Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2358–2367
Marseille, 11–16 May 2020

Representing Multiword Term Variation in a Terminological Knowledge Base: a Corpus-Based Study

Pilar León-Araúz, Melania Cabezas-García, Arianne Reimerink
Department of Translation and Interpreting, University of Granada
Buensuceso 11, Granada (Spain)
{pleon, melaniacabezas, arianne}@ugr.es

Abstract

In scientific and technical communication, multiword terms are the most frequent type of lexical units. Rendering them in another language is not an easy task due to their cognitive complexity, the proliferation of different forms, and their unsystematic representation in terminographic resources. This often results in a broad spectrum of translations for multiword terms, which also foment term variation since they consist of two or more constituents. In this study we carried out a quantitative and qualitative analysis of Spanish translation variants of a set of environment-related concepts by evaluating equivalents in three parallel corpora, two comparable corpora and two terminological resources. Our results showed that MWTs exhibit a significant degree of term variation of different characteristics, which were used to establish a set of criteria according to which term variants should be selected, organized and described in terminological knowledge bases.

Keywords: multiword expressions and collocations, information extraction, lexical database

1. Introduction

Multiword terms (MWTs) are especially inclined to term variation or the coexistence of different denominations. Rendering them in other languages is not easy because of their cognitive and structural complexity as well as their unsystematic treatment in terminographic resources. This results in a wide variety of translation solutions, which reflect a significant degree of term variation when translating MWTs.

We carried out a quantitative and qualitative analysis of the Spanish translations of a set of 98 English environment-related MWTs and their variants, with a focus on translation variation. For this purpose, we evaluated translation equivalents in three parallel corpora, two comparable corpora and two terminological resources. The objectives of this study were the following: (1) to define a typology of term and translation variants of MWTs in the environmental domain; (2) to evaluate term variation of Spanish MWTs in different contexts (translation, bilingual or multilingual lexicography, and original production in Spanish); and (3) to propose a way of representing MWT variation in terminological knowledge bases (TKBs). Our results showed that MWTs exhibit a high degree of term variation of different characteristics, which is particularly present in translation scenarios (i.e. parallel corpora). These characteristics were used to establish a set of criteria according to which variants should be selected, organized and described in a TKB such as EcoLexicon (https://ecolexicon.ugr.es; Faber et al. 2014, San Martín et al. 2016).

The rest of this article is organized as follows. Section 2 gives a brief account of term variation. Section 3 describes the resources and methodology of this study. Section 4 discusses MWT term variation in the environmental domain; and Section 5 presents a proposal for the representation of MWT variation. Finally, Section 6 lists the conclusions derived from this research and plans for future work.

2. Term Variation in Terminology and Translation

Term variation occurs when different designations are used to name the same concept (e.g. *anemometer* and *wind gauge*). This phenomenon often occurs in MWTs (Bowker 1998; Fernández-Silva and Kerremans 2011; Daille 2017; Giacomini 2018; Gledhill and Pecman 2018; *inter alia*). In MWTs, a head is complemented by one or several modifiers (e.g. *wind velocity meter*), which means that the longer the term, the more variants it may generate.

Although the use of one variant or the other might seem arbitrary, there are certain patterns of variation that need to be explained (Rogers 1997). On the one hand, variation may be a purposeful activity (Bowker 1998; Kerremans 2017; Gledhill and Pecman 2018); and on the other, it may reveal the neological nature of certain concepts (Cabré 1993; Picton 2011).

Traditionally, variation has been explained by means of user-based reasons (i.e. temporal, geographic, or social variation) or usage-based motivations (i.e. field, tenor, and channel) (Gregory and Carroll 1978). Nevertheless, additional reasons can be involved. As Freixa (2006) points out, causes for term variation can be (1) dialectal, caused by different origins of the authors; (2) functional, resulting from different communicative registers; (3) discursive, due to different stylistic and expressive needs of the authors; (4) interlinguistic, caused by contact between languages; and (5) cognitive, resulting from different conceptualizations and motivations. Several of these reasons can also co-occur.

Although term variation may not have cognitive consequences (e.g. *river deposit* and *fluvial deposit*), this is not always the case, for example, when there is a shift in perception along with the change in form (Fernández-Silva 2018) (e.g. *river arch* and *navigation arch*). In this line, Aguado de Cea and Montiel-Ponsoda (2012), and Fernández-Silva (2018) assert that term variants can exhibit minimum, medium, or maximum semantic distance.

As for the different types of term variant, several taxonomies have been developed. The following is Faber and León-Araúz's (2016), which encompasses different proposals found in the literature:

(A) Orthographic variants that are not affected by geographic origin and do not alter semantics or the communicative situation, e.g. *groundwater, ground water*.

(B) Diatopic variants:

(i) Orthographic variants that do not affect semantics, e.g. *fecal*, *faecal*.

(ii) Dialectal variants, which may affect semantics if cultural factors are involved, e.g. *gasoline*, *petrol*.

(iii) Culture-specific variants, which affect semantics and the communicative situation, e.g. *dry lake*, *sabkha*.

(iv) Calques, which may affect semantics and the communicative situation, e.g. *environmentally hazardous substance > sustancia ambientalmente peligrosa*, *sustancia peligrosa para el medio ambiente*.

(v) Borrowings, which may affect semantics and the communicative situation and may be adapted or not, e.g. *smog > smog*, *esmog*

(C) Short form variants, which affect the communicative situation:

(i) Abbreviation, e.g. *greenhouse gas*, *GHG*.

(ii) Acronym, e.g. *laser, Light Amplification by Stimulated Emission of Radiation*.

(D) Diaphasic variants:

(i) Scientific variants, which affect the communicative situation:

-Scientific names, e.g. *Dracaena draco*, *drago*.

-Expert neutral variants, e.g. *Ocellaris clownfish*, *Amphiprion ocellaris*.

-Jargon, e.g. *lap-appy*, *laparoscopic appendectomy*.

-Formulas, e.g. *carbon dioxide*, CO_2.

-Symbols, e.g. *€*, *euro*.

(ii) Informal variants, which affect the communicative situation and may also influence semantics:

-Lay user variants, e.g. *Dragon tree*, *drago*.

-Colloquial variants, e.g. *motor vehicle pollution*, *car pollution*.

-Generic variants, e.g. *pollution, contamination*.

(iii) Domain-specific variants, which may affect semantics or the communicative situation if the domains have different term preferences, e.g. *mud* and *sludge* represent the same concept except that the first one is used in Geology and the latter in Water Treatment.

(E) Cognitive variants, which are usually MWTs:

(i) Dimensional variants, they affect semantics because they express different dimensions of the same concept, e.g. *esmog fotoquímico* [*photochemical smog*], *niebla tóxica estival* [*summer smog*].

(ii) Intentional variants, which affect the semantics and/or the communicative situation, since they are used to cause a reaction in the receiver, e.g. *climate change*, *climate emergency*.

(F) Metonymic variants, which affect semantics by alluding to a part or material of the concept, e.g. *accidental water pollution, accidental marine pollution*.

(G) Diachronic variants, e.g. *anhídrido carbónico* [*carbonic anhydride*], *dióxido de carbono* [*carbon dioxide*].

(H) Non-recommended variants, e.g. since *mental retardation* now has negative connotations, it has been substituted by *intellectual disability*.

(I) Morphosyntactic variants, which do not usually affect semantics but depend on the communicative situation, as well as on term preferences and collocations, e.g. *contaminación acústica* [*acoustic pollution*], *contaminación de ruido* [*noise pollution*].

When exploring term variation in interlinguistic contexts, the notion of 'equivalence' becomes central since terminologists and translators may follow different criteria. While terminologists usually pursue equivalence at the term level with a view to including correspondences in terminographic resources, translators look for equivalence at the sentence or text level. The functional equivalence of their translations is thus essential (Reiss and Vermeer 1984; Nord 1997), instead of a direct term correspondence. For this reason, the use of hypernyms or other variants reflecting different conceptualizations may be justified in a translation equivalence context. Therefore, terminological equivalence does not always correspond to translation equivalence (Kerremans and Temmerman 2016: 59).

Term variation in translation contexts has been explored in Fernández-Silva et al. (2009), who investigate, among other aspects, the role of the cultural system as reflected in French and Galician term variants. Kerremans (2010, 2016) also studies term variation in specialized translation, focusing on the reflection of the English source language variants in the target languages (Dutch and French). Fernández-Silva and Kerremans (2011) explore cognitive term variants, and affirm that source language variants in Galician are reflected in the English target texts. Miyata and Kageura (2016) argue that translated texts (from Japanese into English) show more term variants due to the different translation possibilities. This is also confirmed by Sanz-Vicente (2011), who focuses on the translation of English MWTs into Spanish. Jiménez-Crespo and Tercedor-Sánchez (2017) investigate term variation in translated (EN>ES) and non-translated texts (Spanish), with a focus on register, determinologization, explicitation, and term variation in translated documents.

Even though term variation is frequent in translation, terminographic resources do not usually describe the different possibilities and the criteria guiding their selection (Kerremans 2010). This is why translators often resort to "unstructured resources": i.e. texts originally written in the source and target languages or previously translated texts (Kerremans 2017). Users need to know the different variants as well as their conceptual and communicative implications, since this will affect the receiver's interpretation of the message. Frequency alone cannot be the sole criterion of classification, since other motivations can be involved in term selection. Therefore, in addition to including the different equivalence possibilities, those data should be enhanced with structural, semantic, pragmatic, and usage information in terminographic resources (Faber and León-Araúz 2016; Giacomini 2018). This would improve a sound use of variation in texts.

3. Materials and Methods

3.1 Extraction of MWT variants from different resources

Since one of our goals was to explore MWT variation patterns in translation situations, the OPUS2 English

corpus (Tiedemann 2012) was used to extract a set of MWTs, which would then be compared with their equivalents in the OPUS2 Spanish parallel corpus (Tiedemann 2012). The OPUS2 English corpus is an open source parallel corpus that can be accessed at Sketch Engine (Kilgarriff et al. 2014) and has 1,139,515,048 words. Aware of the scarcity of specialized parallel corpora, we selected this parallel corpus, despite the fact that it includes both general and specialized texts. It was thus decided to focus on specialized terms of general interest. This is the case of POLLUTION, a specialized concept that is present in everyday communication because of increasing climate awareness.

Starting from the term *pollution*, a conceptual analysis was manually carried out in the corpus, which allowed us to identify pollution-related concepts, such as *emission*, *ozone*, *gas*, *pollutant*, *substance*, *contamination*, *smog*, *air*, etc. These terms were then used as MWT heads in CQL (Corpus Query Language) queries in Sketch Engine to search for specific morphosyntactic patterns, such as MWTs premodified by different elements (Table 1):

[tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]{1,}[lemma="pollution"][tag!="N.*|JJ.*"]

Table 1: CQL query to extract MWTs whose head is pollution.

The CQL expression in Table 1 elicits MWTs such as *indoor air pollution*. It searches for the lemma *pollution* ([lemma="pollution"]) preceded by nouns, adjectives, adverbs, past participles, or present participles ([tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]) appearing one or more times ({1,}). On the right of the head *pollution*, a restriction is included in order to exclude nouns or adjectives ([tag!="N.*|JJ.*"]), which could indicate that *pollution* is not the MWT head but part of a longer MWT. This query was repeated for the other possible heads (*emission*, *ozone*, *gas*, etc.), and the 234 most frequent MWTs were selected. Several of these MWTs were term variants, and were thus grouped by concept. In the end, we obtained a set of 98 pollution-related concepts.

To identify the Spanish term variants, different resources were used. Firstly, with a view to investigating MWT variants in translation contexts, we queried three parallel corpora: (i) the OPUS2 English-Spanish corpus (Tiedemann 2012); (ii) the EurLex English-Spanish corpus (635,187,126 words; Vaisa et al. 2016); (iii) Linguee. Even though Linguee does not allow specific CQL queries and shows just a summary of the possible translations, it complemented the alignment mismatches that were often found in OPUS2 and EurLex.

Secondly, in order to compare term variants found in translation contexts with those present in bilingual or multilingual lexicographic scenarios, Spanish equivalents of English MWTs were also looked up in two terminological databases: TERMIUM Plus and IATE. The entries consulted in these resources also allowed us to expand the collection of English source terms, since many of their entries contain synonyms. Therefore, a new set of terms was researched in the parallel corpora in order to expand the collection of Spanish translation variants. The final set of terms (a total of 277) ranged from two-word terms (e.g. *oil pollution*) to six-word terms (e.g. *aggregate anthropogenic carbon dioxide equivalent emissions*).

Finally, with a view to analyzing term variants in a context of original production in Spanish, we used a Spanish comparable corpus of specialized texts on the environment. It was compiled by the LexiCon research group of the University of Granada while building EcoLexicon and consists of approximately 10 million words. Since the size of the corpus cannot compete with the size of the parallel ones, this was compensated by the use of Google Scholar as a second comparable corpus.

Strictly speaking, Google Scholar is not a comparable corpus nor does it allow for flexible searches, such as lemmatized or CQL queries. However, it was useful to obtain more results and measure the frequency of all variants found in the previous resources. We decided to only retain those terms that occurred a minimum of 10 times. Restricting the queries to Google Scholar ensured a corpus of specialized language since Google Scholar is limited to research work.

The sequence of resources presented (i.e. parallel corpora, terminographic resources, and comparable corpora) was not random. We started with resources that provided direct access to interlinguistic variants (i.e. parallel corpora and terminographic databases), and then the last step was querying the comparable corpora, which required specific strategies to find equivalences since the searching process was more complex.

Our equivalence identification strategy in the comparable corpora involved the following queries. The terms found in parallel corpora (e.g. *contaminante atmosférico*, *contaminante aéreo* [*atmospheric pollutant*]) were literally searched for to confirm their presence in the comparable corpora. However, some of the variants obtained from parallel corpora were not queried in the comparable corpora, since they could bias the results. For examples, this was the case for hypernyms used as term variants, such as *contaminación acústica* [*acoustic pollution*] and *ruido* [*noise*], and *ad hoc* variants (*daños medioambientales* [*environmental damages*] as a term variant of *contaminación medioambiental* [*environmental pollution*]), which do not point to exactly the same concept.

Additionally, the MWT heads and modifiers found in the parallel corpora were used with a 5-element span in between (Table 2), in order to allow for different possibilities, though without being too broad, since concepts in an MWT usually do not have a wider distance, as found in previous studies. Exceptionally, in larger MWTs, such as those including participles or relative sentences, a 10-element span was used. Since CQL queries are not possible in Google Scholar, the * wildcard was employed to indicate the span. This allowed us to obtain new possibilities of extended MWTs, such as *contaminantes liberados a la atmósfera* or *contaminantes vertidos a la atmósfera* [*pollutants released into the air*].

[lemma="contaminante"][]{0,5}[lemma="aire|atmósfera|atmosférico|aéreo"]

Table 2: CQL query for eliciting MWTs of the type *contaminante * aire/atmósfera/atmosférico/aéreo*.

Other queries in the specialized corpus of EcoLexicon included searching for the head and sorting the results by the right context so as to easily distinguish the different modifiers. The same strategy was applied to modifiers, sorting by the left context in that case to discover the different possible heads. Since this type of queries is more time-consuming, they were only used in the EcoLexicon corpus, given the corpus size and the restriction possibilities, but not in Google Scholar. Figure 1 shows some of the modifiers that follow *sustancia* [*substance*] in the EcoLexicon Spanish corpus: *sustancia contaminante* [*polluting substance*], *sustancia explosiva* [*explosive substance*], *sustancia nociva* [*harmful substance*], *sustancia peligrosa* [*dangerous substance*], *sustancia que agota el ozono* [*ozone-depleting substance*], *sustancia que agota la capa de ozono* [*ozone-depleting substance*], its abbreviation *SAO* [*ODS*], *sustancia química* [*chemical substance*], and *sustancia tóxica* [*toxic substance*].

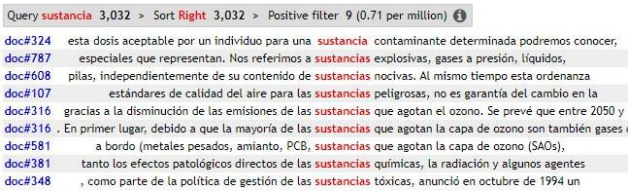


Figure 1: Sample of the modifiers of *sustancia* in the query sustancia + sort by right.

Concordance lines can also reveal knowledge-rich contexts that provide new variants, such as *SAO* (*sustancia que agota la capa de ozono*) [*ODS, ozone-depleting substance*]. Consequently, other variants that had not been elicited in parallel corpora or terminographic resources could be identified. After consulting all different resources, 1428 Spanish term variants were finally retrieved.

3.2 A point system for MWT variants

After extracting the terms in English and their variants in Spanish, an *ad hoc* point system was developed to decide which variants to include –and in which order– in a terminographic resource. Furthermore, this system allowed us to decide which term should act as the main entry term (often arbitrarily based on absolute frequency), since variation requires an anchor point for the establishment of comparisons. The system is based on criteria considered valid by the translation community as well as on our own observations while comparing the data. In consonance with these criteria, points were awarded to each variant:

1. The higher the number of resources where a certain variant was found, the more established and usable in more contexts a variant was. If a variant was found in only one resource, it was awarded 0 points. If it was found in two, then it was awarded 1 point, etc.
2. If a variant was found in a terminographic resource (Termium Plus or IATE), it was more established (2 points) than if it was only found in parallel or comparable corpora.
3. If a variant was found in comparable corpora (EcoLexicon or Google Scholar), it was more established (1 point) than if it was only found in parallel corpora.
4. The higher the frequency of a variant was in each resource compared to the other variants, the more established it was. The frequency of a variant in each resource was calculated in comparison with the total number of appearances of all variants, and that number was then multiplied by 10 to give it the appropriate weight as compared to the other points awarded to the variant.
5. Finally, the points of 1, 2, 3 and 4 were added to give the final score.

For example, the Spanish variant for *anthropogenic pollution*, i.e. *contaminación de origen humano* [lit. pollution of human origin], was found in six resources, and was therefore awarded 5 points. Since it appears in Termium Plus and in both comparable corpora, it was awarded 3 additional points. It appeared three times in OPUS2 out of a total of 39 appearances of all variants, and so came to 3/3.9 (including the multiplication by 10), which equals 0.679. The same procedure was applied to the other corpora and all results were added up. The final score for *contaminación de origen humano* was 13.97. Of course, this final score was only comparable to the final scores of the other Spanish variants of *anthropogenic pollution* and could not be used to compare it with variants of other concepts. Although this system is arbitrary in the sense that there is no objective ground for awarding 1 point for appearing in the EcoLexicon corpus and not 1.5 or 3 points, all of the criteria mentioned have been taken into account and are considered to have a weight according to their importance.

Regarding the final score of all Spanish variants of *anthropogenic pollution*, *contaminación de origen humano* clearly obtained the highest score and was therefore chosen as the main entry term. The rest of the variants are represented and described in relation to this main term. In order to decide on the lowest scores that should be included in the TKB, it is necessary to look at how these scores compare to the highest scores among the variants of the same term. For example, in the case of *anthropogenic pollution*, the highest score is 13.97. This means that variants with scores of around 3 should probably be included. On the contrary, in the case of *greenhouse gas*, the highest score (of the Spanish variant *gas de efecto invernadero*) is 33.86. A variant with a score of 3 would thus not be considered for inclusion in the TKB. Of course, where a terminographer decides to draw the line depends on the aims and intended audience of each terminographic resource.

Accordingly, in the case of *anthropogenic pollution*, *contaminación antropogénica* (13.49), *contaminación antrópica* (9.52), *contaminación humana* (8.83), *contaminación antropógena* (7.84), *contaminación artificial* (5.73), *contaminación de origen antropogénico* (4.29), *contaminación de origen antrópico* (3.85), *contaminación provocada por el hombre* (3.73), *polución*

humana (3.08) should definitely be included. In comparison, variants such as *contaminación debida al hombre*, *contaminación por factores antropogénicos*, *contaminación por factores antrópicos*, *contaminación procedente de actividades humanas*, *contaminación provocada por actividades humanas*, and *contaminación causada por actividades del hombre* only obtained 0.25 points and could thus be discarded for representation, as they are most probably *ad hoc* writing or translation options.

4. Describing MWT Variation in the Environmental Domain

The concepts analyzed show a high degree of variation in Spanish, ranging from 2 term variants (e.g. *sectorial emission*) to more than 46 (e.g. *aircraft noise*). Few of them appear to be highly lexicalized, but among those who are, having an acronym indicates stability (e.g. *compuesto orgánico volátil* (*COV*) [*volatile organic compound*, *VOC*]). The codification of a causal relation, though, was found to make MWTs more prone to variation, since they usually present multiple periphrastic structures making the semantics of the concept explicit. For example, *anthropogenic emissions* can be rendered as *emisiones antropogénicas*, but also as *emisiones procedentes de fuentes humanas* or *emisiones provocadas por el hombre*, among other variants.

It is also worth noting that the more specific the concepts are, the more stable their designations were, even if their hypernyms show many more variants. For instance, *contaminación atmosférica transfronteriza* [*transboundary air pollution*] can be rephrased as *contaminación transfronteriza del aire*, but its hyponym *contaminación atmosférica transfronteriza a larga distancia* [*long-range transboundary air pollution*] is not rephrased as *contaminación transfronteriza del aire a larga distancia* or *contaminación transfronteriza a larga distancia del aire*. Consequently, depending on the bracketing structure (i.e. internal dependencies in the MWT), the distance between the different elements of the MWT seems to present a limited span.

Although many of the variants described in Faber and León-Araúz (2016; Section 2) occurred, most of the variants found in this study were morphosyntactic and cognitive, which calls for an expansion of these categories when it comes to analyzing translation-related variation. The variation found in the different resources has been parametrized in a typology (León-Araúz and Cabezas-García in press) specifically conceived to characterize translation correspondences. Although this classification may also be applied to monolingual term variants, some of its types are especially encountered in translated MWTs. It is divided in three main groups: omissions, changes, and inaccuracies.

1. Omissions
   a. Omission of articles (*total de **las** emisiones agregadas de GEI*, *total de emisiones agregadas de GEI*)
   b. Omission of formants (modifiers: *contaminación atmosférica **transfronteriza***, *contaminación atmosférica* or head: ***sustancia** tóxica*, *tóxico*)
2. Changes
   a. Change of prepositions (expressing cause, location or affected entity: *contaminación **del** agua*, *contaminación **en** el agua*)
   b. Permutations (of modifiers or head and modifiers: *emisiones **totales** de GEI*, ***total** de emisiones de GEI*)
   c. Change of noun within modifier (by synonym, hypernym, metonym or conceptual dimension: *smog de **verano*** (time), *smog de **Los Ángeles*** (location))
   d. Change of noun within head (by synonym, near-synonym: ***destrucción** de la capa de ozono*, ***empobrecimiento** de la capa de ozono*, metonym or conceptual dimension)
   e. Change of both modifiers and head (*smog fotoquímico*, *niebla tóxica estival*)
   f. Change of modifying adjectives (*contaminación **no localizada***, *contaminación **dispersa***)
   g. Change of adjective by periphrasis (*cambio climático **antropógeno***, *cambio climático **producido por el hombre***)
   h. Change of adjective by "of + noun" (*ozono **superficial***, *ozono **de superficie***)
   i. Change of the number of one of the formants (*contaminación **del agua***, *contaminación **de las aguas***)
   j. Introduction of explanatory elements (*contaminación del ozono*, *contaminación **provocada por** el ozono*)
3. Inaccuracies
   a. Inaccuracies related to the semantics of one of the formants (*contaminación **por fuente no localizada***, *contaminación **que no viene de fuente***)
   b. Inaccuracies related to the semantic relation between the formants (*contaminación **del terreno***, *contaminación **de origen terrestre***)
   c. Inaccuracies related to bracketing (*contaminación del aire **urbano***, *contaminación **urbana** del aire*)
   d. Inaccuracies due to style and redundancy (***contaminación** por **contaminantes** orgánicos*)
   e. Inaccuracies due to ad hoc translations (***desastre ecológico** marino*, *contaminación marina*)

As can be inferred from the classification, there are various structural changes that also convey a difference in meaning (i.e. when omissions affect the head, in permutations between head and modifiers, when changes introduce hypernyms, metonyms or different dimensions in modifiers or heads, etc.). Most cases of cognitive variants are related to changes affecting nouns, whether they are in the modifier or in the head, but especially in the latter. As for term opacity, different structures convey more transparent meanings thanks to explicitation. For instance, the preposition *por* (by) is more specific than *de* (of) for making causal relations explicit (e.g. *contaminación **por** petróleo*, *contaminación **de** petróleo*), since *de* is naturally more ambiguous in Spanish.

The most frequent types of variation found in our study were: (1) the omission of articles; (2) the changes in modifiers (reflecting structural or semantic modifications); and (3) the introduction of periphrastic

structures, often through participles such as *causado*, *producido*, *provocado*, etc. (e.g. *cambio climático* ***producido/provocado*** *por el hombre*), and relative clauses followed by verbs such as *provocar*, *causar*, and *producir* (e.g. *emisión de gases* ***que provocan*** *el efecto invernadero*).

Quite often, several of these types coincide within the same set of variants conveying the same concept, as in *contaminación transfronteriza a gran distancia*, where only structural changes and synonyms apply (e.g. *contaminación atmosférica transfronteriza* ***a larga*** *distancia*, *contaminación atmosférica transfronteriza* ***de*** *largo* ***alcance***).

Among the variants for *smog fotoquímico* cognitive changes stand out. For instance, different dimensions are highlighted in the modifiers, whether they are adjectives or nouns: (1) some of them point to the time when this type of pollution usually occurs (*bruma de verano*, *contaminación de verano* [*summer smog*]); (2) others show the city where it was first described (*smog de Los Ángeles* [*Los Angeles smog*]); (3) they can also introduce the agent producing this pollution (*contaminación por ozono*, *esmog de ozono* [*ozone smog*]); or (4) even the process that causes it, the chemical reaction of ozone and light (*bruma fotoquímica*, *contaminación fotoquímica*).

Heads also show cognitive variation, which can entail changes in the general conceptual category, as in *esmog de ozono* and *ozono fotoquímico* or less complex and less accurate conceptualizations, such as *bruma de verano*, *niebla sucia*, and *polución de verano*. Additional structural aspects were also observed in the variants for *smog fotoquímico*, such as the use of the adapted borrowing *esmog* and its non-adapted variant *smog*. SMOG FOTOQUÍMICO is thus a clear example of the richness of term variation.

Therefore, the representation of term variation in terminological resources should be adapted to the different types and consequences of variants.

5. Representing MWT Variation in EcoLexicon

Traditionally, and based on the TBX standard, TKBs are organized in three levels: (1) entry level (for information related to the whole concept), (2) language level (for information pertaining to each language, no matter which term variant is described), and (3) term level (for information differentiating each term). When representing term variation in a TKB, terminographers need to decide at which level they will record each type of information. The previous classifications of term variation (Section 2) are not specifically conceived for the design of a TKB, because the patterns observed refer to both the description of a single term (e.g. borrowings, scientific name, etc.) or the description by comparison to a particular form (e.g. reductions, lexical changes, etc.). Therefore, from the types, causes and consequences of variation analyzed in this paper, we derived a set of descriptive fields that should be included as part of the description of individual terms (i.e. term level) and a set of criteria according to which all term variants of a concept could be grouped and compared (thus, at the language level).

5.1 Variation fields in individual term entries

So far term entries in EcoLexicon contain the following fields: language, term type (main entry term, variant, acronym), part-of-speech (noun, verb, adjective), gender (feminine, masculine, neuter), and note. However, for an accurate representation of term variation, other values and fields need to be added. Table 3 shows the structure of a new term entry proposal for the TKB, including data categories and their values (their type and possible options, whether mandatory or optional, and whether they admit single or multiple values).

Data category	Values
Term type	Picklist (*main term*, *variant*); single value, mandatory
Formation device	Picklist (*borrowing*, *adapted borrowing*, *calque*, *blending*, *acronym*, *abbreviation*, *formula*, *symbol*, *eponym*, *culture-specific*); multiple values, optional
Source	Free text (e.g. UN, corpus EurLex); multiple values; optional
Use_geographical	Free text (e.g. Spain, Mexico, Australia, etc.); multiple; optional
Use_status	Picklist (*admitted*, *deprecated*, *standardized*, *non-recommended*); single value, optional
Use_register	Picklist (*scientific name*, *jargon*, *formal specialized*, *formal semi-specialized*, *informal*); single value; optional
Use_context	Free text; multiple values; optional
Use_translation context	Free text; multiple values; optional
Notes	Free text; multiple values; optional

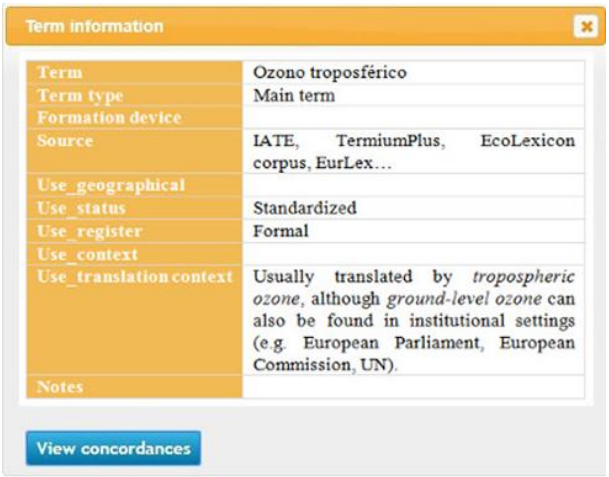
Table 3: Data fields at the term level.

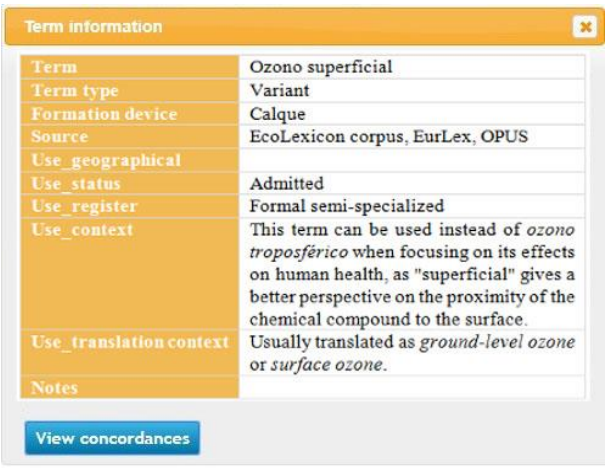
The information pertaining to some of these fields (i.e. Source, Use_geographical, Use_status_context and Use_translation context) can be extracted from the analysis of both comparable and parallel corpora, based on their metadata or on the analysis of a term's local context. For instance, the Use_context is used to include any information about the nuances that a particular variant may have in comparable corpora, whereas the Use_translation context is filled when clear patterns are found regarding the asymmetries of equivalence in parallel corpora.

For example, although *ozono troposférico* is clearly the most frequent Spanish variant designating TROPOSPHERIC OZONE, in the comparable corpora, *ozono a nivel del suelo* and *ozono superficial* seem to be preferred when the term is related to human health issues. In turn, in the parallel corpora, we found that while *ozono troposférico* was usually translated by *tropospheric ozone* or *ground-level ozone*, *ozono a nivel del suelo* and *ozono superficial* clearly preferred *ground-level ozone*, even though it was exactly the same concept. Consequently, the field Use_translation context allows us to establish

interlinguistic variation preferences whereas the field Use_context serves the same purpose for intralinguistic variation.

Figures 2-4 are examples of how the new fields would describe at the term level the terms *ozono troposférico*, *ozono a nivel del suelo* and *ozono superficial*, all variants of *tropospheric ozone*, also known as *ground-level ozone*, *surface ozone*, and *low level ozone*.





Figures 2-4: term entries for *ozono troposférico*, *ozono a nivel de suelo* and *ozono superficial*.

5.2 Variation groupings in contrastive views

When confronted with a vast collection of term variants, translators should perform a contrastive analysis based on parameters, such as frequency, meaning, form and usage trends over time. This kind of information can only be obtained by comparing all of the different choices and not researching individual terms. Therefore, in a TKB this information cannot be represented at the term level. In contrast with the fields in Section 5.1, these parameters were used to enhance the representation of term variants by grouping them together and highlighting their differences.

In EcoLexicon, this will give rise to a new module at the language level divided into five tabs. The first tab contains a summary of the comparison, always with regards to the main term. Figure 5 summarizes the variants of *contaminación de origen humano* (*anthropogenic pollution*). Figures 6-9 show the representation of each type: frequency-ranked variants, formal variants, cognitive variants, and diachronic variants. All examples are illustrated with different groups of term variants.

Figure 6 ranks the selected term variants of *gas de efecto invernadero* (*greenhouse gas*), according to their score (Section 3.2). Figure 7 classifies term variants of *emisión de gases de efecto invernadero* (*greenhouse gas emission*), based on the type of formal changes as compared to the main term, highlighting their differences. In this case, only morphological and morphosyntactic changes, reductions and extensions apply, but other classifying parameters could also be used, such as lexical or graphical changes.

Figure 8 shows the term variants of *smog fotoquímico* (*photochemical smog*), in regard to the conceptual categories and semantic relations codified in the MWT. Finally, Figure 9 depicts the term variants of *agotamiento del ozono* (*ozone depletion*) in a diachronic graph drawn from the Google N-gram viewer.

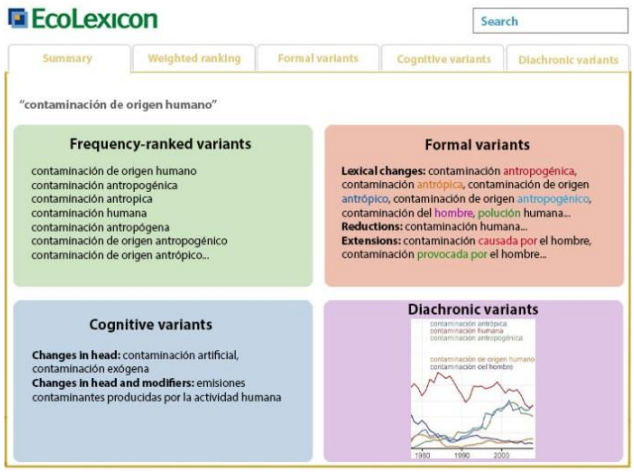


Figure 5: Summary view for *contaminación de origen humano*.

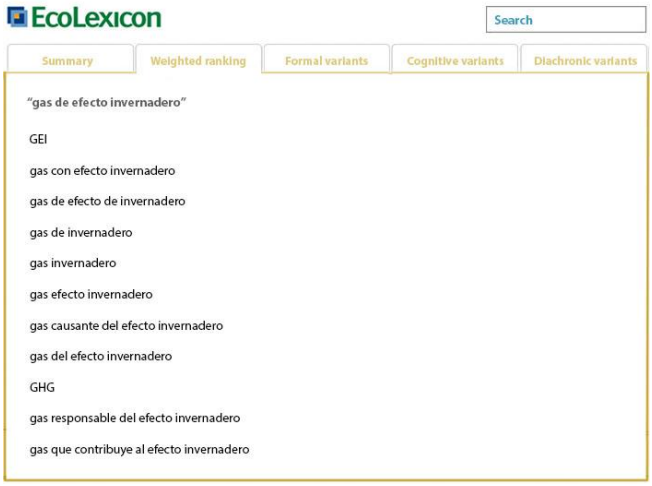


Figure 6: Frequency-ranked *greenhouse effect* Spanish variants.

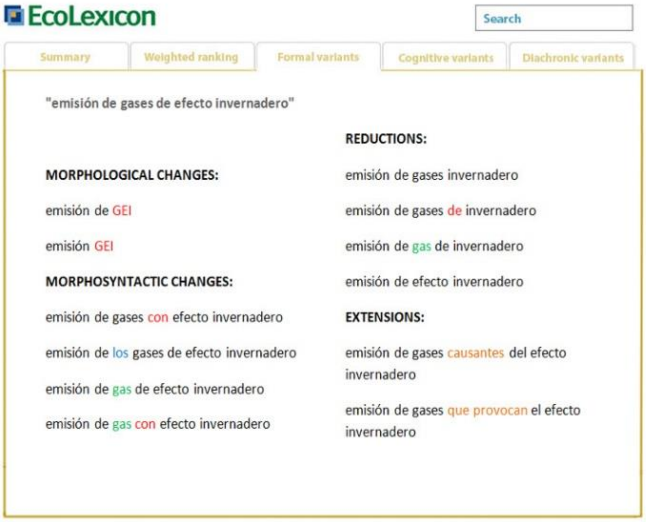


Figure 7: Formal *greenhouse gas emission* Spanish variants.

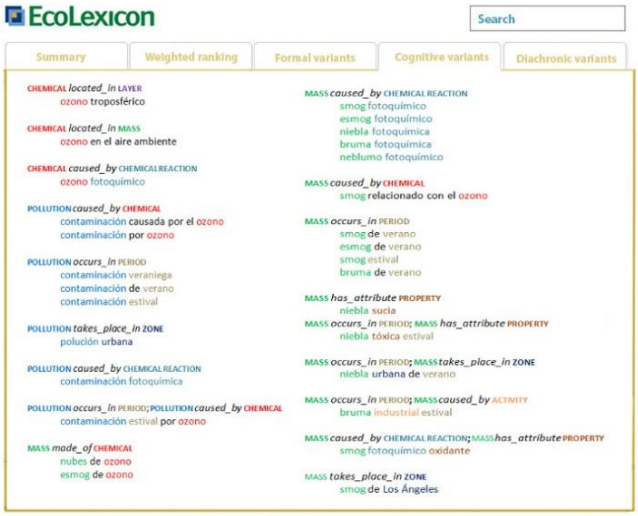


Figure 8: Cognitive *photochemical smog* Spanish variants.

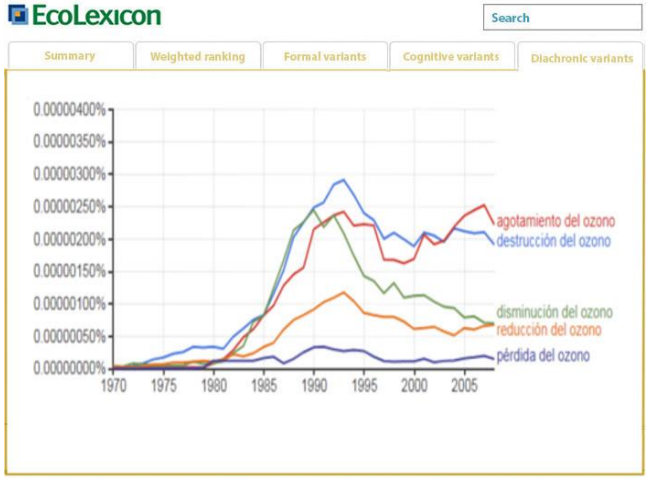


Figure 9: Diachronic view for *ozone depletion* Spanish variants.

6. Conclusions

Although TKBs are primarily conceived as an aid for translation, they rarely contain enough information to help translators make sound choices. In this paper we analyzed and evaluated Spanish translation variants of a set of environmental concepts across different linguistic resources (i.e. parallel and comparable corpora, and terminographic resources). The observations made regarding the types, causes and consequences of term variation led us to design a new representation for it in EcoLexicon. The proposal involves expanding the data in term entries as well as creating a new module where different sets of variants are compared. The comparisons are based on frequency, formal, cognitive and diachronic changes, whereas term entries are enhanced with user- and usage-based information, covering different parameters related to term preference in monolingual and bilingual contexts.

We also developed a system to measure the weight of each variant in order to firstly select the variants (and their order) of representation in the TKB, and secondly, to choose the main term entry. The first aim would establish a delicate balance between thoroughness and over-information because entries with 46 different variants (as found for *noise pollution*) could be counter-productive for users. The second aim would designate the term to be used as the standard of comparison in the contrastive views. Future plans include the study of the causes of diachronic variation (i.e. whether these variants arise because of changes in the conceptualization of specialized entities or whether they stem from lexical preferences influenced by translation or monolingual trends).

7. Acknowledgements

This research was carried out as part of project FF2017–52740-P, funded by the Spanish Ministry of Economy.

8. Bibliographical References

Aguado de Cea, G. and Montiel-Ponsoda, E. (2012). "Term variants in ontologies." In: *XXX Congreso*

Internacional de AESLA 2012. Lleida: AESLA, pp. 436–443.
Bowker, L. (1998). Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study. *Meta* 43(4): 631–651.
Cabré, M. T. (1993). *La terminología. Teoría, metodología, aplicaciones.* Barcelona: Antártida, Empúries.
Daille, B. (2017). *Term Variation in Specialised Corpora: Characterisation, Automatic Discovery and Applications*. Amsterdam: John Benjamins.
Daille, B. and Morin, E. (2005). French-English Terminology Extraction from Comparable Corpora. In: Dale, R. et al. (Eds.) *IJCNLP 2005*. Lecture Notes in Computer Science 3651. Berlin, Heidelberg: Springer, pp. 707–718.
Dejean, H. and Gaussier, É. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica* 1–22.
Faber, P. and León-Araúz, P. (2016). Specialized knowledge representation and the parameterization of context. *Frontiers in Psychology* 7(196):1–20.
Faber, P., León Araúz, P. & Reimerink, A. (2014). Representing environmental knowledge in EcoLexicon. In *Languages for Specific Purposes in the Digital Era*. Educational Linguistics, 19:267-301. Springer.
Fernández-Silva, S. (2018). The cognitive and communicative functions of term variation in research articles: a comparative study in Psychology and Geology. *Applied Linguistics* 1–23.
Fernández-Silva, S., Freixa, J. and Cabré, M. T. (2009). The multiple motivation in the denomination of concepts. *Journal of Terminology Science and Research* 20:1–24.
Fernández-Silva, S. and Kerremans, K. (2011). Terminological variation in source texts and translations: a pilot study. *Meta* 56(2):318–335.
Freixa, J. (2006). Causes of Denominative Variation in Terminology: A typology proposal. *Terminology* 12(1):51–77.
Giacomini, L. (2018). Frame-based Lexicography: Presenting Multiword Terms in a Technical E-dictionary. In: *Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts*. Ljubljana: EURALEX, pp. 309–318.
Gledhill, C. and Pecman, M. (2018). On Alternating Pre-Modified and Post-Modified Nominals Such As Aspirin Synthesis Versus Synthesis of Aspirin: Rhetorical and Cognitive Packing in English Science Writing. *Fachsprache* 40(1-2):26–48.
Gregory, M. and Carroll, S. (1978). *Language and Situation: Language Varieties and their Social Contexts*. London: Routledge.
Jiménez-Crespo, M. A. and Tercedor-Sánchez, M. (2017). Lexical variation, register and explicitation in medical translation: A comparable corpus study of medical terminology in US websites translated into Spanish. *Translation and Interpreting Studies* 12(3): 405–426.
Kerremans, K. (2017). Towards a resource of semantically and contextually structured term variants and their translations. In: P. Drouin et al. (Eds.) *Multiple Perspectives on Terminological Variation*. Amsterdam: John Benjamins, pp. 83–108.
Kerremans, K. (2016). Variation in the translation of terms: corpus-driven terminology research. In: B. Lewandowska-Tomaszczyk et al. (Eds.) *Translation and Meaning*. Frankfurt am Main: Peter Lang, pp. 217–227.
Kerremans, K. (2010). A comparative study of terminological variation in specialised translation. In: C. Heine & J. Engberg (Eds.) *Reconceptualizing LSP. Online proceedings of the XVII European LSP Symposium 2009*. Aarhus: Aarhus University, pp. 1–14.
Kerremans, K. and Temmerman, R. (2016). Finding the (un)expected? Quantitative and qualitative comparisons of term variants and their translations in a parallel corpus of EU texts. In: G. Corpas-Pastor & M. Seghiri (Eds.) *Corpus-based Approaches to Translation and Interpreting: from theory to applications*. Frankfurt am Main: Peter Lang, pp. 43–63.
León-Araúz, P. and Cabezas-García, M. (in press) Term and translation variation of multiword terms. *MonTI*.
Miyata, R. and Kageura, K. (2016). Constructing and Evaluating Controlled Bilingual Terminologies. In: *Proceedings of the Fifth International Workshop on Computational Terminology (Computerm2016)*. Osaka: ACL, pp. 83–93.
Nord, C. (1997). *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St. Jerome Publishing.
Picton, A. (2011). Picturing short-period diachronic phenomena in specialised corpora: A textual terminology description of the dynamics of knowledge in space technologies. *Terminology* 17(1):134–156.
Reiss, K. and Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie*. Tübingen: Niemeyer.
Rogers, M. (1997). Synonymy and equivalence in special-language texts. A case study in German and English texts on Genetic Engineering. In: A. Trosborg (Ed.) *Text Typology and Translation*. Amsterdam, Philadelphia: John Benjamins, pp. 217–245.
San Martín, A., Cabezas-García, M., Buendía, M., Sánchez-Cárdenas, B., León-Araúz, P. & Faber, P. (2017). Recent Advances in EcoLexicon. *Dictionaries: Journal of the Dictionary Society of North America*, 38(1):96-115.
Sanz-Vicente, L. (2011). *Análisis contrastivo de la terminología de la teledetección. La traducción de compuestos sintagmáticos nominales del inglés al español*. Salamanca: Universidad de Salamanca. PhD dissertation.

9. Language Resource References

Google Scholar. Available at: https://scholar.google.es/.
Government of Canada. Termium Plus. Available at: https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography* 1(1):7–36.
LexiCon. Spanish EcoLexicon Corpus. Not yet available for the public.
Linguee. Diccionario inglés-español. Available at: https://www.linguee.es/.
Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In: *Proceedings of the Eighth International*

Conference on Language Resources and Evaluation (LREC'12). Istanbul: ELRA, pp. 2214–2218.

Translation Centre for the Bodies of the European Union. IATE (Interactive Terminology for Europe). Available at: https://iate.europa.eu/home.

Vaisa, V., Michelfeit, J., Medved, M. And Jakubíček, M. (2016). European Union Language Resources in Sketch Engine. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: ELRA, pp. 2799–2803.