Collocational Constructions in Translated Spanish: What Corpora Reveal

Gloria Corpas Pastor[1,2(✉)]

[1] Department of Translation and Interpreting, University of Malaga, Málaga, Spain
gcorpas@uma.es
[2] RIILP, University of Wolverhampton, Wolverhampton, UK

Abstract. In recent years, Construction Grammar has emerged as an enhanced theoretical framework for studies on phraseology in general, and particularly for collocational analysis. This paper aims at contributing to the study of collocational constructions in translated Spanish. To this end, the construction [V PP$_{_de\ miedo}$] is analysed in detail. Our methodology is corpus-based and compares subtitled translations with general Spanish, American Spanish and Peninsular Spanish. The findings suggest that collocational constructions in translated Spanish have a clear preference for the Peninsular standard. They reflect features of translationese, as well as universal traits such as simplification, normalisation, and convergence. Another interesting finding refers to corpus selection, as giga-token corpora appear to provide more fine-grained analysis that conventional, balanced corpora.

Keywords: Collocation · Collocational construction · Translated spanish · Simplification · Normalisation · Subtitled translations · Corpus

1 Introduction

In Phraseology, the term *collocation* refers to a linguistic phenomenon by which words tend to occur together and exhibit idiosyncratic combinatory and semantic properties. The term also covers a type of phraseological unit (*collocation*) and the actual instances (*collocations*). By default, collocations are arbitrary and non-isomorphic, semantically transparent but formally unpredictable. Some examples are *outright insult* (*absolute insult) and *highly intelligent* (*highly unintelligent but *remarkably unintelligent*). Collocation components show a different semantic status: bases are semantically autonomous (*insult, intelligent*), whereas collocates tend to be determined, their actual senses being selected by their bases. Even though collocations undergo some degree of semantic specialisation or grammaticalisation, they differ from other types of phraseological units that exhibit a fixed form and a non-decomposable, unitary meaning. Idioms like *fly off the handle, on cloud nine* or *bite the dust* are semantically opaque and pose both comprehension and production problems.

In computational approaches, the term *collocation* has been used to refer to a distinct type of multiword expression (MWE) that is statistically idiomatic, i.e. a particular combination of words that "occurs with markedly high frequency, relative to the component words or alternative phrasings of the same expression" [1]. For instance, the verbs

R. Mitkov (Ed.): Europhras 2017, LNAI 10596, pp. 29–40, 2017.
https://doi.org/10.1007/978-3-319-69805-2_3

inflict and *impose* are more likely to combine with the noun object *punishment* than *administer*, which in turn is more probable than *prescribe* or *sanction.* Statistical idiomaticity is reminiscent of Halliday's probabilistic definition of collocation [10] and is deeply rooted in the early work on computer-assisted analysis of collocational patterns in large corpora based on frequency [12, 17].

Quantitative methods for the automatic identification and extraction of collocations require large corpora. Corpus-based methods that are based on n-gram frequency can only indentify continuous co-occurrences. This type of collocations have been variously termed *collocational networks*, *lexical bundles*, *clusters*, *recurrent word combinations*, etc. Statistical corpus-based methods use various association measures in order to uncover discontinuous co-occurrences. Hybrid methods rely on linguistic analysis and annotation for refining of results [15]. A more sophisticated version of a hybrid method is *collostructional analysis*, i.e. 'a family of quantitative corpus linguistic methods for studying the relationship between words and the grammatical structures they occur in' [16]. These methods can detect not only discontinuous occurrences of words in various syntactic relationships within a given pattern, but they can also identify words significantly attracted by a particular grammar structure (akin to the notion of colligation) or compare the association strengths of all collocates of two partially synonymous patterns.

Different approaches to collocation agree on co-occurrence and frequency as distinctive features, whether semantically-based, statistically-based or psychologically-based [5]. However, none of those approaches is integrative enough or sufficiently explanatory; nor is there a set of defining features or proper definition of collocation that is generally accepted. In this paper we explore some aspects of the relationship between collocations, idioms, linguistic constructions and grammaticalisation.

The organisation of the paper is as follows. We start with a characterisation of the construal nature of collocations (Sect. 2), with special reference to cross-lingual anisomorphism and potential consequences for translation. Then we provide a case-study in translated Spanish (Sect. 3). Our methodology is corpus-based and compares subtitled translations with general Spanish, American Spanish and Peninsular Spanish. In Sect. 4 we summarise the main findings of the study and some thoughts are presented as how the material discussed might be relevant for further studies on collocational constructions in translated Spanish.

2 Rationale and Background

In recent years, Construction Grammar has emerged as an enhanced theoretical framework for studies on phraseology in general, and particularly for collocational analysis. The constructionist approach views language as an idiomatic continuum of which constructions are the building blocks. Constructions are defined as usage-based pairings of form and (semantic or discourse) function that exhibit different degrees of complexity, schematisation and entrenchment [6, 8]. These symbolic units emerge through repeated experience with actual instances and their generalisations [9]. Frequency plays a key role in the mental representations and storage strength of constructions in the neural network [11].

Collocations possess a distinctive construal nature, as evidenced by their internal lexical restrictions and interpretative accommodation. In this light, collocations can be conceived as partially specified constructions that are semantically predictable. In other words, *collocational constructions* could be described as symbolic units that span various phrasal patterns and contain slots to be filled by a restricted set of lexical items (slot fillers) in a cline of bondness and coercion [3]. The actual instances of collocational constructions would be termed *collocations*. Such a flexible framework fosters a powerful explanatory model of idiomaticity that allows idioms, collocations and other related phenomena to count as constructions in their own right, linked to each other within complex networks.

Collocational peculiarities have serious consequences for cross-language analysis and translation. Monolingual anisomorphism can be observed in (partial) synonyms and lexical sets, as seen in the above examples. It is especially relevant in certain types of semantic processes that are particularly liable to collocational idiosyncrasies, such as intensification. Degree modifiers that refer to a high degree or a high level on a scale are usually lexically restricted to their bases. For instance, adjectives like *huge, tremendous, overwhelming*, *enormous* collocate with *success* ('big success'), but not with *failure*, that usually combines with other intensifiers, such as *complete, utter* and *dismal* in the same sense ('big failure'). Further intervarietal differences arise from discipline-specific collocations and levels of formality: e.g., *give an injection* (general) versus *administer an injection* (medicine), *swear an oath* (formal) versus *take an oath* (neutral), as well as language varieties: e.g., *have a bath, have a rest* (British English) and *take a bath, take a rest* (American English).

To complicate the picture even more, collocational differences are also affected by crosslingual anisomorphism. This phenomenon occurs when the direct translation equivalents of the individual elements of a given collocation in the source language do not constitute collocations in the target language. By way of illustration, consider collocations with *commit*: while *commit a crime/a sin* translate word-for-word in Spanish (*cometer un delito/un pecado*), *commit a mistake* does not translate as **commit a mistake* but as *make a mistake.* As we have previously stated [5], "even completely transparent collocations can pose problems in translation due to the arbitrary, non-isomorfic nature of collocates". This is frequently the case, as collocates are usually polysemous items that depend on their bases for disambiguation and translation (collocation translational equivalents). For instance, the translation into Spanish of collocations with the verb *gain* will depend on its object nouns collocates: ***gain*** *advantage* (***sacar*** *ventaja*), ***gain*** *control* (***hacerse con*** *el control*), ***gain*** *independence* (***conseguir/obtener*** *la independencia*), ***gain*** *port* (***llegar/arrivar*** *a puerto*), ***gain*** *strength* (***cobrar*** *fuerza*), ***gain*** *weight* (***coger*** *peso*), etc. This is the reason why straightforward equivalents (system translation equivalents), such as *gain* ≈ *ganar,* do not hold in translation [5]. In other words, straightforward equivalents when used as individual lexical items may turn into potential false friends as slot fillers of collocational constructions. See, for instance, the large number of bitexts in Linguee where *gain advantage* has been wrongly translated as **ganar ventaja*.

When metaphor is at play, translation choices appear to be even more diverse and complex. For instance, verb-noun collocations with the verbs *kindle* and *spark* are based

on the 'lightning/start a fire' metaphor. However, collocations with *kindle* usually have a positive prosody (e.g., *kindle enthusiasm, interest*), whereas the prosody associated with *sparkle* tends to be negative (e.g., *sparkle outrage, controversy*). Neither figurative metaphors nor prosodies are easily conveyed in the target language. For instance, both verb-noun collocations are primarily translated by the same set of prosody neutral and non-figurative collocates: *causar/suscitar/provocar* + e*ntusiasmo/interés/controversia/indignación*; and secondarily by the collocate verb *despertar* ('wake up, awake'): *despertar entusiasmo/interés/controversia/indignación*. In the second case, the verb *despertar* is prosody-neutral (cf. *despertar* + negative feelings and emotions: *odio/recelos/envidia*) but figurative, although with a different underlying metaphor ('awakening', as opposed to the 'lighting' source metaphor). In addition, other types of differences (diatopic, diastratic, diaphasic) and degree of equivalence may result in cases of infra- or overtranslation. For example, this is the case when collocations pertaining to particular language varieties, levels of formality or specific domains or disciplines are rendered by neutral collocations, and vice versa.

3 Methodology

As a consequence of the translation process, translated texts tend to exhibit characteristic linguistic features, regardless of the source and the target languages. Translations are believed to be simpler, more explicit, closer to the standard prototype and more 'typical' than non-translated texts. These distinctive lexico-grammatical and syntactic characteristics are attributable to widespread translation trends (*universals*) and have been explained by Toury's laws of growing standardisation and interference [20]. The tension between these two laws gives rise to the unique nature of translated language (*translationese*).

This paper contributes to the study of collocational constructions in translated Spanish. To this end, the construction [V $PP_{\_de\ miedo}$] will be analysed in detail. Our starting point will be the lexicographical information provided about this construction by the *Diccionario combinatorio práctico del español contemporáneo* (DCPEC) [2]. This Spanish combinatory dictionary provides a separate entry for the lemma **de miedo**, which is classed as a polysemous idiom with adverbial or adjectival function ("loc. adv./loc. adj."). When combined with verbs *de miedo* has an adverbial function and two main senses: "[de terror]" (lit., 'out of fear') and "[muy bien] (lit., 'very well'). The DCPEC indicates that the first sense is actualised with the verbs *morirse, cagarse, descomponerse*, *encogerse, temblar*; and the second sense, with the verbs *estar*, *pasárse(lo)* and *sentar (a alguien).*

These two types of disambiguating verbs indicated in DCPEC will constitute the list of verbal slot fillers to be analysed against the various corpora of translated and non-traslated Spanish used in this study (see below).

In a previous study [4], we have reported patterns of simplification and normalisation in translated Spanish as regards idiomaticity and diatopy. A similar corpus-based research protocol will be adopted in this study. For the purpose of this study, non-translated Spanish data will be collated from giga-token Web (sub)corpora and then compared

with data stemming from a balanced, conventional reference corpus of Spanish and two subcorpora. Translated Spanish data will be retrieved from a giga-token parallel corpus of fiction subtitles. Slot fillers will be extracted (semi)automatically.

3.1 Corpora

Several (sub)corpora have been selected for the study:

1. ***OpenSubtitles*** – a 8.31 giga-token multilingual parallel corpus that has been downloaded from the *OpenSubtitles.org* repository in 2011 [19]. It comprises 54 languages, but only the bilingual parallel subcorpus has been analysed (50 million aligned sentences of English-Spanish film subtitles). The Spanish component size is over 870 million words.
2. **esTenTen** – a 10.99 giga-token Web corpus of global, standardised Spanish. It was created automatically in 2011 [13]. It comprises the *esEuTenTen* [2011] and the *esAmTenTen* [2011], plus some other documents not classified by their national top level domain (Wikipedia, some Spanish newspapers, etc.).[1]
3. **esEuTenTen** – a 2.3 GT subcorpus of European Spanish (Peninsular variety, 21%).
4. **esAmTenTen** – a 8.6 GT subcorpus of Latin American Spanish (American Variety, 79%). It comprises 18 different varieties that have been identified by their national top-level domains (.ar,.es,.uy,.ve, etc.): Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay and Venezuela.[2]
5. **CORPES XXI** – a pan-Spanish reference corpus of over 225 million words (1975-2017) [14]. It includes Peninsular and American varieties *(esEuCORPES* and *esAmCORPES*).
6. **esEuCORPES** – a subcorpus of Peninsular Spanish (67 million words).
7. **esAmCORPES** – a subcorpus of Latin American Spanish (168 million words). It comprises the 18 Spanish varieties included in esAmTenTen, plus the varieties spoken/written in Puerto Rico, southern parts of United States, Philipines and Ecuatorial Guinea.

Corpora 1–4 are available through SketchEngine [18], whereas corpora 5–7 can be web-searched through an in-built corpus query system.

Web-crawled (sub)corpora (1–4) and conventional (sub)corpora (5–7) offer advantages and disadvantages. The CORPES XXI corpus and its subcorpora have been carefully designed and compiled in order to be representative of the global, standard language spoken/written across the Spanish-speaking world. However, they present several problems [4]: their size is too small to study low-frequency collocational constructions and phraseological units in general, and not all national varieties are sufficiently covered. In addition, the CORPES in-built corpus system is rather unstable and slow in terms of processing, data downloading is not possible and access to the data is not flexible enough. Another shortcoming is that this corpus is under construction, which could compromise

[1] It has been web-crawled with Spiderling, (pre)processed and tagged with Freeling 4.0.
[2] The Spanish varieties spoken in Puerto Rico or southwestern United States are not covered.

data stability and the results since they may vary significantly according to the access date (it is expected to reach over 500 million words in 2018).

By contrast, Web corpora provide a wealth of information thanks to their giga-token size, the stability of the data, the reproducibility of the research, and the reliability of the results [7]. Major drawbacks of corpora 1-4 are the question of 'representativeness' and 'balance' (document selection) and the number of (pre)processing problems they present.

3.2 Results and Discussion

The selected collocational construction (V PP$_{_de\ miedo}$) has been studied in all corpora. This section will discuss the main findings of the study. In order to establish whether Web crawled (sub)corpora provide reliable data, the slot fillers licensed by this particular construction have been checked against the TenTen corpora and the CORPES XXI.

Tables 1, 2 and 3 illustrate raw and normalised frequencies of selected verbal slots (senses 1 and 2) in the esTenTen corpora, the CORPES XXI (general, American and Peninsular Spanish) and the OpenSubtitles corpus. Verbs have been ordered according to normalised frequencies; raw frequencies have been taken into account only when normalised frequencies coincided.

Table 1. Verbal slot fillers in the TenTen corpora (raw and normalised frequencies).

V PP$_{_DE\ MIEDO}$	esTenTen	esEuTenTen	esAmTenTen
[SENSE 1]			
Morirse	2,564	563	1,994
	0.23	**0.24**	**0.23**
Cagarse	1,072	250	820
	0.10	**0.11**	**0.10**
Descomponerse	4[a]	–	4
	0.00	**–**	**0.00**
Encogerse	43	6	37
	0.00	**0.0**	**0.00**
Temblar	1,283	245	1,070
	0.12	**0.10**	**0.12**
[SENSE 2]			
Estar	288	59	3
	0.03	**0.03**	**0.00**
Pasar(se)(lo)[b]	140	302	68
	0.01	**0.13**	**0.01**
Sentar [a alg]	371	59	6
	0.03	**0.03**	**0.00**

[a]The four examples in esAmTenTen and esTenTen are even the same ones.

[b]There are also some cases of *pasar(se)(la)*

Table 2. Verbal slot fillers in CORPES XXI (raw and normalised frequencies).

V PP_DE MIEDO	CORPES XXI	esEuCORPES	esAmCORPES
[SENSE 1]			
Morirse	268	34	163
	0.97	**0.13**	**0.65**
Cagarse	93	10	44
	0.33	**0.04**	**0.17**
Descomponerse	_[a]	–	–
	0.00	**–**	**0.00**
Encogerse	2	–	1
	0.00	**0.0**	**0.00**
Temblar	57	13	28
	0.48	**0.05**	**0.11**
[SENSE 2]			
Estar	3	1	3
	0.01	**0.00**	**0.01**
Pasar(se)(lo)[b]	9	4	68
	0.03	**0.01**	**0.27**
Sentar [a alg.]	4	3	6
	0.01	**0.01**	**0.02**

[a]The four examples in esAmTenTen and esTenTen are even the same ones.

[b]There are also some cases of pasar(se) (la).

Table 3. Verbal slot fillers in OpenSubtitles (Raw and normalised frequencies).

V PP_DE MIEDO		OPENSUBTITLES	
[SENSE 1]		[SENSE 2]	
Morirse	207[a]	*Estar*	14
	0.23		**0.01**
Cagarse	134[b]	*Pasar(se)(lo)*	35
	0.14		**0.04**
Descomponerse	–	*Sentar [a alg.]*	–
Encogerse	9		
	0.01		
Temblar	75		
	0.08		

[a]*Morirse de miedo* (207 occurrences); *estar muerto de miedo* (209 occurrences).

[b]*Cagarse de miedo* (134 occurrences); *estar cagado de miedo* (4 occurrences).

The TenTen corpora provide far more occurrences of individual slot fillers than the CORPES. For instance, there are 2,564 cases of *morirse* in the esTenTen corpus as compared to 188 in the CORPES; or 43 of *encogerse* in the esTenTen, and only 2 in the

CORPES XXI. In general, the American variety (esAmTenTen) appears to be closer to general Spanish than the Peninsular variety. In fact, the general and American varieties share the same rankings:

– [SENSE 1]. 1. *morirse*; 2. *temblar*; 3. *cagarse*; 4. *encogerse*; 5. *descomponerse*
– [SENSE 2]: 1. *sentar*; 2. *estar*; 3. *pasar(lo/la)*

In the Peninsular variety (esEuTenTen) this construction licenses only 4 of the verbal slot fillers (*descomponerse* in not found) for sense 1. The ranking is similar to the other two at the top and bottom positions (1. *morirse*; 4. *encogerse*), but changes at the middle positions: 2. *cagarse*; 3. *temblar* (0.02 difference in normalised frequencies). In the esTenTen the ranking appears completely different for sense 2: 1. *pasar(lo/la)*; 2.*sentar*; 3.*estar*.

The similarities between general Spanish and the American variety in the TenTen corpora might be explained by the high proportion of American Spanish documents (seven billion words) in the general corpus, as compared to less than two billion words of Peninsular Spanish.

Not all 5 verbal slot fillers appear to be licensed for this construction in the CORPES family. The CORPES XXI retrieves only 4, ranked as in the esTenTen: 1. *morirse*; 2. *temblar*; 3. *cagarse*; 4. *encogerse* (*descomponerse* is missing). The American variety contains the same 4 verbal slots as general Spanish; it also coincides in the top and bottom positions of the rank (1. *morirse*, 4. *encogerse*), with a slightly difference in the middle positions (2. *cagarse*; 3. *temblar*; 0.15 difference in normalised frequencies). The Peninsular Spanish variety exhibits less lexical richness and different rank of verbal slot fillers, with the only coincidence of *morir* at the top position: 1. *morirse*; 2. *temblar*; 3. *cagarse* (*descomponerse* and *encogerse* are missing).

The ranking of verbal slot fillers for sense 2 is identical for general Spanish and the two varieties analysed: [SENSE 2] 1. *pasar(lo/la)*; 2. *sentar*; 3. *estar*. This rank coincides with the Spanish variety in the TenTen corpus. A possible explanation could be the different composition of the general corpus, as it also includes varieties spoken in Philippines, Ecuatorial Guinea, Puerto Rico and southern parts of United States. Those corpus components might be closer to the European standard. Or else, it could be explained because of a low number of occurrences (and consequently very low normalised frequencies), which might have compromised the results, due to small coverage and lack of representativeness. For this reason, comparative results below will only take into account the data from the TenTen corpora.

The picture depicted by the OpenSubtitles corpus is quite suggestive (see Table 3).

The rank of slot fillers for sense 1 is identical to the euEsTenTen rank: 1. *morirse*; 2. *cagarse*; 3. *temblar*; 4. *encogerse*. This means that Peninsular Spanish would be the variety preferred in subtitled translations. The ranking of verbal slot fillers licensed by the construction for sense 2 points in the same direction. Compare esEuTenTen: 1. *pasar(lo/la)*; 2.*sentar*; 3.*estar* and OpenSubtitles: 1. *pasar(lo)*; 2.*estar*. The main difference is that subtitled translations do not contain the filler *sentar* and only the pronoun *lo* can be found as direct object of *pasar* in the construction under study. This could be also seen as a trait of simplification (lower lexical richness of subtitled translations).

When individual fillers are considered, the data also suggest that translated Spanish tends to gear towards the Peninsular standard, with some exceptions. For instance, the verb *morir* (OpenSubtitles: 0.23) presents a uniform distribution in all three varieties, that is identical to general Spanish and American Spanish (0.23), just 0.01 less than Peninsular Spanish. By contrast, *encogerse* appears slightly higher (+ 0.01) than in all three TenTen corpora. This might be indicating that *morirse de miedo* could be truly considered pan-Spanish, whereas *encogerse* could be suggesting translationese. Normalised frequencies for the rest of fillers in sense 1 are closer to Peninsular Spanish (*cagarse*: 0.14/0.11, + 0.3 difference; *temblar*: 0.08/0.10, −0.02 difference). And *descomponerse* does not occur as filler in both OpenSubtitles and esEuTenTen, possibly because it indexes general and American Spanish.

A similar situation is presented by the fillers for sense 2. Their normalised frequencies are 0.00 in American Spanish, as they seem to be restricted to the Spanish variety (*estar*: 0.03 and *pasárselo*: 0.03), and, therefore, are present with the same values in general Spanish. Their normalised frequencies in the OpenSubtitles corpus simply show minor differences as regards to Peninsular Spanish: *estar* (−0.02) and *pasárselo* (+0.01). The verb *sentar* is not licensed by this construction in the OpenSubtitles corpus, possibly because it is more frequent in the American variety (0.02) than in Peninsular and general Spanish (0.01). It could be the case that its distributional area be covered by *pasárselo* in subtitled translations.

Other examples of normalisation and simplification can be found as regards the lexical richness of the verbal slot fillers licensed by the collocational construction (V $PP_{\_de\ miedo}$). As we have already mentioned, two slot fillers are missing in the Spanish subtitled translations: *descomponerse* (sense 1) and *sentar* (for sense 2). This could be indicative of lower lexical richness in subtitled translations. A comparison between the fillers licensed in translated and not translated Spanish confirms this assumption. The esTenTen corpus registers up to 25 different verbal types (6 happax legomena) for this construction (sense 1); 21 types in American Spanish (10 happax legomena) and 15 types (6 happax legomena) in Peninsular Spanish. Those verbal fillers in non-translated Spanish function as intensifiers that refer to body reactions to fear, such as shivering, sweating, crying, mictioning, etc. (e.g., *tiritar*, *llorar*, *gritar*, *sudar, estremecerse*, *mearse* ….) or metaphorical ways of expressing having experienced emotions of intense fear (*paralizarse, desmembrarse, agarrotarse, disolverse, desfallecer*, etc.). In OpenSubtitles there are only 9 fillers (1 happax legomena) which refer to body reactions (*orinarse, sudar, chillar*, etc.), and only one refers to consequences after having experienced extreme fear (*paralizarse*).

As to the number of alternative verbs found for sense 2 of this construction, the situation is as follows. The esTenTen corpus registers 10 (4 happax legomena) for general Spanish, 5 (4 happax legomena) for Peninsular Spanish and 9 (6 happax legomena) for American Spanish. Some verbs exhibit a much stronger bond, as they seem to select secondary, figurative senses in this construction: *ir* and *venir*; *caer,* and *quedar* (synonyms of *sentar* [a alg. *algo de miedo*]; and *dar* (*dársele* a alg. algo *de miedo*). Others appear to be used in their literal senses (*jugar, besar, venderse*, etc.) which are then intensified by the fixed part of the construction. In OpenSubtitles there are 6 more verbal types (5 happax legomena), but the choice of fillers is more restricted: the verb

ir (*irle* a alg. algo *de miedo*) seems to occupy the area of *sentar*, and together with *pasar* and *estar* are the verbs primarily licensed by this construction in translated Spanish.

The collocational construction [V PP$_{_de\ miedo}$] shows a process of grammaticalisation by which the PP functions as an adverbial modifier, substitutable by an intensifier adverb or adverbial phrase (e.g., 'very much', 'terribly well'). The PP is perspectivised as in the foreground: an extremely intense emotion which was originally negative but of which only the intensity remains. The degree of lexicalisation of slot fillers is proportional to the degree of grammaticalisation and coerced meanings of this semi-schematic construction.

In this light, *de miedo* is not just an idiom with two different senses, but a semi-schematic collocational construction V PP composed of a variable slot (verb fillers) and a fixed slot (*de miedo*). The choice of verbal slot fillers determines the meaning accommodation of both variable and fixed components. Verbs which denote a physical reaction of weakness or unwellness to the feeling of the emotion trigger a metonymic interpretation of intense fear: e.g., *temblar de miedo* ('tremble with fear') → *tener mucho miedo* ('to be very frightened'). The more intense the fear, the more intense the physical reaction (e.g. *morirse de miedo*, lit. 'die out of fear'). In this case, the interpretation of the verbal filler is coerced by the fixed slot (*morirse* does not literally mean 'die', but be terribly frightened). Bondness and grammaticalisation also affect the interpretation of the fixed slot, which undergoes a process of delexicalisation toward intensification ('in high degree'). Once the fixed slot denotes intensification, it is ready for other verbal slot fillers and further lexicalisations (e.g. *pasarlo de miedo*).

Translated Spanish also reflects this intricate process but in a more restricted way, as regards the lower number slot fillers and the degree of bondness of lexicalised and non-lexicalised verbs (simplification). This grammaticalisation process can be seen in the actual choice of lexical fillers for the construction in Spanish subtitled translations. In this respect, a marked preference for the Peninsular Spanish standard is observed (normalisation).

Simplification seems to be also at work when translation choices and procedures are examined. For instance, we have identified over 50 different ways to express the meaning of 'getting/being terribly frightened' in the English component of OpenSubtitles: e.g., *be shitting, be scared shitless, be scared to death, take a shit, wet one's pants, scare the muggers stiff, shit one's pants, be plain chicken shit, be fucking scared, shit oneself, wait for shit to happen, be fucking scared, pee one's kilt, shit bricks, piss on oneself, be shit-scared, be scared out of one's wits, be piss-scared, crap one's pants, be chicken shit, chicken out, crap in a sock, be really afraid*, etc. The number of examples illustrates the lexical richness of the English subtitles. Many of them also represent creative uses of the language. Interestingly enough, all of them have been translated systematically as *cagarse de miedo*. This makes the Spanish subtitled translations look not only simpler (simplification), but also more homogeneous and closer to the standard (convergence) and more 'typical' or less creative (normalisation).

Finally, this study presents a series of limitations as regards corpora. In addition to the (pre)processing errors of Web-crawled corpora (parsing errors, incomplete deduplication, misrecognition of characters, etc.), *OpenSubtitles* presents problems concerning bitexts (alignment across language pairs only). Another issue is the degree

of comparability and/or (a)symmetry between the Spanish corpora, both translated and non-translated. Besides, the technical constrains of subtitling can influence translators' choices. Elements such as the number of lines in a subtitle, the length of subtitles, the structure of line breaks, the number of characters (per second/line) allowed, etc. are an essential facet of subtitled translations that should be regarded as a differential factor.

4 Conclusion

Construction Grammar provides a powerful explanatory model of idiomaticity that caters for different clines of complexity, formal restrictions, semantic coercion and grammaticalisation processes. In this framework, traditional concepts such as collocation, idiom or phraseological unit converge into collocational constructions.

This paper has examined a particular collocational construction in both translated and non-translated corpora. The analysis reveals that [V PP$_{_de\ miedo}$] has undergone a process of grammaticalisation that has affected bondness and meaning accommodation (coercion) of slots (and fillers) in a gradual way. This provides the basis for the creative choice of lexical fillers, bondness and subsequent semantic change. Translated Spanish also reflects this process but in a more restricted way, as the number of lexicalised slot fillers and choice of actual fillers unveil simplification and normalisation traits. Within this process, translated Spanish tends to shows a clear preference for the Peninsular Spanish standard, as well as other features of translationese.

Finally, the corpus-based analysis has revealed that Web-crawled giga-token corpora, like the TenTen family, enable researchers to perform more fine-grained analyses and get more representative results than a balanced, reference corpus like CORPES XXI. The future lies with big data.

Acknowledgements. The research presented in this paper has been partially carried out in the framework of the research projects INTELITERM (FFI2012–38881), TERMITUR (HUM2754) and VIP (FFI2016-75831-P).

References

1. Baldwin, T., Kim, S.N.: Multiword expressions. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, 2nd edn, pp. 267–292. CRC Press, Boca Raton (2010)
2. Bosque, I. (dir.) Diccionario combinatorio práctico del español contemporáneo: las palabras en su contexto. Madrid, SM (2006)
3. Pastor, G.C.: Register-specific collocational constructions in english and spanish: a usage-based approach. J. Soc. Sci. **11**(3), 139–151 (2015)
4. Pastor, G.C.: Translating English verbal collocations into spanish: on distribution and other relevant differences related to diatopic variation. Lingvisticæ Investigationes **38**(2), 229–262 (2015)

5. Pastor, G.C.: Collocations in e-bilingual dictionaries: from underlying theoretical assumptions to practical lexicography and translation issues. In: Torner, S., Bernal, E. (eds.) Collocations and Other Lexical Combinations in Spanish. Theoretical and Applied Approaches, pp. 139–160. Routledge, London (2017)
6. Croft, W.: Radical Construction Grammar. In: Hoffmann, T., Trousdale, G. (eds.) The Oxford Handbook of Construction Grammar, pp. 211–232. Oxford University Press, Oxford (2013)
7. Gatto, M.: The 'body' and the 'web': The web as corpus ten years on. ICAME J. **35**, 35–58 (2011)
8. Goldberg, A.E.: Constructions: a construction grammar approach to argument structure. University of Chicago Press, Chicago (1995)
9. Goldberg, A.E.: Constructions at Work: The Nature of Generalization in Language. Oxford University Press, New York (2006)
10. Halliday, M.A.K.: Lexis as a linguistic level. In: Bazell, C.E., Catford, J.C., Halliday, M.A.K., Robins, R.H. (eds.) Memory of John Firth, pp. 148–162. Longman, London
11. Hoffmann, T.: Abstract phrasal and clausal constructions. In: Hoffmann, T., Trousdale, G. (eds.) The Oxford Handbook of Construction Grammar, pp. 307–328. Oxford University Press, Oxford (2013)
12. Jones, S., Sinclair, J.: English lexical collocations. A study in computational linguistics. Cahiers de Lexicology **24**, 15–61 (1974)
13. Kilgarriff, A., Renau, I.: esTenTen, a Vast Web Corpus of Peninsular and American Spanish. Procedia Soc. Behav. Sci. **95**, 12–19 (2013)
14. Real Academia Española (n.d.). Banco de datos (CORPES XXI). Corpus del español del siglo XXI. http://www.rae.es. Last accessed 10 Aug 2017
15. Seretan, V.: Syntax-Based Collocation Extraction. Springer, Dordrecht (2011)
16. Stefanowitsch, A.: Collostructional analysis. In: Hoffmann, T., Trousdale, G. (eds.) The Oxford Handbook of Construction Grammar, pp. 290–306. Oxford University Press, Oxford (2013)
17. Stubbs, M.: Two quantitative methods of studying phraseology in English. Int. J. Corpus Linguist. **7**(12), 215–244 (2002)
18. SketchEngine Homepage. https://www.sketchengine.co.uk. Last accessed 16 Aug 2017
19. Tiedemann, J.: News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) Recent Advances in Natural Language Processing V. Selected Papers from RANLP 2007, pp. 237–248. Amsterdam and Philadelphia, John Benjamins (1999)
20. Toury, G.: Descriptive Translation Studies and Beyond. John Benjamins, Ámsterdam (1995)