipo de particularidad vinculada con el uso de las mismas, tal y como puede comprobarse en la información que ofrecemos a continuación, donde aparece para cada diccionario el número de las locuciones no registradas, las locuciones recogidas sin ninguna marca y las locuciones documentadas y con restricción.

	No registradas	Registradas y sin marca	Registradas y con restricción
Diccionario de uso del español	62	57	11
Diccionario esencial Santillana de la lengua española	90	37	3
Diccionario para la enseñanza de la lengua española	118	11	1
Diccionario Salamanca de la lengua española	94	20	16
Gran diccionario de uso del español actual	74	55	1

Para las 24 locuciones adverbiales documentadas con algún tipo de restricción se han empleado las siguientes marcas:

a) *Diccionario de uso del español*: culto, familiar, literario, popular y Teatro.
b) *Diccionario esencial Santillana de la lengua española*: culto, familiar y literario.
c) *Diccionario para la enseñanza de la lengua española*: formal.
d) *Diccionario Salamanca de la lengua española*: coloquial, elevado y literario.
e) *Gran diccionario de uso del español actual*: coloquial.

De este listado de marcas y de su aplicación, queremos resaltar dos hechos principales:

1.- La casi inexistente correspondencia que hay entre los diferentes diccionarios a la hora de informar sobre el uso de una locución, teniendo en cuenta el contenido que abarca la marca *literario* en el *Diccionario fraseológico*. Esta coincidencia la hemos observado en un solo ejemplo:

- *de hinojos*: lengua literaria (*Diccionario de uso del español*, *Diccionario esencial Santillana de la lengua española* y *Diccionario Salamanca de la lengua española*).

En otros casos, la correspondencia ha sido manifestada solamente por un diccionario:

- *por ende*: lengua culta (*Diccionario esencial Santillana de la lengua española*).

2.- Llama la atención encontrar locuciones marcadas por algunos diccionarios con una restricción que difiere del contenido incluido por la marca *literario*:

- *muy mucho*: familiar (*Diccionario de uso del español)* y coloquial (*Diccionario Salamanca de la lengua española* y *Gran diccionario de uso del español actual*).

B) En cuanto al grupo de diccionarios que no emplean la marca *literario*, los resultados proporcionados con relación al número de las locuciones no registradas, las locuciones recogidas sin ninguna marca y las locuciones documentadas y con restricción, son los siguientes:

	No registradas	Registradas y sin marca	Registradas y con restricción
Diccionario de la lengua española	49	67	14
Diccionario de locuciones adverbiales para la enseñanza del español	93	35	2
Diccionario fraseológico del español moderno	102	0	28

De las 40 locuciones adverbiales documentadas y con restricción, las marcas que las acompañan son las que ofrecemos a continuación:

a) *Diccionario de la lengua española*: coloquial, culto y Teatro.
b) *Diccionario fraseológico del español moderno*: formal e informal.
c) *Diccionario de locuciones adverbiales para la enseñanza del español*: informal.

Tras la revisión de este listado de marcas y de su aplicación, podemos afirmar que:

1.- Tan solo el *Diccionario de la lengua española* muestra correspondencia en una locución con uno de los contenidos que abarca la marca *literario*:

- *al paño*: Teatro.

2.- Vuelve a llamar la atención la ocurrencia de locuciones marcadas con una particularidad que nada tiene que ver con la información proporcionada por el *Diccionario fraseológico*:

- *a humo de pajas*: coloquial (*Diccionario de la lengua española*) e informal (*Diccionario de locuciones adverbiales para la enseñanza del español*).

C) Tras los resultados presentados en A) y en B), podemos concluir que:

1.- Solamente en dos locuciones (*al paño* y *de hinojos*) aparece explícitamente en los diccionarios la marca *literario* para indicar que la acepción de la locución es propia de los textos literarios.

2.- En cuanto al número de locuciones relacionadas con un estilo de lengua culta (diafasía), este asciende a nueve: *a ultranza* o *en gran manera.*

3.- La misma frecuencia absoluta se ha documentado también para las locuciones vinculadas con un estilo de lengua que no tiene nada que ver con el contenido que alberga la marca *literario*: *a humo de pajas* o *muy mucho* (familiar, informal y coloquial).

4.- Por último, conviene hacer alusión al elevado número de locuciones que no han sido registradas por los diccionarios, o para las que estos, aun documentándolas, no las acompañan de ningún tipo de restricción: *a buen seguro* o *de buena mañana.*

3. LA MARCACIÓN DE LAS UNIDADES FRASEOLÓGICAS A PARTIR DEL EXAMEN DE CORPUS

Para resolver esta falta de homogeneidad en la información proporcionada por los diccionarios examinados, consideramos que son de consulta obligada aquellos corpus que recolectan grandes muestras de la lengua oral y escrita, puesto que nos permiten esclarecer qué tipo de textos y qué situaciones comunicativas favorecen la ocurrencia de una determinada locución. En nuestra investigación, los textos y contextos examinados proceden de tres corpus: el *Corpus de Referencia del Español Actual* (CREA) y el *Corpus Diacrónico del Español* (CORDE) del banco de datos de la Real Academia Española[197], y el corpus DAVIES/NEH de la Universidad de Brigham Young (Utah)[198]. Los criterios manejados para la selección, ordenación y etiquetado del material que presentan los tres corpus pone a nuestra disposición muestras totalmente representativas de la lengua española, puesto que los textos pertenecen a la lengua escrita y oral, se integran en diferentes géneros y subgéneros (novela, teatro, entrevista, etc.) y en ellos se desarrollan diferentes temáticas (política, ocio, asuntos de la vida diaria, etc.). Y la utilidad de estas tres herramientas lingüísticas en nuestra investigación queda justificada porque pensamos que el análisis de los diferentes textos y contextos en los aparece una locución en cuestión es fundamental a la hora de tomar decisiones sobre el tipo de lengua (oral o escrita), el tipo de texto (literario o no) y el estilo de lengua con el que puede quedar restringido el uso de una determinada unidad.

Por lo que se refiere a los textos que han configurado la muestra que nos ha servido de base para el estudio de las locuciones adverbiales, queremos insistir en lo siguiente:

a) todos los textos que hemos utilizado los hemos restringido al español de España y al siglo XX,

b) no hemos partido de un porcentaje previo de tipologías textuales, pues hemos analizado todos los textos proporcionados por los corpus en los que se documentaba la locución en cuestión,

c) y el fragmento que hemos seleccionado de cada texto responde a una extensión diferente. Lógicamente, este hecho es consecuencia de los criterios manejados en la confección particular de cada corpus.

La consulta en los corpus de los cuatro grupos de locuciones adverbiales que hemos obtenido y presentado en el apartado anterior, evidencia que:

[197] Los dos corpus son complementarios. El CREA es representativo del estado actual de la lengua, y cuenta con más de 160 millones de formas procedentes de textos fechados a partir de 1975. Los textos escritos, tomados de libros, periódicos y misceláneas (prospectos, correos electrónicos, blogs, etc.) abarcan materias muy distintas (salud, artes, finanzas, etc.). La lengua hablada, por su parte, está representada por transcripciones de documentos sonoros, que provienen, mayoritariamente, de la radio y la televisión. A ellos se unen también materiales procedentes de discursos políticos, conversaciones telefónicas, diálogos informales, etc. En cuanto al CORDE, este corpus recoge en la actualidad 250 millones de registros desde los orígenes de la lengua hasta 1974, y contiene textos escritos de muy diferente género (en prosa y en verso): textos narrativos, líricos, jurídicos, religiosos, etc. Los dos corpus se encuentran disponibles en la página electrónica de la Real Academia Española (<http://www.rae.es>).

[198] Este corpus ha sido creado por Mark Davies y contiene más de 100 millones de palabras. Los 20 millones que pertenecen al siglo XX provienen de diversas fuentes: enciclopedias, periódicos, literatura y textos orales. El corpus está disponible y es de acceso gratuito en la siguiente página electrónica: <http://corpusdelespanol.org>.

1.- Con relación a las dos locuciones que aparecen acompañadas en algunos diccionarios por la marca *literario* para indicar que la acepción es propia de los textos literarios, los corpus manejados indican que:

- *al paño* y *de hinojos* son locuciones propias de la lengua literaria en la que está presente el estilo formal.

2.- En cuanto al grupo de locuciones relacionadas con un estilo de lengua elevada (diafasía), la contextualización de estas locuciones en los corpus pone de manifiesto que:

- *en gran manera* se documenta en la lengua escrita en la que hay un estilo formal,
- *por ventura* es propia de la lengua literaria en la que domina el estilo neutro,
- *en derredor* aparece en la lengua literaria cuando el estilo es formal.

3.- Aunque ha resultado muy significativo el número de locuciones vinculadas con un estilo de lengua que se aleja del contenido que alberga la marca *literario* en el *Diccionario fraseológico*, los resultados ofrecidos por los corpus han sido heterogéneos:

- *a humo de pajas* es propia del estilo informal y de la lengua escrita y oral,
- *muy mucho* se vincula con el estilo neutro y con la lengua escrita y oral.

4.- Y por último, con relación al elevado número de locuciones que no han sido registradas por los diccionarios o para las que estos, aun documentándolas, no las acompañan de ningún tipo de restricción, su documentación en los corpus pone de manifiesto que:

- las locuciones se utilizan tanto en la lengua escrita como en la lengua hablada, y estas son propias de un estilo de lengua neutro (*de buena mañana*) o informal (*como un forzado*),
- otras locuciones se corresponden con alguno de los contenidos de la marca *literario* en el *Diccionario fraseológico*: lengua literaria y estilo formal (*a banderas desplegadas*) o lengua escrita y estilo formal (*a todo evento*).

4. CONCLUSIONES

De la investigación que acabamos de presentar se desprende, en primer lugar, que la información proporcionada por diferentes diccionarios sobre las 130 locuciones adverbiales examinadas no tiene demasiada correspondencia con el contenido que encierra la marca *literario* en el *Diccionario fraseológico documentado del español actual:* uso de obras literarias, de la lengua escrita o de situaciones formales.

En segundo lugar, consideramos que el manejo de los corpus CREA, CORDE y DAVIES ha sido una herramienta muy útil al permitirnos esclarecer cómo se emplea una serie de locuciones adverbiales en el discurso, tras el desacuerdo o la falta de correspondencia que han mostrado los diccionarios. Esta utilidad de los corpus también la extendemos a la hora de examinar el funcionamiento del gran número de locuciones que no están contenidas en la microestructura de los diccionarios.

Y en tercer lugar, la revisión de los diferentes contextos que nos proporcionan los corpus para cada locución analizada ha evidenciado que, aunque hayamos podido relacionar el empleo de la locución con alguno de los contenidos ofrecidos por el *Diccionario*

fraseológico, ha sido bastante significativo el número de locuciones adverbiales que no se vinculan con ninguno de los contenidos que encierra la marca *literario*, puesto que son propias tanto de la lengua escrita como de la lengua oral, y se adscriben a un estilo neutro o a un estilo informal.

Bibliografía

CASARES, J., 1969. *Introducción a la lexicografía moderna*. Madrid: CSIC.

FAJARDO, A., 1996-1997. Las marcas lexicográficas: concepto y aplicación práctica en la Lexicografía española. *Revista de Lexicografía*, III, pp.31-57.

GARRIGA, C., 1997. Las marcas de uso en los diccionarios del español. *Revista de Investigación Lingüística*, 1, pp.75-110.

GUTIÉRREZ CUADRADO, J. dir., 1996. *Diccionario Salamanca de la lengua española*. Madrid: Santillana.

MOLINER, M., 1966 [1998]. *Diccionario de uso del español*. 2ª ed. Madrid: Gredos, 2 vols.

MORENO FERNÁNDEZ, F. dir., 1995. *Diccionario para la enseñanza de la lengua española*. Barcelona: Biblograf / Alcalá de Henares, Universidad de Alcalá de Henares.

PENADÉS MARTÍNEZ, I., 2005. *Diccionario de locuciones adverbiales para la enseñanza del español*. Madrid: Arco/Libros.

PORTO DAPENA, J. Á., 2002. *Manual de técnica lexicográfica*. Madrid: Arco/Libros.

REAL ACADEMIA ESPAÑOLA, 2014. *Diccionario de la lengua española*. 23ª ed. Madrid: Espasa-Calpe.

SÁNCHEZ CEREZO, S. dir., 1994. *Diccionario esencial Santillana de la lengua española*. Salamanca: Santillana.

SÁNCHEZ PÉREZ, A. dir., 2001. *Gran diccionario de uso del español actual*. Madrid: SGEL.

SECO, M., ANDRÉS, O. y RAMOS, G., 2004. *Diccionario fraseológico documentado del español actual*. Madrid: Aguilar.

VARELA, F. y KUBARTH, H., 1994. *Diccionario fraseológico del español moderno*. Madrid: Gredos.

UNE APPROCHE MIXTE DE L'ALIGNEMENT SOUS-PHRASTIQUE DE CORPUS PARALLÈLES ARABES-ANGLAIS

Abdelghani Yahiaoui
Université Lumière Lyon 2, France
abdelghani.y@hotmail.fr

Joseph Dichy
Laboratoire ICAR (CNRS/Lyon 2, ENS-Lyon), France
joseph.dichy@univ-lyon2.fr

Résumé

Nous présentons ici une approche mixte pour l'alignement automatique d'unités inférieures à la phrase dans des corpus parallèles arabes-anglais. L'utilisation de méthodes statistiques pour l'alignement des corpus parallèles arabes-anglais donne souvent des résultats peu satisfaisants en comparaison de la qualité d'alignement des paires de langues à écriture latine. Cela est dû principalement aux spécificités de langue arabe, notamment l'écriture non-voyellée et la structure agglutinante du mot. Nous proposons une approche mixte combinant (a) une méthode statistique basée sur les modèles IBM et le modèle de Markov caché et (b) une méthode linguistique basée sur une segmentation morphologique des mots arabes faisant appel à une base de données lexicale. Le cadre de recherche est, au sein du laboratoire ICAR, celui de l'équipe SILAT, de son cadre théorique et de ses réalisations (http://silat.univ-lyon2.fr). L'évaluation de notre approche est effectuée sur deux corpus parallèles qui sont différents en style de traduction et en longueur de phrases. Les résultats montrent une amélioration significative dans la qualité d'alignement par rapport à la méthode statistique standard.

1. INTRODUCTION

Cet article aborde le sujet de l'alignement automatique des corpus parallèles arabes–anglais au niveau du mot. L'objectif est d'améliorer la qualité d'alignement de ces corpus qui est souvent de moins bonne qualité que pour des corpus en écriture latine. Si l'alignement sous-phrastique des corpus de langues à écriture latine peut donner des résultats que certains estiment satisfaisants à partir des seules méthodes statistiques courantes, ce n'est généralement pas le cas pour des corpus mettant en parallèle l'arabe et des langues à écriture latines. C'est le cas des corpus arabes-anglais que nous traitons ici. C'est pourquoi il est préférable à notre sens de recourir à une approche mixte dans laquelle une méthode linguistique permettant notamment de surmonter les difficultés de l'arabe peut compléter une méthode statistique.

Nous dirons pour commencer un mot de l'alignement des corpus parallèles en général et des corpus arabes-anglais. Nous mettrons en lumière les spécificités de la langue arabe qui sont source de difficultés pour l'alignement sous-phrastique. Ensuite, Nous présenterons ensuite la méthode statistique choisie et la méthode linguistique développée dans notre équipe, ce qui nous amènera à montrer comment la fusion des résultats des deux méthodes a été effectuée. Enfin, nous procèderons à une évaluation de cette approche en l'appliquant sur deux corpus différents, tant par leur style de traduction que par la longueur des phrases.

2. L'ALIGNEMENT DES CORPUS PARALLÈLES

Un corpus parallèle est un ensemble de textes dans une langue source accompagnés de leur traduction dans une ou plusieurs langues cibles (OLOHAN 2004, BOWKER and PEARSON 2002, p. 24, 92). L'alignement est défini comme un objet pour indiquer les correspondances entre les mots dans un texte parallèle (BROWN et al. 1993). Celui-ci, en d'autres termes, consiste à mettre en correspondance chaque mot ou groupe de mot dans le texte source avec un homologue dans le texte cible. Cette définition correspond plus particulièrement à l'alignement sous-phrastique.

2.1. L'alignement sous-phrastique

L'alignement sous-phrastique automatique des corpus parallèles est une opération qui connait plusieurs applications dans le traitement automatique des langues naturelles (SMADJA, MCKEOWN et HATZIVASSILOGLOU 1996; MELAMED 2000 ; pour les liens avec la traduction automatique, v. OCH and NEY 2000 ; pour la recherche d'information multilingue, v. ROGATI et YANG 2004).

L'alignement sous-phrastique manuel des corpus parallèles, est couteux en temps et en ressources humaines (notamment dans les corpus volumineux), d'où l'utilité d'automatiser au moins partiellement cette opération. Pour développer des programmes d'alignement sous-phrastique, il existe plusieurs types de méthodes : les méthodes statistiques qui s'appuient sur des modèles probabilistes comme les modèles IBM (BROWN et al. 1993) et les chaînes de Markov cachées. Les méthodes statistiques sont souvent utilisées en raison de leurs résultats qui paraissent satisfaisants avec plusieurs paires de langues d'une part, et du fait qu'elles semblent ne pas dépendre d'une langue particulière d'autre part. Il existe aussi des méthodes heuristiques, qui utilisent des fonctions de similarité comme le coefficient de Dice (DICE 1945). Les méthodes heuristiques sont simples à implémenter et à comprendre, bien que le choix de la fonction de similarité semble arbitraire (OCH et NEY 2003). Le troisième type de méthodes porte sur les approches linguistiques qui s'appuient sur des analyses syntaxiques, morphologiques, sémantiques, lexicales, ainsi que sur l'utilisation de dictionnaires bilingues. Les méthodes linguistiques dépendent – bien évidemment – des langues des corpus à aligner.

2.2. Les difficultés de l'alignement arabe-anglais

L'alignement sous-phrastique des corpus parallèles arabes-anglais est confronté aux contraintes de la langue arabe, qui rendent la qualité de l'alignement en utilisant les méthodes standards moins satisfaisantes. Le cas de l'arabe, mais aussi d'autres langues sémitiques comme l'hébreu, conduit d'ailleurs à s'interroger sur les naïvetés que l'on peut lire çà et là sur les possibilités qu'offriraient des méthodes

exclusivement statistiques (POIBEAU, 2014 ; on trouvera un exemple de comparaison entre une approche purement statistique et une approche optimisant les résultats grâce à une analyse linguistique dans RAHEEL, 2010). Rappelons brièvement les caractéristiques de la langue arabe qui impactent le plus la qualité d'alignement sous-phrastique.

2.2.1. L'absence dans l'écriture arabe courante des signes diacritiques secondaires, dits de « voyellation »

L'absence de voyelles dans l'écriture courante pose d'énormes problèmes au traitement automatique de l'arabe. Elle entraine la non-notation des voyelles brèves, de la gémination des consonnes, des marques casuelles incluant une consone «ن» (*tanwîn*), etc., ce qui pose des problèmes d'ambiguïté graphique. Le tableau suivant présente le mot علم (sans voyelles), qui a au moins six valeurs différentes selon ses voyellations (sans préjuger des cas de polysémie, comme en 3 et 4 ci-dessous). Dans l'alignement sous-phrastique des corpus parallèles, l'absence des voyelles abaisse la qualité des résultats, notamment dans le cas où l'alignement prend en compte uniquement le résultat le plus probable.

	Mot voyellé	**Traduction**
1	عَلِمَ (*ʿalima*)	« il a su »
2	عُلِمَ (*ʿulima*)	« il a été su »
3	عَلَّمَ (*ʿallama*)	« il a enseigné » ou « il a marqué »
4	عُلِّمَ (*ʿullima*)	« il a été enseigné » ou « il a été marqué »
5	عِلْم (*ʿilm*)	science, connaissance
6	عَلَم (*ʿalam*)	drapeau

Tableau 1. Equivalents français du mot graphique علم (*ʿlm*) selon ses voyellations

2.2.2. Caractère agglutinant du mot et nécessité de spécificateurs morphosyntaxiques

Le mot graphique arabe comporte une structure d'objet complexe (DICHY et HASSOUN, 1989) : proclitiques, préfixes, suffixes et enclitiques sont attachés au radical (ou base) du mot et forment ensemble un *mot maximal* (COHEN 1970 ; DICHY 1997 ; pour une description du mot graphique et de ses structures, voir DICHY, 1990). Cette caractéristique nous met souvent dans des situations où pour un seul mot arabe, l'équivalent en anglais est plusieurs mots voire une petite phrase (DICHY et HASSOUN, eds., 1989). Cela complique le processus d'alignement sous-phrastique où l'unité principale d'alignement est le mot. Ci-dessous, des exemples de mot graphiques en arabe.

Mot graphique en arabe	Traduction en anglais
المعرفة (al-mʿrfṯ)	the knowledge
نبّهته (nbbhth)	I warned him
أستتذكّرونهم ؟ (asttḏkkrūnhm)	will you remember them?

Tableau 2. Exemples de mots illustrant le caractère « agglutinant » de la langue arabe.

Ces deux caractéristiques entraînent la nécessité, pour assurer l'analyse du mot, d'une base de données (DICHY et al., 2002 ; DICHY et HASSOUN, 2005) dans laquelle les entrées sont associées à des *spécificateurs morphosyntaxiques* « gérant » la relation entre le noyau lexical du mot et les morphèmes situés avant ou après lui dans le mot-forme (DICHY et HASSOUN, 1989 ; DICHY 1990 ; 1997 ; 2000. Pour les formes conjuguées, v. AMMAR et DICHY, 2008). Dans l'évaluation d'un analyseur automatique de l'arabe, l'absence d'une telle base de données est généralement considérée comme un point faible (DICHY et KANOUN, 2013).

3. UNE APPROCHE MIXTE

Présentons maintenant l'approche mixte conçue par nous pour améliorer l'alignement sous-phrastique des corpus parallèles arabes-anglais. Celle-ci s'appuie sur deux types de méthodes : une méthode statistique standard qui offre un premier alignement de corpus parallèles arabes-anglais, dont qualité est plus au moins satisfaisante. Puis, une méthode linguistique qui tient compte des contraintes linguistiques liées à la langue arabe et qui ne sont généralement pas prises en compte par les méthodes statistiques. Enfin, les résultats d'alignement des deux méthodes sont fusionnés selon un ensemble de critères qui seront présentés par la suite.

3.1. L'approche statistique

La méthode statistique utilisée est basée principalement sur les modèles statistiques IBM : 1, 2, 3, 4, 5 (BROWN et al. 1993) et aussi sur le modèle de Markov caché (MMC) (VOGEL 1996). Pour aligner les corpus parallèles, nous utilisons un outil qui s'appelle GIZA++ et qui implémente tous ces modèles. GIZA++ est une extension de programme GIZA qui fait partie de la boite à outil pour la traduction automatique statistique Egypt (AL-ONAIZAN et al). Les extensions de GIZA++ ont été conçues et implémentées par Franz Josef Och (OCH et NEY 2003). GIZA++ est l'un des outils les plus outils utilisés actuellement pour l'alignement sous-phrastique des corpus parallèles. GIZA++ peut donner des alignements de 1 à n ou de n à 1 entre le texte source et le texte cible. En l'appliquant dans les deux sens arabe-anglais puis anglais-arabe, nous pouvons obtenir un alignement de n-n, susceptible de grouper les unités polylexicales qui doivent rester solidaires lors du passage à la traduction, ce qui devrait permettre de diminuer le bruit concernant l'alignement (ABDULHAY 2012). Il existe plusieurs heuristiques pour fusionner les résultats d'alignement de GIZA++, nous avons choisi l'intersection où nous prenons en compte seulement les alignements qui sont présents dans les résultats d'alignements des deux sens. Ce choix est justifié par le fait que cette heuristique donne la meilleur précision ce qui est le critère le plus favorisé dans cette approche.

3.2. L'approche linguistique

Notre approche linguistique a pour objectif de renforcer et compléter la méthode statistique en alignant les mots d'un point de vue purement linguistique, en prenant en compte les spécificités de la langue arabe. L'alignement est effectué dans le sens arabe-anglais et peut donner des alignements de 1 à n ou de n à n. Pour aligner les mots arabes, il y a trois étapes principales:

A) Réalisation de toutes les segmentations morphologiques possibles du mot arabe en composants linguistiques basiques (proclitiques et préfixe, base ou noyau, suffixes et enclitiques) en prenant en compte les différentes voyellations possibles. Pour cela, un analyseur morphologique de l'arabe est nécessaire (voir, pour la structure de celui-ci, ZAAFRANI, 2002) (BUCKWALTER, T 2004).

B) Pour chaque segmentation morphologique – c'est-à-dire, pour chaque analyse du mot arabe (ANIZI et DICHY, 2009, 2011) –, nous essayons d'aligner le résultat avec son équivalent dans la phrase anglaise en se basant sur les traductions des composants linguistiques basiques de ce mot. Plus précisément, nous effectuons une recherche des équivalents des composants linguistiques basiques dans un intervalle de trois mots avant et cinq mots après la position qui correspond au mot arabe dans la phrase anglaise. Pour l'implémentation de cette étape, il est nécessaire de se référer à un dictionnaire arabe-anglais qui comporte aussi les traductions des préfixes et des suffixes arabes.

C) Choix du meilleur alignement pour un mot arabe donné parmi tous les alignements effectués dans l'étape précédente. Le meilleur alignement est celui où nous pouvons aligner le maximum des composants linguistiques basiques du mot arabe.

3.3. « Fusion » des résultats

Après avoir explicité l'alignement sous-phrastique statistique et linguistique, détaillons la fusion des résultats des deux méthodes. Cette fusion est effectuée selon trois critères :

- L'intersection : tous les alignements présents dans les résultats des deux méthodes sont considérés ;
- Tous les alignements de la méthode statistique sont pris en compte ;
- Les alignements de la méthode linguistique sont pris en compte s'ils sont compatibles avec leurs voisins dans la méthode statistique. C'est-à-dire, si par exemple un mot arabe X est aligné avec un mot anglais de positon P dans la phrase anglaise, les voisins du mot X devraient être alignés –si c'est le cas- avec des mots anglais de positions proches de P.

4. EVALUATION

4.1. Le recours, pour l'évaluation

Pour l'évaluation de notre approche, nous avons utilisé deux corpus parallèles arabes-anglais. Ces corpus sont extraits à partir de la plateforme des corpus parallèles open source OPUS (TIEDEMANN 2012) et ils sont déjà alignés au niveau de la phrase. La description de ces deux corpus est détaillée dans le tableau suivant.

Nom de corpus	**OpenSubtitles2012**	**UN**
Description	Une collection de documents de sous-titres de films, séries ... etc. de http://www.opensubtitles.org/	Une collection de documents traduits de l'organisation des Nations Unies à l'origine compilé dans une mémoire de traduction.
Nombre de phrases	7 600 000	74 100
Nombre de mots arabes	46 300 000	2 700 000
Nombre de mots anglais	57 000 000	3 000 000
Nombre de phrases (filtrées et tokenisées)	7 593 701	67 491
Moyenne de la longueur des phrases (mots/phrase)	6	40
Style général de traduction	Traduction direct.	Le sens est bien soigné mais les mots utilisés sont parfois différents et l'ordre des expressions dans certaines longues phrases n'est pas toujours le même.

Tableau 3. Caractéristiques des corpus parallèles utilisés pour l'évaluation.

Pour calculer les critères d'évaluation, nous avons choisi un échantillon de 100 paires de phrases par corpus. Ce choix est effectué arbitrairement parmi les paires de phrases dont la partie arabe ne contient que des mots écrits en lettres arabes. Cela est justifié par le fait de vouloir montrer l'impact de la méthode linguistique sur l'amélioration de l'alignement statistique, et aussi du fait que le mélange des caractères latins et arabes pose certains problèmes d'affichage et de traitement.

Nous avons ensuite procédé à un alignement manuel de ces deux échantillons, afin de constituer un échantillon aligné de référence pour évaluer l'alignement automatique effectué avec notre approche.

4.2. Critères d'évaluation

Les critères d'évaluation utilisés sont les mêmes critères définis par Franz Josef Och (OCH and NEY 2003) : le rappel, la précision et le taux d'erreur AER (Alignemnt Error Rate). Ce dernier combine les deux autres critères et reflète le plus la qualité d'alignement si elle est bonne ou non. Plus la valeur de l'AER est petite, plus la qualité d'alitement est bonne. Les trois critères d'évaluation sont calculés de la façon suivante :

A : l'ensemble de tous les liens entre les mots dans l'alignement à évaluer.

S , P : ensembles des liens sûrs et possibles entre les mots dans l'alignement de référence.

| | : désigne la cardinalité d'un ensemble.

$$\textbf{Rappel} = \frac{|A \cap S|}{|S|} \qquad \textbf{Précision} = \frac{|A \cap P|}{|A|}$$

$$\textbf{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

4.3. Résulats d'évaluation

Les résultats d'évaluation des deux corpus sont exprimés en pourcentage dans les deux tableaux 4 et 5. Nous avons d'abord effectué l'évaluation de la méthode statistique standard, puis l'évaluation de la méthode linguistique qu'on a conçue, et enfin, l'évaluation de la méthode mixte que nous proposons pour améliorer l'alignement.

	Rappel	Précision	AER
Méthode statistique	58.79	**90.56**	28.39
Méthode linguistique	36.90	76.76	50
La méthode mixte	**71.38**	86.69	**21.56**

Tableau 4. Résultats d'évaluation pour le corpus OpenSubtitles2012 (valeurs exprimées en pourcentage).

Les résultats d'évaluation concernant le corpus parallèle OpenSubtitles2012 montre une amélioration significative dans la qualité d'alignement avec la méthode mixte par apport à la méthode statistique utilisée généralement pour l'alignement sous-phrastique. Même si les résultats de la méthode linguistique sont largement insuffisants mais l'objectif n'était pas d'aligner les corpus uniquement avec cette méthode mais plutôt la conception d'une méthode linguistique qui permet en combinaison avec la méthode statistique d'améliorer la qualité d'alignement sous-phrastique. La méthode mixte a largement amélioré le rappel et la précision est bonne ce qui donne un AER de 21.56 % soit une amélioration d'environ 7 % par apport à la méthode statistique.

	Rappel	Précision	AER
Méthode statistique	49.19	**90.11**	36.25
Méthode linguistique	33.37	57.95	57.15
La méthode mixte	**58.87**	74.34	**34.09**

Tableau 5. Résultats d'évaluation pour le corpus UN
(valeurs exprimées en pourcentage).

Les résultats d'évaluation pour le corpus parallèle UN montre aussi une amélioration dans la qualité d'alignement avec la méthode mixte par apport à la méthode statistique. Comme nous pouvons le constater, la méthode mixte a significativement amélioré le rappel, la précision reste relativement bonne ce qui donne une amélioration dans l'AER d'environ 2 % par apport à la méthode statistique.

La baisse dans le taux d'amélioration de l'alignement entre le corpus OpenSubtitles2012 où il était environ 7 % et le corpus parallèle UN où il était environ 2 % est principalement due au style de traduction différent. A la différence du corpus OpenSubtitles2012 où les traductions sont plutôt directes et les phrases sont courtes, les mots dans le corpus UN sont parfois différents et l'ordre des expressions dans certaines longues phrases n'est pas toujours le même entre le texte arabe et le texte anglais ce qui impacte négativement les résultats de la méthode linguistique qui est basée sur une segmentation morphologique et un dictionnaire bilingue. Par conséquent, cela baisse le taux d'amélioration dans la méthode mixte qui est la fusion des deux méthodes.

5. CONCLUSION

Cet article, présente une approche mixte pour l'amélioration de l'alignement sous-phrastique automatique des corpus parallèles arabes-anglais. Avec les seules méthodes statistiques standards, les résultats d'alignement des corpus parallèles arabes-anglais ne sont pas assez satisfaisants à cause de spécificités de la langue arabe comme la structure complexe du mot-forme, qu'il est malaisé de prendre en compte par ces méthodes statistiques. Nous avons donc proposé une méthode mixte dans laquelle nous avons effectué d'abord un alignement avec une méthode statistique basée sur les modèles IBM et le modèle de Markov caché, puis nous avons effectué un alignement avec une méthode linguistique basée sur une segmentation morphologique des mots arabes et l'utilisation d'un dictionnaire arabe-anglais. Ensuite, nous avons fusionné les résultats des deux méthodes pour arriver à l'alignement final.

Pour évaluer notre approche, nous avons utilisé deux corpus parallèles : le corpus OpenSubtitles2012 qui se caractérise par un style de traduction direct et des phrases d'une longueur moyenne et courte, et le corpus UN qui se caractérise par un style de traduction où le sens est bien soigné mais pas les mots, et l'ordre des expressions peut parfois être différent. Les résultats de cette évaluation et de ce travail sont résumés dans les deux points suivants :

- L'approche mixte proposée dans cet article améliore significativement la qualité d'alignement sous-phrastique des corpus parallèles arabes-anglais et en particulier les corpus ayant un style de traduction direct.
- La segmentation morphologique des mots arabes joue un rôle important dans l'amélioration de la qualité de l'alignement sous-phrastique dans cette approche.

Une suite envisageable de ce travail est l'optimisation de la méthode linguistique expliquée dans cette approche afin d'améliorer l'alignement encore plus, ainsi que la vérification de cette approche avec d'autres paires de langues notamment avec les corpus arabes-français.

Annexe sur la translitération des caractères arabes

La translitération des mots arabes a été effectuée selon les traditions arabisantes françaises et internationales (revue *Arabica*, éditeur Brill, Leiden).

Remerciements

Les auteurs remercient le laboratoire ICAR (Interactions, Corpus, Apprentissages, Représentations, UMR 5191, CNRS / Université Lumière-Lyon 2 et ENS-Lyon) au sein lequel ce travail a été effectué, et le Labex ASLAN (Lyon) pour son soutien. Merci également à Mohammed Amin Bouchoukh.

Références

ABDULHAY, A., 2012. *Constitution d'une ressource sémantique arabe à partir d'un corpus multilingue aligné*. Thèse, Université de Grenoble. https://tel.archives-ouvertes.fr

ALEXANDRE, A. and GUILLAUME, W., 2009. Modèles discriminants pour l'alignement mot à mot in *TAL*-2009-3-06-Allauzen. http://www.atala.org/IMG/pdf

AL-ONAIZAN, Y., CURIN, J. ABDULHAY, JAHR, M., KNIGHT, K., LAFFERTY, J., MELAMED, D., OCH, F.-J., PURDY, D., SMITH, N.A., and YAROWSKY, D., 1999. Statistical machine translation. In: *Final Report, JHU Summer Workshop* [online]. Available from: http://mt-archive.info/JHU-1999-AlOnaizan.pdf [Accessed 22 Jul 2015].

AMMAR, S. et DICHY, J., 2008, *Bescherelle des verbes arabes* (2e édit., revue et augmentée de *Les verbes arabes,* 1999)*, Paris, Hatier (coll. Bescherelle).*

ANIZI M. and DICHY, J., 2009, Assessing Word-form based Search for Information in Arabic: Towards a New Type of Lexical Resource, in: Khalid Choukri and Bente Maegaard, *Proceedings of the Second International Conference on Arabic Language Resources and*

Tools, 22-23 April 2009, Cairo, Egypt, The MEDAR Consortium. http://silat.univ-lyon2.fr ou: http://www.elda.org/medar-conference/pdf/75.pdf

ANIZI M. and DICHY, J., 2011, "Improving information retrieval in Arabic through a multi agent approach and a rich lexical resource", *Knowledge Organisation journal*, v. 38 (2011) No. 5. Würzburg: Ergon. http://www.isko.org/ko.html

BOWKER, L. and PEARSON, J., 2002. *Working with specialized language: a practical guide to using corpora*. Routledge.

BROWN, P.F., PIETRA, V.J.D., PIETRA, S.A.D., and MERCER, R.L., 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19 (2), pp. 263–311.

BUCKWALTER, T., 2004. Issues in Arabic Orthography and Morphology Analysis. In: *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 31–34. Available from: http://dl.acm.org/citation.cfm?id=1621804.1621813 [Accessed 28 Jul 2015].

COHEN, D., 1961/70. Essai d'une analyse automatique de l'arabe. 1961, *TA information, in* D. Cohen, *Etude de linguistique sémitique et arabe*, Paris, Mouton, 1970.

DICE, L.R., 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 26 (3), pp. 297–302.

DICHY, J., 1990, *L'écriture dans la représentation de la langue: la lettre et le mot en arabe,* Thèse pour le doctorat d'État, Univ. Lyon 2, 2 vol., 660 p.

DICHY, J., 1997. Pour une lexicomatique de l'arabe: l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. *Meta: Journal des traducteurs/ Meta:Translators' Journal.* 42 (2), pp. 291–306. www.erudit.org/revue/meta/1997/v42/n2/002564ar.pdf

DICHY, J., 2000. Morphosyntactic Specifiers to be associated to Arabic Lexical Entries - Methodological and Theoretical Aspects. Proceedings of ACIDA' 2000 (Monastir, Tunisia, 22-24.03.00), Corpora and Natural Language Processing volume: 55-60. http://silat.univ-lyon2.fr

DICHY, J., BRAHAM, A., GHAZALI, S., and HASSOUN, M., 2002. La base de connaissances linguistiques DIINAR. 1 (DIctionnaire INformatisé de l'Arabe, version 1). In: *Proceedings of the International Symposium on The Processing of Arabic, Tunis (La Manouba University)* [online]. pp. 18–20. Available from: http://silat.univ-lyon2.fr or http://ghazali.freehomepage.com/publications/Joint%20paper%20on%20Diinar.pdf

DICHY, J. et HASSOUN, M., éds., 1989, *Simulation de modèles linguistiques et Enseignement assisté par ordinateur de l'arabe - Travaux SAMIA I,* Paris, Conseil international à la langue française (CILF), 256 p.

DICHY, J. et HASSOUN, M., 2005, "The DIINAR.1-« معالي » Arabic Lexical Resource, an outline of contents and methodology", *The ELRA Newsletter*, Vol. 10, n°2, April-June 2005, p. 5-10. http://silat.univ-lyon2.fr

DICHY, J. et KANOUN, S., éds., 2013, *Linguistic Information Integration in Arabic Text and Character Recognition*, special issue of *Linguistica Communicatio, Al-TawâsSul al-lisânî,* vol. 15, n°1-2 , Fez (Maroc).

MELAMED, I.D., 2000. Models of Translational Equivalence among Words. *Computational Linguistics*. 26 (2), pp. 221–249.

OCH, F.J. and NEY, H., 2000. A Comparison of Alignment Models for Statistical Machine Translation. In: *Proceedings of the 18th Conference on Computational Linguistics - Volume 2* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 1086–1090. Available from: http://dx.doi.org/ 10.3115/992730.992810 [Accessed 24 Jul 2015].

OCH, F.J. and NEY, H., 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*. 29 (1), pp. 19–51.

OLOHAN, M., 2004. *Introducing corpora in translation studies*. Routledge.

POIBEAU, T., 2014, « La linguistique est-elle soluble dans la statistique ? », *Revue Sciences/Lettres* [En ligne], 2 | 2014, mis en ligne le 07 octobre 2013, consulté le 21 juillet 2015. URL : http://rsl.revues.org/ 402 ; DOI : 10.4000/rsl.402

RAHEEL, S., 2010, *Apprentissage artificiel et fouille de données multilingues : le cas des documents arabes*, doctorat en Sciences de l'information et de la communication, Université Lumière-Lyon 2.

RAHEEL, S. et DICHY, J., 2010, "An Empirical Study on the Feature's Type Effect on the Automatic Classification of Arabic Documents", in Gelbukh, Alexander (ed.), in *Computational linguistics and intelligent text processing,* proceedings of the 11th international conference, *CICling 2010* (Iaşi, Romania, March 2010) Berlin, Heidelberg, New-York: Springer Verlag, p. 673-686.

ROGATI, M. and YANG, Y., 2004. Multilingual Information Retrieval Using Open, Transparent Resources in CLEF 2003. In: C. PETERS, J. GONZALO, M. BRASCHLER and M. KLUCK, eds. *Comparative Evaluation of Multilingual Information Access Systems* [online]. Springer Berlin Heidelberg. pp. 133–139. Available from: http://link.springer.com/chapter/10.1007/978-3-540-30222-3_12 [Accessed 20 Jul 2015].

SMADJA, F., MCKEOWN, K.R., and HATZIVASSILOGLOU, V., 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Comput. Linguist.* 22 (1), pp. 1–38.

TIEDEMANN, J., 2012. Parallel Data, Tools and Interfaces in OPUS. In: *LREC* [online]. pp. 2214–2218. Available from: http://lrec.elra.info/proceedings/lrec2012/pdf/463_Paper.pdf [Accessed 31 Mar 2015].

VOGEL, S., NEY, H., and TILLMANN, C., 1996. HMM-based Word Alignment in Statistical Translation. In: *Proceedings of the 16th Conference on Computational Linguistics - Volume 2* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 836–841. Available from: http://dx.doi.org/10.3115/993268.993313 [Accessed 22 Jul 2015].

ZAAFRANI, R., 2002, *Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère*, doctorat en Sciences de l'information et de la communication, Université Lumière-Lyon 2.