Introducing *ProTermino*: A New Tool Aimed at Translators and Terminologists

[ProTermino: *una nueva herramienta dirigida a traductores y terminólogos]*

Isabel Durán Muñoz, Gloria Corpas Pastor (1)
Le An Ha y Ruslan Mitkov (2)
iduran@uma.es
(1) Universidad de Málaga
(2) University of Wolverhampton

Abstract

The aim of this paper is to introduce *ProTermino*, a comprehensive terminological management system that has been recently developed in the framework of a Spanish R&D project.[1] As a terminological management tool its target users are terminographers or translators working in the terminology domain. This system supports English, German, Spanish, Italian and French and presents a very user-friendly interface. In this paper, we present the main functionalities and specifications of *ProTermino* and the reasons that launch us to work on such a tool. Subsequently, we examine similar systems and compare them with the advantages that *ProTermino* brings to the terminology domain. And finally we depict some research results achieved with this tool.

Keywords

terminological management tool; terminographers; translators; *ProTermino*

Resumen

El objetivo de este trabajo es presentar *ProTermino*, un sistema de gestión terminológica integral que ha sido desarrollado recientemente en el seno de un proyecto de I+D+i de ámbito nacional.[1] Como herramienta de gestión terminológica está dirigido principalmente a terminógrafos o traductores que trabajan en el campo de la terminografía. Esta aplicación cuenta con las siguientes lenguas de trabajo: inglés, alemán, español, italiano y francés y presenta una interfaz amigable e intuitiva. En este trabajo, exponemos las principales funcionalidades que presenta *ProTermino* y las razones por las que emprendimos su diseño e implementación. A continuación, examinamos sistemas similares y comparamos las ventajas que ofrece nuestra herramienta y, finalmente, describimos algunos resultados obtenidos con *ProTermino* hasta el momento.

Palabras clave

gestor de terminología; terminógrafos; traductores; *ProTermino*

1. Introduction

AT PRESENT, the influence of computational and corpus linguistics is observed in almost every terminological project, and terminographers' tasks have been facilitated thanks to the introduction of computational linguistic technologies. However, they do not cover all terminographers' needs, and thus they are obliged to combine several tools and technologies to carry out a terminological project. For example, they need a terminological management tool to create a terminology database, along with term extractors, concordancers and concept map editors, among other tools, so as to accomplish the different phases and steps of any terminological project. This situation provokes a mushrooming of these tools, which hampers terminographers' tasks. As a consequence, technology, on the one hand, facilitates these tasks by avoiding time-consuming manual processing but, on the other, the great number of needed applications hinders their tasks by requiring the combination of different tools to carry out different purposes.

In this light, we have designed and implemented *ProTermino*, a comprehensive and flexible tool that provides corpus, terminological and ontological modules so as to carry out a multilingual terminological project based on knowledge representation (Durán-Muñoz 2012). As a terminological management tool its target users are terminographers or translators working in the terminology domain. However, a key point of this tool is that the target group of the terminological products are mainly translators and terminologists, since the structure, the terminological fields available, the ontological information contained as well as the different exportation options are based on their needs and expectations. This system supports English, German, Spanish, Italian and French and presents a very user-friendly interface. It also provides different exportation options and a selection of fields to the term entry along with other possibilities, such as the multi-user access to work simultaneously, remote access, the creation of different projects, which foster a collaborative working environment and facilitate terminographers' tasks as much as possible.

This paper is structured as follows: first, we analyse the requirements of a typical terminological project; then, we examine previous work and applications close to *ProTermino*; third, we depict the main functionalities and specifications of our tool and the reasons that launch us to work in it. And finally we present some research results achieved with this tool.

2. Meeting the needs: Selection of Working Tools

Before a project commences, the terminological tools to be employed are determined according to the available budget and according to the requirements arising from the project objectives. In some cases, terminographers would have to employ a set of tools that cover different steps in a terminological projects, but in other cases, comprehensive tools would be available and ready to use, i.e., terminographers would be able to employ an application that provides several tools integrated on the same platform without the need of using a range of different applications or even of some manual work. It is clear that the second situation is always preferable but, as said, it is not always possible.

Owing to the possible economic constraints that might encounter, a terminological project it is recommended to choose a tool that meets at least 80 % of the initial requirements, as indicated by Pavel and Nolet (2001:xx). Following these authors, we need to answer a number of questions when selecting the tool with which we will be working to select the one that best fits our needs. Basically, these issues relate to the following points:

- ***Capacity***, which refers to the volume of data that is expected to manage the tool, i.e., we need to select a tool that will be able to manage the volume of information in accordance with our project.
- ***Access***, which makes reference to the type of access you require when working with the tool and the number of people working in it. For example, we should have to take into account whether the access is simultaneous by different researchers, remote, etc.
- ***Flexibility***. The tool should provide the flexibility to meet the needs of the project with respect to the number of terminological fields you want to include in the database, the selection of these fields, the working languages, etc. That is, the selected tool should comply with the pragmatic-linguistic variables that are set at the beginning of any terminological project and either it directly provides the fields, languages, etc. that have been selected for the project, or it allows the creation or modification of such information in a flexible and easy way.
- ***Specifications***. At this point you need to determine whether specific functionalities to the terminological project are required, such as the need to create ontologies, managing textual corpus, to include audio material, etc. In any case, the tool selected should take into account the exact specifications of each project, or at least allow customisation and adaption of the tool to meet these requirements.
- ***Exchange***. This point might be relevant depending on the goals set for the terminological project. In case terminographers' preferences lead to the reusability and exchange of the information gathered in the database, the application selected should provide a format suitable for such exchange or reuse.
- ***Export***. Related to the previous point, the tool selection is determined by the exportation format that provides the application. Thus, depending on the initial goals terminographers should select a tool that allows them to export in .rtf, .pdf, .html, .xml, or other.

These items listed above are not intended to be exhaustive, but they can be considered as basic points, which should underpin decisions when selecting one or other terminology management tool for the project to be performed.

3. Meeting our needs

As part of our research, we established our own needs according to the above issues to select the suitable tool that meets our project requirements, or at least part of them.

- ***Capacity.*** The intention of our research was to carry out a terminological tool for the tourism domain, including different tourist segments, and also to employ it in forthcoming projects. In this sense, the chosen application would have to manage and store a large amount of terminological information.
- ***Access***. The access should be remote (via the Internet) and allow simultaneous researchers working on the same or different projects. Therefore, the tool needs to provide remote and multiple-user access.
- ***Flexibility.*** The tool should provide flexibility in selecting the terminology fields in which the project is involved and allow the selection of different working languages. At present, the working languages are Spanish, English, German, Italian and French, but other languages could be added and also the terminographers are not obliged to work with all the languages in different projects.
- ***Specifications.*** Specific functionalities for our project are mainly based on the possibility of developing and managing ontologies as well as large corpora. Also, we considered having a tool that allows the inclusion of graphics such as images as supporting material.
- ***Exchange.*** The selected tool should allow the exchange of information with other applications and, thus, the exchange format should meet the current standard requirements. In this sense, the TBX format would be preferable for being the exchange format standard recommended by international organisations like ISO.

- ***Export.*** As mentioned above, the selected tool should allow the exportation in TBX format. Also, the possibility of exporting to other formats for paper edition, such as .rtf or .pdf would be positively considered.

Once we have established our basic needs regarding the selected tool, we conducted a study of the different terminology management systems that would help us achieve the goals of our project. In the next section, we present these tools and our remarks.

3.1. Possible terminological management systems

In this section we provide a review of different terminological management systems that are available at the moment, either under commercial license or are free-access versions. At this point we need to make a distinction between terminology management systems (TMS) and ontoterminological management systems (OTMS), the latter being systems that allow ontology management.

On the one hand, TMS are widely used for translators and terminological projects of different natures, since they present many interesting features and advantages. These applications enable the management of data, i.e., allow the creation, management and administration of terminological databases, as well as the inclusion of definitions and other linguistic information. Some of the most extended TMS nowadays are *SDL Multiterm, TermStar Star, MemoQ*, which are linked to translation memories (although some may also be used independently), and *TshwaneLex Suite*, among others. None of these programs are freely distributed, but all offer a trial version to check their features. Furthermore, these programs offer great manageability and data storage, different types of access (remote / local, single or multiple users), great flexibility in selecting the fields of terminological entries and working languages, as well as different export formats and interchangeability. However, despite all these advantages, they do not meet the initial requirements raised in our terminological project: they do not allow the development of ontologies nor corpus management; neither do they provide concordancers or term extractors in their systems. Furthermore, the data exchange provided is limited to users with the same systems, i.e., users can exchange data as long as this data is uploaded to the same system (for example, the database created by a user *SDL Multiterm* in the system of another user B).

For this reason, it is recommended to use other management systems that function as comprehensive workstations when performing any systematical terminological work. These workstations are comprehensive terminology management systems that provide all the necessary modules to perform a complete terminological work, from the compilation of the working corpus to the export step for further editing. Here we can also distinguish two subgroups: those that do not include ontology creation module, such as *TERMINUS*, developed by the IULA group at the University Pompeu Fabra, and *System Quirk*, by the University of Surrey; and those which do include ontologies, like *Ontoterm*, *Termontography Tools* or *Corpógrafo*.[2] These systems are usually developed in the framework of research projects and, as such, most of them are not possible to employ out of these projects, but it is worth knowing them and taking them into account.

The authors themselves of *TERMINUS* (Cabré et al. 2012) define it as a workstation for terminology, since it integrates in a single working environment all the phases for a complete terminology project. In other words, this tool allows us to carry out from the compilation of a corpus to the final edition of the resource, passing through corpus management and selecting the terminological fields that make up the terminology entries. It also allows us to create and manage different user profiles, create and manage more than one terminology project and work simultaneously from different locations.

This workstation features a modular structure, which consists of different sections that allow us to perform the terminology work, namely:

Projects	to create one or several terminology projects according to user needs.
Sources	to manage the fonts used in a terminology project.
Structuring concept	to create a conceptual tree to structure the terms of a domain.
Documents	to add text files in different formats to compile the corpus.
Corpus	to compile a corpus directly through searches in the Internet.
Analysis	to analyse corpus frequencies, concordances, n-grams and calculation of association.
Glossaries	to create glossaries.
Terms	to introduce data in the terminology entry, to consult and to modify (if necessary).
Export	to export in the following formats: .html, .pdf, .txt and .xml.

In general, it appears that this workstation is a useful resource for conducting terminology projects, since it assists the terminographers in completing all the stages of a terminology project (cf. Cabré Castellví, 1993) and allows customisation to different user profiles.

On the other hand, *System Quirk* is also a flexible and inclusive package of tools for creating and managing terminological databases, but it includes some more features that *TERMINUS*. Like the previous one, this system assists the terminologist during the phases of the terminology work, from the compilation of the corpus to the edition of dictionaries, and it is flexible regarding the number of users (single or multiple users), the number of projects, the terminological entry fields, etc. However, it offers some differences that distinguish it from earlier and makes it a bit more complex. The main difference noticed is the possibility of automatic term extraction from the working textual corpus uploaded into the application with minimal interaction from the user.

The architecture of this system is also modular, to easily organise work and allow greater flexibility. Thus, we find different applications integrated within the system that allow us to carry out the tasks of any terminology work:

Virtual Corpus	to create and management corpora.
Kontext	to analyse texts by generating word lists, concordance search, collocations, etc.
Ferret	to carry out statistical analysis to the corpus to locate terms, especially compound terms.
Browser / Refiner	to create and modify terminology databases.

It also offers other additional modules, such as automatic summarisation or word aligner, with the aim of completing the needs of terminologists. Finally, it seems relevant to point out that, despite being a very complete, flexible and easy to handle, does not allow the export of the terminology database, i.e., the system is designed to develop and query the data within the same program. Consequently, it does not currently allow export to any format.

Considering the initial requirements of our terminology project, these tools would be useful to some extent, since they would need the combination of independent ontology editors, such as *Protégé* or similar. Therefore, we are forced to continue analysing other systems that include ontology creation and management in their workstation. As stated before, examples of this kind of tools are *Ontoterm*, *Corpógrafo* and *Termontography Tools*.

First, *Ontoterm* (Moreno Ortiz, 2000, 2004) is a terminology management system based on open access use and addressed to research and academic users. It allows the development of domain ontologies and the creation of terminological databases based on those ontologies, as well as it provides remote and simultaneous access. This tool has been used in several terminology projects in Spain, as in the genome-KB project[3] and the OncoTerm project.[4]

This application is divided into two main modules: first, an ontology editor, which allows the creation of ontologies from scratch or based on previous ontologies and thus, the knowledge representation of a domain of expertise in question; and, secondly, a terminology database manager, which enables the creation and editing of information terminology of the concepts previously introduced in the ontology. Besides these two main modules, *Ontoterm* also consists of an ontology browser, which allows the query and display of the ontology, as well as a generator of reports in HTML, which allows users to export term entries in this format. Despite the advantages observed in this tool, it also has some disadvantages that prevent us from employing it in our project: the most important one is the lack of technical support and maintenance of the tool, but also the inability to export the database created with the tool to formats different from .html, like .rtf or .xml, and the complexity of the information shown once the entries are published in .html (cf. Durán Muñoz 2010:7–8).

The second tool discussed under this section is *Termontography Tools*,[5] which is a set of three interrelated applications that are framed in the Socio-cognitive theory of Terminology (Temmerman 2000). These tools are developed in JAVA and were created within the European project FFPOIROT (de Baer et al. 2006), at the Centrum voor Communicatie Vaktaal in (Erasmushogeschool, Brussels). These tools aim at creating ontoterminographical resources in three phases, one for each tool: first, the domain conceptualisation is performed at a conceptual level (i.e. language independent); secondly, the terminological database is filled with information taken from the corpus and previously developed conceptual representation at a terminological level (i.e. language dependent) and, finally, the last application allow users to consult the terms included in the database. These tools are easy to use and user-friendly as they have been designed by and for terminologists. The main problems encountered in this tool are related to the by default semantic relations, which all are part-whole relations. Consequently, terminologists are obliged to change all relations that are not part-whole and create the required relations, which involve several unnecessary steps. On the other hand, it establishes a division of meta-categories and categories that are not easy to deal with at the conceptual level and provoke misinterpretations. Finally, the tool that allows users to check the information included in the database is quite limited, since it does not include all the information that has been added in the previous tools neither does it give the possibility to choose the information to be displayed, but just several fixed fields: definition, related terms and language.

In general, we would conclude that the two programs previously discussed, both freely available for academic use, have great strengths in their design and use, however, their weaknesses make their selection limited, especially *Termontography Workbench*, as they would need deep revision of some of their features to provide a more useful tool in the field of ontoterminography.

Another application that can be discussed under this section is *Corpógrafo*,[6] a freeware application developed at the University of Oporto with the aim of supporting the work of linguistic researchers, especially terminologists. Unlike the previous two applications, which had a conceptual-oriented approach, this tool is term-oriented and, thus, it does not require a domain conceptualisation as starting point. It is also a set of several tools integrated on a common platform addressed at the needs of a terminological project and, thus, it provides some similarity with aforementioned *TERMINUS* program, or the *Spaterm* prototype, which will be discussed below. However, we must emphasise that it is a system that enables semantic work and, hence, it differs from *TERMINUS*.

Corpógrafo includes a variety of tools ranging from format converters to concordance searchers and semiautomatic semantic relationships extractors, through tools that allow users to create their own databases. The structure is divided into four main areas of work, namely:

1. The manager *(Manager)*, which includes editing tools and pre-processing applications, as well as comparisons between corpora;
2. The analyser *(Research)*, which automatically extracts term candidates using n-grams, and search for collocations and concordances;
3. The knowledge centre *(Centro de Conhecimento)*, where the tools to generate and organise knowledge are found;
4. The communication centre *(Centro de Comunicação)*, where all the documentation for the application is listed, together with the received and sent messages.

Along with these options, the program also offers other features, such as text aligner, document uploading, etc. Finally, it also allows exporting the results to various formats and applications, following the standard terminology database (XML) and translation memory.

Despite the advantages presented in this tool, its use is entirely limited to terminology work in Portuguese. In other words, the system has been developed by and to Portuguese researchers and, thus, all the applications are implied for the Portuguese language. This fact allows us to use the tool and check how it works and is structured, but it is unfortunately unsuitable for our project.

4. *ProTermino* as a need

After the thorough review carried out on several possible management systems, we concluded that none of them were suitable to accurately meet our needs. Therefore, inspired by them and led by our needs, we designed and developed our own tool, *ProTermino*, a comprehensive tool to carry out ontology-based terminological projects.

4.1. Functionalities of *ProTermino*

ProTermino is a modular tool which consists of three main modules: corpus management, terminological database and knowledge patterns, allowing terminographers to carry out the different phases of a multilingual terminological project based on knowledge representation (Durán-Muñoz 2012), namely, corpus management, concordance search, term and equivalent extraction, ontology creation, creation of term entry and export to edition.

These three modules are also divided into several sub-modules, which facilitate terminographers' tasks: the first one, corpus management includes *a)* corpus uploading and *b)* term extraction; the second module integrates *a)* terminological entry templates according to the ISO standard, *b)* cognate identification to extract equivalent candidates, *c)* a concordancer, *d)* ontology building, and *d)* terminology export in several formats (.PDF, .RTF, .HTML and .TBX); and finally the third module includes the option to upload semantic patterns and provide candidate semantic relationships and instances taken from the corpus. In Table 1 the different functionalities are classified in the corresponding modules and sub-modules:

MODULE	SUB-MODULES
Corpus module	Upload a corpus (compressed, text by text, copy-paste)
	Consult the corpus (by terms or by documents)
	Term extractor
Terminological database module	Seek terms
	Validate term candidates
	Eliminate term candidates
	Consult validated, eliminated and extracted candidate terms
	Fill in term entries
	Manually introduce terms
	Export (.rtf, .pdf, .html, .xml)
Knowledge patterns module	Add knowledge patterns and relations
	Search semantic relations in the corpus
	Represent detected semantic relations
	Edit detected semantic relations

Table 1. Classification of modules and sub-modules

The *ProTermino* tool is also multilingual (currently working with English, German, Spanish, Italian and French), and it displays a very user-friendly interface. Besides, it is a very flexible tool as it also provides different exportation options and selection of fields to the term entry along with other possibilities, such as the multi-user access to work simultaneously, remote access, the creation of different projects, so as to foster a collaborative working environment and facilitate terminographers' tasks as much as possible.

4.2. *ProTermino* Workflow

ProTermino is organised in projects, which are created by users according to their own needs. This step is necessary since this tool provides the possibility to create different working projects simultaneously, and allows remote access to multiple users. Thus, by allowing simultaneous jobs and users, the tool requires the identification of on-going projects so as to avoid confusion and provide support in subsequent phases. In this step, users must provide basic information about the project to be created, namely: the title and a brief description of the project, the working languages, etc. in order to identify it. Once the project is been created, it will be added to the list of projects that have been created so far.

Besides creating projects, the user has also the ability to delete previous projects by selecting them and pressing the delete button. Before doing it, the user must be aware that, once removed the selected project, all the information (documents, extracted terms, full term entries, etc.) in those projects will be also deleted.

Once we have accessed the tool and created the corresponding project, a corpus must be uploaded. To do so, the user is provided with different options:

1. a compressed corpus in .zip or .7zip format containing documents in plain text (.txt) can be uploaded and automatically decompressed by *ProTermino*,
2. users can upload document by document in .txt format, and
3. users can copy-paste the text to be considered for the tool.

Accordingly, users are able to select the most suitable option for their needs. Along with the corpus upload, the corpus language must be specified so as to be correctly incorporated into the application. Hence, in case of working with several languages we need to upload a corpus for each language, specifying the language of each corpus for a correct processing. Once the corpora (in case of working with more than one language) are uploaded, the application automatically decompresses the files and gives us a count of all uploaded documents and the number of words (tokens) contained in the documents.

The general workflow in *ProTermino* is shown in Figure 1.

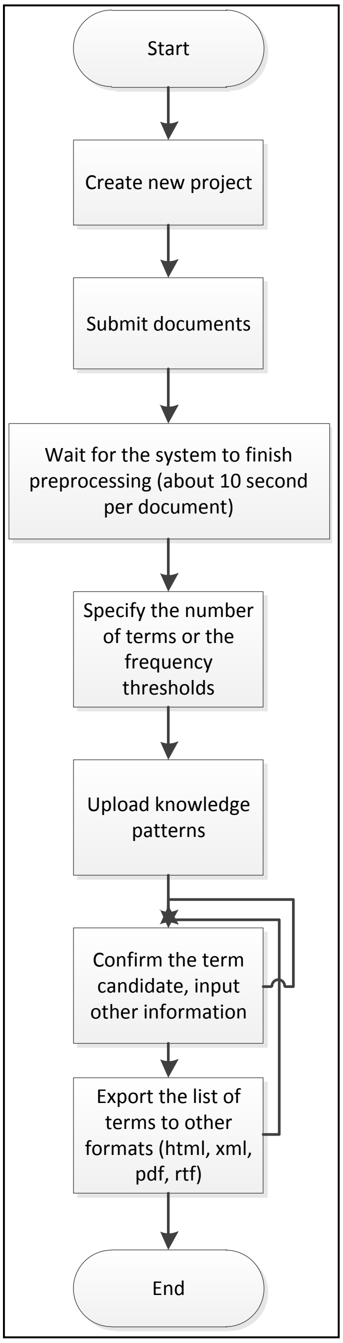

Start
Create new project
Submit documents
Wait for the system to finish preprocessing (about 10 second per document)
Specify the number of terms or the frequency thresholds
Upload knowledge patterns
Confirm the term candidate, input other information
Export the list of terms to other formats (html, xml, pdf, rtf)
End

Figure 1. *ProTermino* workflow

After the completion of these stages, the user can proceed in the corpus module, either by checking the uploaded documents or by extracting term candidates.

4.2.1. *Extraction of term candidates*

The possibility of terminology extraction within the same application greatly facilitates the work of terminographers and, thus, gives advantage compared to other similar applications that currently exist and which do not offer this option.

This tool automatically extracts term candidates, both complex and single words in the selected working languages. The extraction is carried out on comparable corpora, i.e. corpus containing solely original texts, by using a hybrid technique, which combines a statistical approach technique and a linguistic approach (cf. Mitkov et al. 2007). On the one hand, the technique based on the statistical approach is the TF.IDF (*Term Frequency-Inverse Document Frequency)*, a technique for assessing the importance of a candidate term by its frequency relative to a body of texts containing the selected unit. The frequency of a term (TF) corresponds to the number of occurrences of this term in a document corpus of work, and, meanwhile, the inverse document frequency (ITF) filters and discards units with more repetitions in the documents, which are usually articles, prepositions, conjunctions, etc., so they are not taken into account during the extraction process and units with fewer repetitions are given greater weight. Thus, the larger the number of appearances of a unit, the lower the score obtained with this statistical measure. Furthermore, the technique of linguistic approach, which must be carried out prior to the application of the statistical technique, consists in the morphological labelling of units by means of the *TreeTagger* application. Accordingly, the units of the corpus are labelled according to their grammatical category and, thus, allowing subsequent removal from a set of syntactic patterns in combination with the TF.IDF statistical technique. These syntactic patterns included in the application are referenced to nouns, which are considered the units with higher semantic content. As a matter of fact, all units automatically extracted correspond to a noun, if concerned with the extraction of 1-gram, or a noun phrase to be extracted if 2 or more -grams. In this second case, we find the noun phrase: 1. N (noun) + prep (preposition) + N (noun), N + Adj. (adjective) or Adj. + N, among others.

Despite the complexity of the automatic terminology extraction process, users only have to make a very simple action to implement it. Once the corpora have been uploaded to the tool, users must click on the *Recalculate term score* tab, and then select one of the two extraction options available, namely:

1. the minimum number of term candidates after extraction for each language, that is, users must indicate the total number of term candidates they want to have at the end of the extraction (without taking into account other aspects), or
2. minimum frequency threshold of occurrence of each term candidate, i.e., users must specify the number of repetitions the terms extracted must have (units under this number of repetition will be discarded).

These two options will provide different results, so it may be also interesting to try both and check results. As an example, if we select the first option and decide that the total number of units extracted for each language should be 2,000, we introduce this number in the corresponding field and, as a result, the application performs an automatic extraction of term candidates and provides 6,000 units, 2,000 for each working language (in case we are working with 3 languages, say English, Spanish and German).

Once extracted, the application will display all the extracted candidate terms organised by languages, either by frequency or alphabetically according to users' preferences. At this step, users must check the extracted units and validate or delete both non-terms or term that are not relevant to the working domain. Some noise can be detected regarding poorly constructed noun phrases, prepositions, conjunctions, irrelevant verbs, among others, and this is why users need to check the results.

In addition to this functionality of automatic term candidate extraction, this application allows the manual inclusion of terminological units in case users deem it necessary to include some units after having performed the extraction. To do so, users would have to manually introduce the term, the language of the unit, together with the relevant information in the fields of the term entry.

Once this step is performed and saved in the database, the manually entered unit appears in the list together with the other units automatically extracted.

At this point, terminographers are able to proceed according to their needs, either they can start by filling in term entries in the Terminological Database module or, preferably, uploading semantic patterns in the Knowledge Pattern Module and launching the search for semantic relations in the corpus.

4.2.2. *Inclusion of semantic patterns and extraction of relations*

As discussed in section 4.1., users can upload knowledge patterns in the Knowledge Pattern module and launch the search for semantic relations in the corpus. These steps can be carried out in all the working languages and to all the comparable corpora uploaded in the project. According to several authors (Termmerman 2007; Roche 2003, Durán-Muñoz 2012) ontologies and knowledge representation bring a great number of advantages to the terminology field and help terminographers organise specialised domains and detect gaps in conceptual information.

The *ProTermino* tool has been designed in the ontoterminography framework and, as such, it aims at facilitating the acquisition and organisation of conceptual knowledge based on the working corpus in several ways.

The first step that must be performed in this phase is the introduction of semantic patterns Knowledge Pattern module, which must have been detected during the term extraction or based on previous analysis. To do so, we simply need to upload a text file format (.txt) containing the patterns for each working language following this specific structure:

```
Language,RELATION, pattern

Spanish,ES_UN,es un
Spanish,ES_UN,es una
Spanish,ES_UN,es una forma de
Spanish,ES_UN,es una variante de
Spanish,ES_UN,es una especialidad de
Spanish,ES_UN,es una modalidad de
Spanish,ES_UN,es un deporte
```

As a result, we automatically obtain a table consisting of three columns in which the information uploaded is displayed according to the three kinds of information: language, relation, and pattern.

From the moment in which these patterns are available in the application, *ProTermino* works fully automatically and proposes semantic relationships between terms encountered in the working corpora. These detected semantic relations are proposed to the user, who is able to confirm, modify or delete them. At this point, users can also add examples that have not been detected in the corpora, in case they consider it necessary.

Once the proposals are confirmed, they are automatically included in the table of relations as confirmed relations, placed in the term entry corresponding to the terms involved, from which they can be also removed.

The user can also visualise the confirmed relations in the form of a graph:

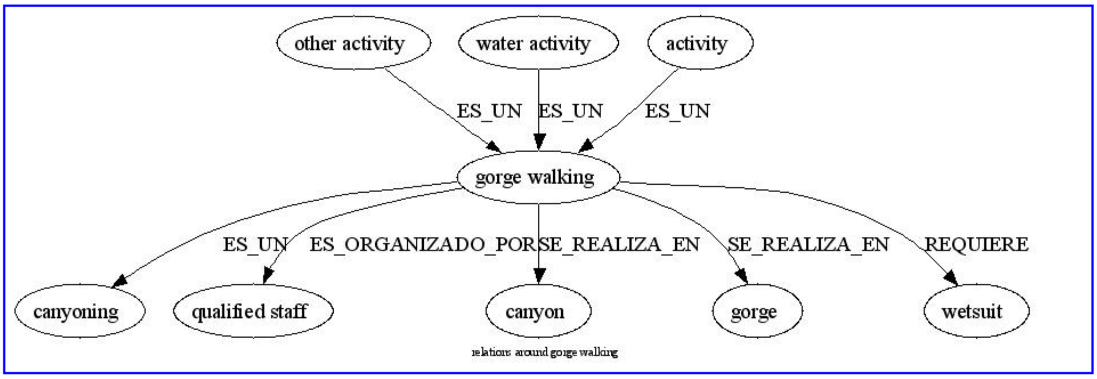


Figure 2. *ProTermino* graph.

Although this is not an accurate ontology graph, since it lacks the appropriate hierarchies as well as the vertical relations between concepts, we believe, however, that it is very helpful when fetching instances (real examples) from the corpus and, especially, when writing definitions. Also, it allows users to acquire deeper knowledge about the working domain by detecting new concepts, concept relations, and conceptual organisation.

4.2.3. *Development of the ontoterminographical database*

In order to develop the ontoterminographical database, we need to follow three main steps: namely, 1. Selection of terminological fields, 2. Writing definitions, and 3. Selection of contextual examples. In this context, *ProTermino* assists terminographers in all these steps in the Terminological Database module, together with the inclusion of information in the term entry according to the target users' needs and the project purposes, i.e. definitions, contexts, equivalence, pictures, semantic relations, among other fields.

According to the first requirement, the selection of terminological fields, we must highlight that the list of fields proposed by the tool are based on the ISO 12620 Computer applications in terminology - Data categories (1999), but also on a previous study conducted to know the target group's preferences and needs, i.e. translators (cf. Durán-Muñoz, 2010). Consequently, the microstructure of the term entry includes the following fields:

1) terminological unit (the term)
2) part of speech
3) gender
4) grammatical number
5) term status
6) standardising entity
7) geographical usage
8) subdomain
9) definition
10) context
11) collocation
12) nontextual illustration
13) equivalence
14) term type*
15) linguistic remark

*whether it is an abbreviation, acronym, full form, scientific term, symbol, synonym, orthographical variant or related term

As stated, these fields are based on translators' preferences and the ISO standard, but they are not compulsory to every terminology project but it is the user who decides which fields are necessary according to their target users and project purposes.

The *Equivalence* field requires some explanation from our behalf, since it is also another innovative functionality by *ProTermino*. This field is intended for the introduction of translation equivalents, but our tool provides equivalent candidates gathered from an automatic extraction. That is, users, instead of searching for appropriate equivalents, are given several proposals by *ProTermino*, which extracts them automatically from the uploaded corpora. As it occurred with term candidates, users can, at any time, validate, delete or modify *ProTermino* proposals, and achieve the best results.

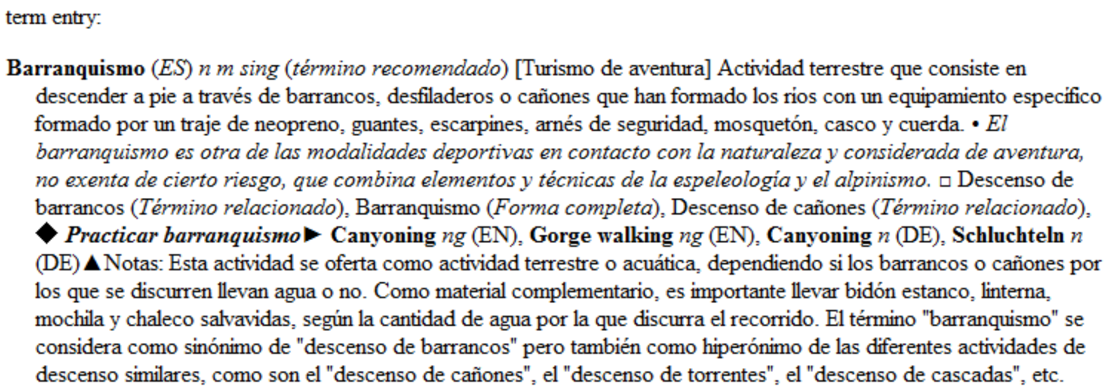
The search and validation of these translation equivalents is one of the most complex parts of a terminology project, since it is here where discrepancies arise between languages. Terms representing concepts vary total or partially due to linguistic, cultural or social aspects. Several techniques are been implemented in NLP to enhance terminographers' tasks dealing with equivalents, but most of them are based on the equivalent extraction from parallel corpora. By contrast, *ProTermino* employs a technique based on cognates, which permits the extraction of equivalent candidates from comparable corpora. The technique used to perform this extraction is called Levenstein distance, also known as edit distance, whose aim is to calculate the differences between two sequences of symbols. More specifically, this technique estimates the number of modifications needed to transform a sequence of symbols or characters into another, either adding, removing or changing one another. In this regard, the smaller the number of modifications required, the greater the similarity between two sequences. Since the development of this technique in the mid-twentieth century, it has been employed in different fields, including computational linguistics and the recognition of cognates between two different languages (cf. McTait, 2001; Mitkov et al., 2008) achieving very positive results.

By implementing this technique in *ProTermino*, equivalent candidates are quicker and more simply encountered thanks to the automatic proposals by the tool. However, we must also highlight the close revision needed after the extraction so as to validate or delete the proposals. Once the extraction has been conducted, the tool provides the different options to the user, who confirm, delete or modify the results. After this step, the results confirmed are automatically incorporated to the term entry in the Equivalence field.

Despite the advantages gained with this functionality, we must also indicate that it is not always possible to locate all equivalents of all units in the given comparable corpus, and neither is it always possible to find the equivalent proposals correct, especially when remote working languages such as German and Spanish are involved. For example, the unit *helmet* in English, the application displayed a number of candidates for Spanish and German, but only in German the correct equivalent was found: *Helm*.

Another important functionality that assists terminographers when working in *ProTermino* is the concordancer application embedded in the tool. The concordancer employed by *ProTermino* is *AntCon*,[7] a freeware concordance program. With this software, users are able to search concordances, collocations, real contexts, and information to write proper definitions and support all the information gathered in the term entry with data from the working corpora. Besides, *ProTermino* makes possible the automatic selection of real contexts, that is, once users encounter good examples in the information provided by the concordancer, this example is automatically included in the Context field of the entry. A simple action that reduces users' manual work and time.

At this point when all the information, according to the target users' and project needs, is included and saved in the database, the definitions are written, the equivalent are confirmed, etc. we reach the final step: validation of the term entries. This is an easy task in *ProTermino*, since the final entry is always displayed on the right side of the screen as follows:



term entry:

Barranquismo (*ES*) *n m sing* (*término recomendado*) [Turismo de aventura] Actividad terrestre que consiste en descender a pie a través de barrancos, desfiladeros o cañones que han formado los ríos con un equipamiento específico formado por un traje de neopreno, guantes, escarpines, arnés de seguridad, mosquetón, casco y cuerda. • *El barranquismo es otra de las modalidades deportivas en contacto con la naturaleza y considerada de aventura, no exenta de cierto riesgo, que combina elementos y técnicas de la espeleología y el alpinismo.* □ Descenso de barrancos (*Término relacionado*), Barranquismo (*Forma completa*), Descenso de cañones (*Término relacionado*), ◆ ***Practicar barranquismo*** ▶ **Canyoning** *ng* (EN), **Gorge walking** *ng* (EN), **Canyoning** *n* (DE), **Schluchteln** *n* (DE) ▲ Notas: Esta actividad se oferta como actividad terrestre o acuática, dependiendo si los barrancos o cañones por los que se discurren llevan agua o no. Como material complementario, es importante llevar bidón estanco, linterna, mochila y chaleco salvavidas, según la cantidad de agua por la que discurra el recorrido. El término "barranquismo" se considera como sinónimo de "descenso de barrancos" pero también como hiperónimo de las diferentes actividades de descenso similares, como son el "descenso de cañones", el "descenso de torrentes", el "descenso de cascadas", etc.

Figure 3. Sample of term entry.

This option greatly facilitates the recruitment of possible errors concerning deficiencies, too much information, misprints, etc. and thus permits a final review of the information quickly and efficiently. Once the review is finished, we need to proceed exporting the database.

4.2.4. *Exporting the database*

To perform this step, *ProTermino* allows multiple exportation formats so that users can determine the required format according to the editing needs and project purposes. In this sense, it permits exportation in .html, .pdf, .html, .rtf and .tbx. In addition, the application offers the possibility to select the terminological fields users want to extract, i.e. it allows users to customise term entries. After having selected the appropriate fields, users select the format for exportation and obtain the exported database in the selected format. From that moment on, this file can be used for editing the end terminological resource, either in paper or electronic format.

In our case, we recommend the TBX format (Termbase eXchange), since it is the terminology exchange standard recommended by international organisations such as ISO. With this format, the possibilities to reuse this information in the future or in another application are greater, as well as the interchange between research groups.

5. Advantages of *ProTermino*

As concluding remarks, we highlight that terminographers are usually obliged to combine a number of tools so as to carry out a terminological project, and even more when their project is knowledge-based. There are some comprehensive systems on the market that intend to satisfy their requirements, such as *Terminus*, *Corpógrafo*, *Ontoterm* or *Termontography Tools*, but they still require the combination of different tools and do not utterly fulfil their needs. *ProTermino* can be considered a suitable solution for knowledge-based terminological projects due to its main advantages:

1. It is a comprehensive tool that allows terminographers to carry out all the main phases of any terminological project based on knowledge representation, from corpus management to export.
2. It is web-based and, as such, it permits remote, simultaneous and multiple-user access.
3. It provides a semantic relation search option based on semantic patterns provided by the user and encountered by means of tools within the system, namely the concordancer and the term extractor.
4. Both term and equivalent extractors provide precise and accurate results in all the working languages.
5. It provides several options in the different modules: choose among languages (Spanish, English, German, Italian, and French), export formats, term fields, corpus uploading, corpus and term checker, among others.
6. It includes a concordancer to search for KWIC in the uploaded corpora.

7. It provides an ontology editor displaying graphs that allow terminographers to visualise knowledge representation and edit it.

We could continue depicting the functionalities and advantages offered by *ProTermino* much further, but due to space constraints this is not possible. In any case, we are concerned about the fact that the previous outline is good enough to prove the validity and suitability of *ProTermino* in the domain of modern terminology.

References

Cabré Castellví, María Teresa. 1993. *La terminología. Teoría, metodología, aplicaciones.* Barcelona: Antártida/Empúries. ISBN 9788475964058.

Cabré Castellví, María Teresa; Rogelio Nazar and Amor Montané. 2012. Computer Assisted Terminology Processing. @ *LREC 2012.* 21st May. Istambul.

De Baer, Peter; Koen Kerremans, and Rita Temmerman. 2006. Developing Special Language Dictionaries with the Termontography Workbench @ *12th EURALEX Conference.* Turín.

Durán-Muñoz, Isabel. 2012. *La ontoterminografía aplicada a la traducción. Propuesta metodológica para la elaboración de recursos terminológicos dirigidos a traductores*. Berlin: Peter Lang. ISBN 9783631631195.

Durán-Muñoz, Isabel. 2010. Herramientas para la gestión y elaboración de recursos ontoterminográficos. @ M. Ibáñez Rodríguez, ed. *Lenguas de especialidad y terminología*. Granada: Comares. 111–123. ISBN 9788498367416.

International Organization for Standarization (ISO). ISO 12620. 1999. *Computer applications in terminology - Data categories*. Ginebra: ISO.

McTait, Kevin. 2001. Linguistic Knowledge and Complexity in an EBMT System Based on Translation Patterns. @ *Proceedings of the Workshop on Example-Based Machine Translation (EBMT)-MT Summit VIII.* Santiago de Compostela.

Mitkov, Ruslan; Richard Evans, Constantin Orasan, Le An Ha and Viktor Pekar. 2007. Anaphora Resolution: To What Extent Does It Help NLP Applications? @ A. Branco, ed. *Anaphora: Analysis, Algorithms and Applications*. Berlín: Springer-Verlag, pp. 179–190. ISBN 9783540714118.

Mitkov, Ruslan; Viktor Pekar, Dimitar Blagoev and Andrea Mulloni. 2008. Methods for extracting and classifying pairs of cognates and false friends. @ *Machine Translation* 21/1: 29–53.

Moreno Ortiz, Antonio. 2000. Diseño e Implementación de un Lexicón Computacional para Lexicografía y Traducción Automática @ *Estudios de Lingüística Española* (*ELiEs*), 9. <http://elies.rediris.es/elies9/index.htm> Retrieved 2011–4–13.

Moreno Ortiz, Antonio. 2004. Representación de la información terminológica en Ontoterm®: Un sistema gestor de bases de datos terminológicas basado en el conocimiento. @ P. Faber Benítez and C. Jiménez, eds. *Investigar en terminología*. Granada: Comares, pp. 25–70. ISBN 8484446328.

Pavel, Silvia and Diane Nolet. 2001. *Handbook of Terminology*. Ottawa-Hull: Translation Bureau. ISBN 0660616165.

Roche, Christoph. 2003. Ontology: A Survey. @ *8th Symposium on Automated Systems Based on Human Skill and Knowledge*, IFAC. 22-24 September. Göteborg (Sweden): 1–6.

Termmerman, Rita. 2000. *Towards New Ways of Terminology Description. The sociocognitive approach.* Amsterdam: John Benjamins. ISBN 9789027223265.

Notes

[1] The research reported in this paper has been carried out in the framework of project BBF2003-04616 (Spanish Ministry of Science and Technology/EU ERDF).

[2] 'For more on these tools, see the relevant websites: *Terminus* (http://igraine.upf.edu/Terminus2/index.html), *Corpógrafo* (http://www.linguateca.pt/corpografo/), *Ontoterm* (http://www.ontoterm.com/), *Termontography Tools* (http://taalkunde.ehb.be/cvc/software) and also see Durán-Muñoz (2010).

[3] <http://www.iula.upf. edu/> Retrieved January 20, 2014.

[4] <http://www.ugr.es/~OncoTerm/> Retrieved January 20, 2014.

[5] <http://taalkunde.ehb.be/cvc/software> Retrieved January 20, 2014.

[6] <http://www.linguateca.pt/corpografo/> Retrieved January 20, 2014.

[7] <http://www.antlab.sci.waseda.ac.jp/software.html> Retrieved January 20, 2014.