Searching for patterns in the transfer of multiword units: a corpus-based contrastive study on secondary term formation

Lara Sanz Vicente

Erasmushogeschool Brussel, Brussels, Belgium
marialara.sanz.vicente@ehb.be

Abstract. The dominance of English in specialised communication is currently emphasizing the importance of secondary term formation. In this respect, studying the way English multiword terms are transposed into other languages becomes of great interest. This paper reports on a corpus-based contrastive study that describes how multiword terms are formed in English and transferred into Spanish in the field of remote sensing of forest fires. The study particularly focuses on identifying patterns among these units and their language equivalents. The results reveal the existence of certain regularities which could be useful when transferring other multiword terms, but also report on the great structural diversity of the equivalents found for each source-language term.

Keywords: secondary term formation, multiword terms, corpus-based study, transferring procedures

1 Introduction

Secondary term formation, defined by Sager [17] as the process that 'occurs when a new term is created for a known concept [...] as a result of knowledge transfer to another linguistic community', is closely related to the transfer of multiword terms. These units derive from different formation procedures, the most frequently used being the addition of modifiers to an already existing term to reflect its specific properties [7, 17]: *infrared* > *near infrared*, *mid-infrared*, *short wavelength infrared*.

Most of the new terms created are multiword units. Preference for these forms in specialised languages has been noted by many authors, including Sager et al. [18], in the case of English, Kocourek [9] for French and Cabré [3] for Spanish. There are studies that quantitatively measure their importance, according to which they represent around 80% or more of the total vocabulary of certain domains [2]. Furthermore, it has been demonstrated that they are more frequent in highly specialised texts, i.e., texts written by and for experts [8, 15].

From a contrastive perspective, it has been noted that these units are a common cause of trouble in specialised translation, specially between Romance and Germanic languages. Some contrastive studies in multiword units between

English and Spanish are those of Salager-Meyer [19] in the medical field, Pugh [14] and Montero Fleta [11] in IT, Ahronian [1] on the Internet specifically, and Quiroz [15] in a body of texts on the genome. The translation of these multiword units from English into Spanish presents great difficulty due to their syntactic-semantic complexity, the differing syntactic natures of the two languages involved and their word formation rules, but also due to the lack of comparative studies and reference sources to understand and solve them.

The study reported in this paper describes and compares these type of terms in a different and recent field, remote sensing of forest fires, where English is the dominant language, i.e., the language of primary term formation. Using an English-Spanish comparable corpus of research articles, the study wants to go deeper into the knowledge of the secondary term formation process in highly specialised texts. The final aim is to assist translators in the identification, understanding and transfer of multiword units by providing strategies and offering a bilingual database which presents the results derived from the analysis.

The study is based on the belief that the description of the structures of multiword terms (MWTs) – of their morphosyntactic patterns and semantic contents – in a body of real texts can allow us to establish generalisations that increase understanding of these units and offer strategies for their translation. Specifically, the basic hypothesis is that there are certain patterns in the formation of MWTs in English and in their transfer and translation into Spanish. The results presented here derive from a contrastive analysis of MWTs carried out as part of a PhD dissertation. A detailed description of the corpus and methodology and a complete account of the results can be found in Sanz [20].

2 Methodology

2.1 Corpus design and term extraction

The research is based on a descriptive analysis of multiword units and their equivalents in a tailor-made English-Spanish comparable corpus composed of highly specialised texts on remote sensing of forest fires. The corpus compiled contains two subcorpora, English and Spanish, of 193,893 and 128,823 tokens respectively. It is composed of several research papers (35 in English and 38 in Spanish) published between 1992 and 2008 in peer-reviewed journals and conference proceedings, dealing with a specific subfield, burned area mapping. In both languages, it contains original texts – not translations – of similar characteristics and compositions as regards text type and origin, topic, size, setting, date of publication, etc., meaning that cross-linguistic comparisons can be drawn. Topic has been a relevant criterion to guarantee comparability between subcorpora. Research papers had to contain a set of keywords in the title, abstract and within the full text in order to be selected.

A collection of 12 glossaries and dictionaries of different types and sizes was also formed manually with the objective of contrasting and complementing the results obtained from the corpus: half of them specialised in the field of remote

sensing, three in forestry sciences and forest fires, and the rest concerning larger fields including or relating to remote sensing, such as geomatics and aerospace sciences and technology. Five of them are monolingual English dictionaries, three Spanish-English, one English-Spanish, two French-English, and one multilingual (French, Catalan, Spanish, Galician, Italian, Portuguese and English).

The selection of MWTs from the corpus was conducted with the aid of WordSmith Tools [21] and was done in a two-step process: first, drawing up a list of English MWTs and then, attempting to identify correspondences in the corpus of texts in Spanish.

The extraction from the English corpus was mainly based on drafting single-word and multiword wordlists (lists of clusters)[1] and concordance searches using both automatic and manual processes at all times.

The first step was the extraction of English word clusters. We computed 2–12 word clusters with a minimum frequency of occurrence of 3 in the corpus and with a relevant distribution through it, i.e., involved in at least three different texts. From the list obtained, only noun sequences were selected with the aid of a single-word frequency list filtered using a stoplist of high frequency words without specific meaning (articles, pronouns...) that were excluded. Generating single-word frequency lists helped in finding the most relevant units in the field and identifying keywords that could possibly work as nuclei or modifiers of multiword term candidates. The resulting candidates were then grouped into lemmas (*active fire*, *active fires*; *burned area*, *burnt area*; etc.), and inflectional and orthographic variants were also detected and grouped under the same heading.

A total of 460 English MWTs with different levels of lexicalisation were subsequently identified, precisely those complying with he characteristics linked to MWTs, which refer to: i) its morphological structure, formed by a noun (nucleus) accompanied by one or several modifiers, ii) its unity and semantic specificity within the conceptual system of the targeted specialised field, iii) its syntactic function as a minimum independent component of a sentence, and iv) its proximity to specialised phraseological units. This was done not only by producing concordances but also with the help of the dictionaries and glossaries collected and by consulting experts in the field.

The process of searching for equivalents within the Spanish subcorpus was based on compiling concordance lists from possible translations into that language of the 460 English MWTs identified and from the possible translations of their nuclei or modifiers or their most representative collocates. This method is closely related to those proposed for automatic extraction of bilingual terminology from comparable corpora. Most of them are based on the idea that across languages there is a semantic correlation between the co-occurrences of words that are translations of each other [4–6, 16]. Searching for equivalents was also supported by single-word and multiword wordlists in Spanish and with the help of the glossaries and dictionaries collected. It enabled us to find corresponding Spanish terms for 80% of the English MWTs.

[1] In WordSmith Tools multiwords are called *clusters* and defined as 'words which are found repeatedly together in each others company, in sequence' [22].

2.2 Data analysis

The analysis performed centered, first, and both for English and Spanish MWTs, on the manual description of the morphologic structure and substructure of each term (*burned area mapping* > Adj+N+N > Adj$_{-ed}$+N+N$_{-ing}$), and on the identification of the role played by each component element (nucleus or modifier) to be able to represent their morphosyntactic scheme and intraterm semantic relation too (*burned area mapping* > [(Adj+N)$_{Mod}$+N$_{Nuc}$] > PATIENT – ACTION). The intraterm semantic relations of the multiword terms were manually identified and classified using Oster's typology of semantic relational schemas [12, 13] – slightly modified to take account of all of the relationships observed in MWTs in the field under study[2]. The analysis and understanding of the internal syntactical-semantic structure of these units was thus considered an essential step which first required the identification and categorisation of the modifiers linked to the nucleus or nuclei. The information on syntactic and semantic relationships could only be recovered by returning to the context (the text) in which the term was produced and is used, taking all extralinguistic parameters involved into account.

A comparative analysis was carried out afterwards between the English MWTs and their equivalents, which were interpreted as translation equivalents. This analysis was performed in the English-Spanish direction by describing the equivalents of the English MWTs as regards their morphosyntactic and semantic structure and the influence of English in them.

We compared the morphological and morphosyntactic structure of the English MWTs and their equivalents, and how the English MWTs' intraterm semantic relationships materialised in Spanish. That involved studying the correlation between the English MWTs' semantic relationship and the form of the equivalent terms in Spanish.

Finally, the Spanish equivalents were classified according to the strategies applied when importing them into Spanish. This classification, specifically defined for this analysis, included ten basic procedures: borrowing, calquing, paraphrasing, adaptation, transposition, modulation, synonymy, clarification, shortening and endogenous formation, and paid special attention to calquing, as the most important procedure regarding the transfer of MWTs. The classification, therefore, differentiates between calques of expression and structural calques, which, in turn, have been subdivided into two groups: full translation (literal or free) and half translations (literal or free). Attention was also drawn to the procedures most frequently used to import each of the elements of the MWTs separately.

[2] Oster [12, 13] defines semantic intraterm relations as the semantic relation between two concepts *a* and *b* expressed through the combination of the functions carried out by *a* and *b* with respect to each other. For example, *burned area mapping* will be understood as a PATIENT – ACTION relation, where *mapping* performs an action on the patient, *burned area.*

3 Results of the English-Spanish contrastive study

The comparison of the English MWTs' structures with those of their Spanish equivalents demonstrated that there are certain regularities in the translation of these units. As shown in Table 1, of the 30 solutions observed in Spanish for the English morphosyntactic structure [$N2_{Mod}$+$N1_{Nuc}$], the most frequent one, four are highly productive, accounting for more than 70% of the equivalents produced using this English construction: [$N1_{Nuc}$+(prep+N2)$_{Mod}$] (EN. *brightness temperature* → ES. *temperatura de brillo*), [$N1_{Nuc}$+(prep+art+N2)$_{Mod}$] (EN. *infrared band* → ES. *banda del infrarrojo*), [N_{Nuc}+Adj_{Mod}] (EN. *cloud pixel* → ES. *píxel nuboso*) and [$N1_{Nuc}$+$N2_{Mod}$] (EN. *difference image* → ES. *imagen diferencia*). Equally, the second most frequently-used structure in English, [Adj_{Mod}+N_{Nuc}], is matched with the reverse structure [N_{Nuc}+Adj/Pp_{Mod}] in 55% of cases in Spanish (EN. *ancillary data* → ES. *datos auxiliares*, EN. *contaminated pixel* → ES. *píxel contaminado*) and in 10% as [$N1_{Nuc}$+(prep+N2)$_{Mod}$] (EN. *contextual algorithm* → ES. *algoritmo de contexto*).

Table 1. English-Spanish structure correspondences of N+N and Adj+N English multiword terms

English multiword terms		**Spanish equivalents**			
Morphological structure	Morphosyntactic structure	Morphological structures	Morphosyntactic structure	N.	%
N+N	[$N2_{Mod}$+$N1_{Nuc}$]	N+prep+N	[$N1_{Nuc}$+(prep+N2)$_{Mod}$]	85	26.23
		N+prep+art+N	[$N1_{Nuc}$+(prep+art+N2)$_{Mod}$]	73	22.53
		N+Adj	[N_{Nuc}+Adj_{Mod}]	43	13.27
		N+N	[$N1_{Nuc}$+$N2_{Mod}$]	40	12.35
		N	[N_{Nuc}]	13	4.01
		other (25)		70	21.61
Adj+N	[Adj_{Mod}+N_{Nuc}]	N+Adj	[N_{Nuc}+Adj_{Mod}]	80	47.06
		N+prep+N	[$N1_{Nuc}$+(prep+N2)$_{Mod}$]	17	10.00
		N+Adv+Pp	[N_{Nuc}+(Adv+Pp)$_{Mod}$]	15	8.82
		N+Pp	[N_{Nuc}+Pp_{Mod}]	14	8.24
		N+prep+art+N	[$N1_{Nuc}$+(prep+art+N2)$_{Mod}$]	6	3.53
		other (17)		38	22.35

N: Noun; Adj: Adjective; prep: preposition; art: article; Pp: Past participle; Adv: Adverb; Mod: Modifier; Nuc: Nucleus

As for the MWT equivalents with three or more elements, it has been observed that their structures vary based on the syntactical dependency shown by the English MWTs. For example, the Adj+N+N MWTs with dependency [(C+B)$_{Mod}$+A_{Nuc}] are generally translated as N+prep+(art)+N+Pp/Adj (EN. *burned area mapping* → ES. *cartografía de (las) áreas quemadas*, EN. *spectral mixture analysis* → ES. *análisis de mezclas espectrales*), while the most frequent solution for compounds Adj+N+N with dependency [C_{Mod}+(B+A)$_{Nuc}$]

is N+Adj+prep+(art)+N (EN. *viewing zenith angle* → ES. *ángulo cenital de observación*).

Furthermore, the analysis by substructures has shown that in those cases where Spanish uses prepositional phrases to add the modifying element to the nucleus, the connecting preposition most often used is *de*, which is used as the wild card preposition sometimes replacing prepositions with a more specific meaning (EN. *omission error* → ES. *error de omisión/error por omisión*).

The analysis of the intraterm semantic relationships showed that the most frequently-used schema in English MWTs, PROPERTY – DETERMINED ENTITY, which is almost always expressed using the structure [Adj_{Mod}+N_{Nuc}], is essentially formulated with the reverse structure in Spanish, [N_{Nuc}+Adj_{Mod}], (EN. *spectral signature* → ES. *firma espectral*). The second most frequent in English, ORIGIN – DETERMINED ENTITY, mainly expressed in that language using the form [N2_{Mod}+N1_{Nuc}] to denominate remote sensing images according to the sensor or satellite they come from, is translated in Spanish as [N1_{Nuc}+N2_{Mod}], using the sensor or satellite's name as a direct modifier (EN. *AVHRR image* → ES. *imagen AVHRR*) or, sometimes, by connecting it with the preposition *de* plus an article (EN. *Landsat imagery* → ES. *imágenes del Landsat*). The third most often used schema, PATIENT – ACTION, expressed in English with N+N compounds (*change detection*) and Adj+N+N (*burned area mapping*), mainly gives rise to prepositional constructions with *de* in Spanish (*detección de cambios, cartografía de áreas quemadas*). In general, it has been observed that prepositional constructions with *de* serve to express all sorts of semantic relationships.

The results of the classification of MWTs by transferring procedures confirmed that the majority of Spanish equivalents (66%) are translated and imported as calques of expression with full translation of the English MWT, literal in most cases (EN. *active fire* → ES. *incendio activo*) and, to a lesser extent, free (EN. *active fire* → ES. *foco activo*). The second most used resource is explicative paraphrasing (13%), which reformulates the meaning of the English term (EN. *burn signal* → ES. *señal procedente de las áreas quemadas*). In third place, with 5%, are calques of expression containing unadapted loans (mainly initialisms and acronyms) which consist of a literal translation (EN. *AVHRR image* → ES. *imagen AVHRR*). These are followed by unadapted loans, which are not very numerous (4%) and which mainly correspond to the proper names of sensors and satellites expressed as initialism compounds (*NOAA-AVHRR, NOAA-11, Landsat ETM+*) and to some image analysis and interpretation techniques (*Maximum Value Composite, Normalized Burn Ratio*).

As regards the procedures most often used to import each of the elements of the MWTs separately, three are noteworthy: i) transpositions, among which changes from singular to plural prevail (EN. *cloud shadow* → ES. *sombra de nubes*) and noun to adjective (EN. *azimuth angle* → ES. *ángulo acimutal*); ii) clarifications, which involve the inclusion of some elements that were implicit in the English forms, such as prepositions (EN. *colour composite* → ES. *composición en color*); and ii) modulation, based, above all, on the use of partial synonyms (EN. *statistic* → ES. *índice*).

4 Conclusions

The results reveal the existence of certain regularities which guide the transposition of these MWTs into Spanish and which could be therefore useful in translation. Generalising greatly, it could be concluded that the prepositional construction N+*de*+N is mainly used to translate N+N English MWTs and N+Adj to translate Adj+N MWTs. This data, set out in this manner, could lead some to believe that a linear translation rule (right to left) exists, as suggested in some English-Spanish translation manuals [23, 10].

However, comparing the structures of the English MWTs with those of their Spanish equivalents clearly shows, that for each English structure there are many divergent structures in Spanish. The English structure N+N alone has up to 30 different corresponding structures in Spanish. Furthermore, where the MWT features two or more premodifiers in English, its Spanish equivalents' structures vary more widely mainly due to an increase in the variety of possible translations for each source-language term (EN. *burned area mapping algorithm* → ES. *algoritmo para la cartografía de áreas quemadas*, *algoritmo para cartografiar áreas quemadas*, *algoritmo para la producción de mapas de área quemada*), and sometimes because of difficulties in understanding English units (EN. *maximum value composite* → ES. *composición del máximo valor*, **máximo valor compuesto*).

Besides, translation arises as the most important procedure in transferring English MWTs to Spanish. The results have shown that the preferred mechanism in importing these units into Spanish is calques of expression, i.e., a mechanism that respects the syntactic structures of the target language and, more specifically, that consists of a literal translation of the English MWT. This demonstrates the influence English has on Spanish formation of these units within the area being studied. This preference for calques (loan translations) means we should consider to what extent they act as a terminologically innovative and enriching element in the language of secondary word formation.

References

1. Ahronian, C.: Les noms composés anglais français et espagnols du domaine d'Internet. PhD thesis, Université Lumière-Lyon 2 (2005)
2. Boulanger, J.C., Nakos-Aupetit, D.: Le syntagme terminologique: bibliographie séléctive et analytique 1960-1988. Reference materials - bibliographies – multilingual/bilingual materials, Centre international de recherche sur le bilinguisme (CIRB), Université Laval, Québec (1988)
3. Cabré, M.: Terminology. Theory, Methods and Applications. John Benjamins, Amsterdam/Philadelphia (1999)
4. Déjean, H., Gaussier, E., Sadat, F.: An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In: COLING'02 Proceedings of the 19th International Conference on Computational Linguistics., Taipei, Taiwan (2002) 1–7
5. Fung, P.: Parallel text processing. In Véronis, J., ed.: A statistical view on bilingual lexicon extraction - from parallel corpora to nonparallel corpora, Dordrecht, Kluwer Academic Publishers (2000) 1–17

6. Gaussier, É., Renders, J.M., Matveeva, I., Goutte, C., Déjean, H.: A geometric view on bilingual lexicon extraction from comparable corpora. In: Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL). (2004) 526–533
7. Guilbert, L.: La dérivation syntagmatique dans les vocabulaires scientifiques et techniques. In Dabène, M., Gaultier, M.T., eds.: Les langues de spécialité. Analyse linguistique et recherche pédagogique. Actes du Stage du Saint-Cloud, 23-30 Nov. 1967, Strasbourg, AIDELA (1970) 116–125
8. Horsella, M., Pérez, F.: Nominal compounds in chemical English literature: Towards an approach to text typology. English for Specific Purposes **10**(2) (1991) 125–138
9. Kocourek, R.: La langue française de la technique et de la science. 2nd (1991) edn. Brandstetter Verlag, Wiesbaden (1982)
10. López-Guix, J.G., Minett, J.: Manual de traducción: inglés/castellano. Gedisa, Barcelona (1999)
11. Montero-Fleta, M.B.: Technical communication: complex nominals used to express new concepts in scientific English. The ESP **17**(1) (1996) 57–72
12. Oster, U.: Las relaciones semánticas de términos polilexemáticos. Peter Lang, Frankfurt am Main (2005)
13. Oster, U.: Classifying domain-specific intraterm relations: a schema-based approach. Terminology **12**(1) (2006) 1–17
14. Pugh, J.: Contrastive conceptual analysis of noun compound terms in English, French and Spanish within a restricted, specialized domain. In Hartmann, R.R.K., ed.: Lexeter'83 proceedings. Papers from the International Conference on Lexicography at Exeter, 9-12 Sep. 1983, Tübingen, Max Niemeyer Verlag (1984) 395–400
15. Quiroz-Herrera, G.Á.: Los sintagmas nominales extensos especializados en inglés y en español: descripción y clasificación en un corpus de genoma. PhD thesis, Universitat Pompeu Fabra (2008)
16. Rapp, R.: Automatic identification of word translations from unrelated English and German corpora. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL), Maryland, Association for Computational Linguistics (1999) 519–526
17. Sager, J.C.: A Practical Course in Terminology Processing. John Benjamins, Amsterdam/Philadelphia (1990)
18. Sager, J.C., Dungworth, D., McDonald, P.F.: English Special Languages: Principles and practice in science and technology. Brandstetter Verlag, Wiesbaden (1980)
19. Salager-Meyer, F.: Syntax and semantics of compound nominal phrases in medical English literature: a comparative study with Spanish. English for Specific Purposes Newsletter **95** (1985) 6–11
20. Sanz-Vicente, M.L.: Análisis contrastivo de la terminología de la teledetección. La traducción de compuestos sintagmáticos nominales del inglés al español. PhD thesis, Universidad de Salamanca (2011)
21. Scott, M.: WordSmith Tools (version 5). Lexical Analysis Software, Liverpool (2008)
22. Scott, M.: WordSmith Tools Help. Lexical Analysis Software, Liverpool (2011)
23. Váquez-Ayora, G.: Introducción a la traductología; Curso básico de traducción. Georgetown U.P. (1977)