EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021

Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

Term variation in terminographic resources: a review and a proposal

Cabezas-García M., León-Araúz P.

University of Granada, Spain

Abstract

Term variation or the coexistence of different terms to name the same concept (e.g. *contamination* and *pollution*) is frequent in specialized language (Fernández-Silva 2018). Since variants are not always interchangeable, language users such as translators or terminologists need to know when and why a variant should be used in preference to another. Terminographic resources should facilitate this task by including different variants as well as the criteria guiding their selection. However, variants are not usually fully covered and, when they are included, indicators regarding semantics, pragmatics, or usage are not often provided. This paper investigates the representation of term variation in terminographic resources. Our goals were (i) to confirm whether term variants are underrepresented and usage indications are not usually provided, (ii) to collect the data categories and fields employed in the description of term variants, and (iii) to propose a model of representation of term variation in the terminological knowledge base EcoLexicon. Our results showed that, despite the prevalence of term variation, terminographic resources do not usually describe the different possibilities and/or the criteria guiding their selection. In contrast, those which attempt to add pragmatic information do not show this kind of data in a parameterized way.

Keywords: term variation; terminology; terminographic resources

1 Introduction

Term variation occurs when different designations are used to name the same concept (e.g. *environmental contamination* and *environmental pollution*). Although to a lesser extent than general language (Freixa 2006; Sanz-Vicente 2011), specialized discourse exhibits a considerable degree of term variation, which has been explored by means of corpora (Fernández-Silva 2018).

Discovering the causes or types of variation may reflect the mental processes involved in the selection of one term instead of another. Furthermore, this information is helpful to terminologists or translators since they need to know when and why one variant should be used in preference to another (Candel-Mora & Carrió-Pastor 2012). Accordingly, terminographic resources should reflect the different variants that may designate each concept as well as their conceptual and communicative implications, since this will affect linguistic productions.

Unfortunately, the different possibilities and the criteria guiding the selection of variants are not usually described (Kerremans 2017). Frequency alone cannot be the sole criterion of classification, since other motivations can be involved in term selection (León-Araúz & Cabezas-García *in press*). Therefore, those data should be enhanced with structural, semantic, pragmatic, and usage information in terminographic resources (Faber & León-Araúz 2016; Giacomini 2018). This would improve a sound use of variation in texts.

This paper explores how term variation is currently represented in terminological resources with a view to (i) confirming whether term variants are underrepresented and usage indications are not usually provided, (ii) collecting the data categories and fields employed in the description of term variants when represented, and (iii) proposing a model of representation that acknowledges the tendencies found and meets the needs in EcoLexicon (a terminological knowledge base on the environment). For this purpose, several multilingual terminological resources, generally aimed at translators (who need to handle variants properly), were examined (e.g. IATE, TERMIUM Plus, MuLex, VariMed), focusing on the presence of variants, their location, and the information provided.

The rest of this paper is organized as follows. Section 2 explains term variation and relevant aspects from the domains of Translation and Terminography. Section 3 analyzes the representation of term variants in terminographic resources. Section 4 presents a proposal for the description of term variation in EcoLexicon, and Section 5 summarizes the conclusions of this study and future research lines.

2 Term Variation in Translation and Terminography

Term variation, or the coexistence of more than one term to name the same concept, can be frequently found in general and specialized language. In both types of discourse, but particularly in the latter, it must be properly understood and treated for the sake of an accurate communication. It thus affects both comprehension and production stages in the translation workflow. On the one hand, different variants in the same text must be understood as pointers to the same concept. On the other hand, the deliberate or random use of variants must be identified and distinguished. Term variation is directly related to the notion of “equivalence”, which can be intra- or interlinguistic (i.e. term variants in the same language, and translations, respectively). While for terminologists, equivalence is present when two or more terms refer to the same concept, other linguists, such as translators, adopt a broader view of equivalence. They consider equivalence at the sentence or text level, rather than at the term level, and thus accept broader mechanisms as equivalent facilitators, such as hypernymy or variants reflecting different conceptualizations. These views often underlie the notion of term variation in terminography, and consequently, its description in terminographic resources.

However, even when variants represent exactly the same concept and thus seem interchangeable, the use of one of them may be preferred depending on different contextual factors. Accordingly, ascertaining the reasons of variation can help us make the correct choice. As Freixa (2006) states, causes for term variation can be (1) dialectal, caused by different origins of the authors (*lift, elevator*); (2) functional, resulting from different communicative registers (*hepatoma, cancer of liver*); (3) discursive, due to different stylistic and expressive needs of the authors (*motor vehicle pollution, car pollution*); (4) interlinguistic, caused by contact between languages (*nursery school, kindergarten*); and (5) cognitive, resulting from different conceptualizations and motivations (*climate change, climate emergency*). Several of these reasons can also co-occur.

Term variation can thus have consequences in communication. When there is only a change in the form but not in the meaning, term variants do not have cognitive effects, as in *marine product* and *sea product*. On the contrary, these consequences occur when term variation implies a shift in perception, as in *sea product* and *fishing product* (Fernández-Silva et al. 2009). Its potential results should be considered when selecting a term variant.

Along these lines, Faber & León-Araúz (2016: 12-13) proposed a classification of term variants that encompasses different proposals and integrates the types, causes and consequences of term variation with a view to describing them in a terminological knowledge base (TKB). This inventory delves into the particularities of the different terms and specifies whether semantics or communicative situation are affected, thus facilitating term selection. The classification includes: (1) orthographic variants (*aesthetics, esthetics*); (2) diatopic variants (*groyne, groin*); (3) short form variants (*laser, Light Amplification by Stimulated Emission of Radiation*); (4) diaphasic variants (*H_2O, water*); (5) dimensional variants (*Gutenberg's discontinuity, core-mantle boundary*); (6) metonymic variants (*water, sea*); (7) diachronic variants (*carbonic anhydride, carbon dioxide*); (8) non-recommended variants (*mental retardation, intellectual disability*); and (9) morphosyntactic variants (*wave action, the action of the waves*). Every category is further specified with additional levels (e.g. diaphasic variants can be scientific, informal, or domain-specific).

Since translators mostly use terminographic resources as their primary tool for finding equivalence, it would be useful for them to find variation-related information, which would accelerate and enhance their translations. However, the different variants that may designate a single concept are not always included in terminographic resources and, when they are, these are not usually enriched with usage information or data are not provided in a systematic fashion (e.g. variants may be lemmatized or not, they may appear in different microstructural positions, etc.) (Giacomini 2018). Different studies have highlighted the need for enhancing the representation of variants by means of semantic, pragmatic, and usage information. One of these studies is Louw's (1998), who argues that, in bilingual dictionaries, the different possible target terms should be accompanied by usage indicators of these variants since contextual constraints may apply. He also emphasizes the role of a coherent marking system of term variants. Precisely this marking system is analyzed in Marí (2017), who states that the tags used in lexicographic resources to account for variation are not often supported by an explanation of how these tags are used. He also highlights the need for systematizing these labels, while acknowledging the complexity of this task.

The importance of enriching the representation of variants for practical users, such as translators or specialized language learners, has also been addressed in Abekawa & Kageura (2008), who complain about the lack of accurate descriptions and exemplifications of variation in electronic terminological English-Japanese dictionaries, despite the large amount of these resources. Since this hinders the translators' tasks, they ask for more usage notes of term variants and develop a tool, which can be accessed from a translation aid system and explores term variants in context using the web. Although these variants are often not very frequent (there is no frequency threshold) and their occurrences are not verified (the system is automatic), as acknowledged by the authors, it is a promising system that may be helpful thanks to the high number of variants provided (mostly based on additions and reductions to the main term).

Along this same line, Alves Costa & Fernández-Silva (2018) explore explicit term variation in Brazilian dictionaries and also complain about its underrepresentation in terminological resources, due to the prescriptive tradition in terminology. They design a proposal for representing term variation in a specialized dictionary for learners, in which they claim that users should be able to: i) understand term variants; ii) distinguish the usage differences of every variant; iii) recognize the possible causes of variation; and iv) verify its cognitive consequences. To this end, the following information is provided: i) term; ii) part of speech; iii) definition; iv) knowledge-rich contexts; v) term variants (followed by an explanation of how variation occurs: relative synonym by inclusion, relative synonym by intersection, no synonym, abbreviation); and vi) explanatory notes. This valuable proposal can guide users when there is no total equivalence.

Giacomini (2018) devises a frame-based model for the representation of multiword terms and their variants in a technical e-dictionary. Aimed at facilitating text production in the native language, her proposal includes multiword terms and their variants, as well as an indication of the type of variation (e.g. partial morphological variation + syntactic variation), the variation pattern followed (e.g. paraphrase + explicitation + transposition), and usage contexts of every variant. This is also a useful representation that does not limit to the mere inclusion of variants.

Accordingly, Kruse & Giacomini (2019) make a proposal for an electronic specialized dictionary, focusing on synonymy relations from an orthographical, morphological and syntactic perspective. Entries will include definitions, abbreviations, equivalents, collocations, and usage examples, as well as information on semantically related terms (e.g. synonyms, antonyms or hyponyms). They claim that, to facilitate text production, terminological resources should include term variants accompanied by information such as the type of text where they are found and its author.

Other studies that have focused on term variation are Janssen's (2006), who explores the representation of orthographic variation, usually presented in the form of cross-references; as well as different works by Freixa (2006), León-Araúz (2015, 2017), Fernández-Silva (2016, 2018), Daille (Daille 2017; Hazem & Daille 2018), and L'Homme (2020), among other authors, on different aspects of term variation.

In conclusion, the representation of the different term variants in terminographic resources becomes central in descriptive settings, contrary to what has traditionally been done. Evidently, users, such as translators, need to know when to use each variant as well as its conceptual and communicative implications, since this will affect the receiver's interpretation of the message. Otherwise, they can actually over-standardize, creating consistency in places where the use of variants was deliberate and well-reasoned (Bowker & Hawkins 2006: 80). Consequently, besides describing different types of variants, which is undoubtedly important, the added value of a linguistic resource lies in the enhancement of those data with additional information, such as semantic, pragmatic, and usage aspects. This thorough representation demands an in-depth analysis as well as a homogenization effort with a view to providing enriched, consistent data.

3 Exploring the Representation of Term Variants in Terminographic Resources

In order to explore how term variation is usually represented in terminographic resources, several publicly available multilingual resources were examined (i.e. IATE, TERMIUM Plus, MuLex, VariMed). Different aspects were analyzed: (i) the presence of term variants and their scope (e.g. what is considered a term variant and how many of them are included in each entry); (ii) their location and prominence in the resource (e.g. whether term variants are described with the same detail as main entry terms); (iii) the information provided for each variant (e.g. type of usage-related constraints included in their description). The analysis was carried out by browsing through term entries where an indexed list was available (i.e. TermiumPlus, MuLex and VariMed) or by searching for a list of terms that were likely to show variants in the domain(s), mostly based on previous knowledge of the concepts (i.e. IATE). The resources analyzed are all aimed at translators, they represent different domains (Law in MuLex and Medicine in VariMed) or multiple domains (i.e. IATE and Termium Plus) are developed by researchers or institutions and are bilingual or multilingual, where English is the only common language. They thus cover different settings in term management and provide an overall view on different ways of dealing with term variation.

3.1 IATE

IATE (https://iate.europa.eu/; Fontenelle & Rummel 2014) is the TKB of the European Union, which includes terms of multiple specialized domains in its official languages. IATE is a concept-oriented resource, that is, entries represent concepts (term variants are thus presumably to be found in the same entry). The entry level in IATE (i.e. the top level, which includes information that applies to the concept and, thus, the whole entry) shows the following data categories: i) concept ID; ii) domain(s); iii) domain note; iv) owner (e.g. Council, European Parliament); v) primary entry (in the case of duplicates); vi) origin (for country-specific concepts); vii) origin note (only in English); viii) life-cycle (historical, proposed or abandoned); ix) cross-references; and x) attachments (documents or graphics). Then, the language level, which holds information relevant to a language, includes the following fields: i) language code; ii) anchor language (according to which all other languages will be attached, usually the source language of the text where the term occurred); iii) definition; iv) definition reference; v) language note; vi) language note references; and vii) attachments. Finally, the term level in IATE (i.e. the basic level, which holds data that is specific to a term) shows the following categories, which are mostly optional and should be used to describe term variants, as will be analyzed below: i) terms; (ii) term type (term, abbreviation, phrase, formula, short form, lookup); iii) evaluation (preferred, admitted, deprecated, obsolete); iv) term reference; v) reliability (not verified, minimum reliability, reliable, very reliable); vi) note; vii) context (i.e. an excerpt); viii) context reference; ix) language usage; x) regional usage; xi) customer (e.g. Translation Centre, European Environment Agency); and ix) grammatical information (part of speech, gender, and number).

After reviewing the structure of this database, we focused on its representation of term variants. To this end, the presence of term variants and their scope was first analyzed. Overall, a "strict" notion of term variation is observed in IATE, that is to say, variants convey the same concept whereas terms conveying slightly different concepts (e.g. hypernyms) are not usually considered term variants, with a few exceptions. Some examples of this vision are *air pollution* and *atmospheric pollution*, which are represented as different denominations of the same concept. However, in some cases a broader vision is adopted. For instance, the term *air pollution episode* is accompanied by its variants *pollution episode* and *episode*, which, strictly speaking, do not convey exactly the same concept although translators could resort to them for stylistic reasons. Nevertheless, it can be concluded that term variants in IATE are most often conceptually equivalent, as confirmed by Kerremans (2015, 2017). As for the number of variants included in IATE entries, we can consider it to be intermediate. In other words, although variants are not extensively covered (as can be the case in TERMIUM Plus, see below), there is a reasonable inclusion of them, with most entries including at least two different denominations in every language.

Regarding the location and prominence of term variants, these appear at the term level (i.e. inside every concept entry, in the language in question) and are described with the same detail as main entry terms. This includes the term-level fields presented above, such as reliability, evaluation, term level note and regional usage, among others. Usage-related constraints can be found in these fields, which must be highlighted since, as mentioned above, this is not the norm. Different types of information are provided: from the evaluation of the term (e.g. preferred, admitted, deprecated) to its geographic details (e.g. UK, internationally), as well as other data, which are very useful in comprehension and production tasks.

However, this information is not provided whenever there are term variants, it is distributed among different fields (despite the fact that a predetermined and thorough set of data categories is available) and it is not always systematically organized. Although this does not prevent users from finding the details, a consistent organization is considered to enhance user experience. For example, some variants (e.g. symbols) are not presented in the *term* field, but in other data categories, such as the *term level note*. This occurs, for instance, with the *O2* symbol. Additionally, certain types of usage-related information do not appear in the corresponding field. This is the case, for example, in the French term *acide carbonique*, which is said to be an inappropriate denomination, as well as in the French term *mongolisme*, which is considered obsolete and thus not recommended. These evaluation-related details, instead of appearing in the *evaluation* field, are presented in the *term level note* category. Accordingly, the Spanish term *anhídrido carbónico* is also considered obsolete and not recommended, which again is not indicated in the *evaluation* field; on the contrary, this time the *regional usage* field is used.

Sometimes, different types of indications are provided in the same data category. For instance, the *language usage* field includes regional information such as the following on *Down's Syndrome*: *Although "Down's Syndrome" is still commonly used in the UK, "Down syndrome" is becoming prevalent internationally and is also found in the UK*. In other cases, this field includes conceptual information that would certainly be more appropriate in this field, such as the following on *ocular ulcer*: *The term ocular ulcer is seldom used when referring to animals, although it occurs in CELEX:32006R1950/EN where it is used for horses. When referring to humans, this term is used to denote a broader concept, encompassing both corneal and eyelid ulcers*. In conclusion, the accurate description of term variation in IATE must be acknowledged. However, even if IATE offers many options to record different types of variants and specify their use, many of the fields are left empty (Kerremans 2015, 2017). It would be ideal if these rich details were provided in as many variants as possible since this would help make better linguistic decisions. Besides, an ordered and homogenous distribution would also be beneficial in such a useful resource.

3.2 TERMIUM Plus

TERMIUM Plus (https://www.btb.termiumplus.gc.ca/) is the TKB of the Government of Canada, which represents millions of concepts from different specialized domains in English, French, and to a lesser extent Spanish and Portuguese. It is one of the largest TKBs in the world. It is also concept-oriented, although the three-level structure is not as clear as in IATE, since some conceptual (i.e. subject field) and term-related information (i.e. usage observations) are stored at the language level.

There is no information at the entry level. At the language level, available categories are what they call textual support: i) definition; (ii) context (where the terms are employed in official sources, especially the main entry term); (iii) observation (where more information is provided regarding the concept or terms); and (iv) phraseologism (for collocations); and v) key terms, which cover spelling or semantic variants and masculine, feminine, singular or plural forms. It is striking that these fields are shown at the language level since, except for the definition, all of them describe the particular use of terms. It is also surprising that key terms are used to include spelling or semantic variants, instead of including them as full-fledged term entries. This is the case of the pollution equipment French entry, which includes four terms as variants (*matériel antipollution*, *matériel de dépollution*, *matériel de lutte contre la pollution* and *équipement antipollution*) and five more as key terms (*matériel anti-pollution*, *équipement de lutte anti-pollution*, *équipement de lutte contre la pollution*, *équipement anti-pollution* and *équipement de lutte antipollution*).

At the term level, available categories are the following: (i) acceptability rating (correct, avoid, unofficial, no rating); (ii) temporal labels (former name, archaic, obsolete); (iii) origin (chemical abstracts service number, classification system code, form code, ISO item number, occupational code, publication code, formula, latin, legal origin, trademark, and proposals, which are equivalents proposed by terminologists, translators or specialists but cannot be found in a written source); iv) linguistic parameters (anglicism, barbarism, calque, deceptive cognate, pleonasm); v) reference (pointing to the field of observation); vi) parts of speech; vii) gender; viii) number; ix) geographic parameters (Africa, Antarctica, Canada, France, Great Britain, NATO, intergovernmental, international, regional); x) frequency (less frequent, rare); xi) sociolinguistic parameters (familiar, jargon); xii) semantic parameters (generic, pejorative, specific); and xiii) official status parameters (standardized, officially approved).

Regarding the presence of variants, Termium Plus covers a wide range of variants per entry. For instance, the entry of *photochemical smog* contains, as opposed to IATE, which does not include any other English variant, three different variants in English (*oxidant smog*, *Los Angeles smog* and *photochemical oxidant smog*) and four in French (*smog photochimique oxydant*, *smog oxydant*, *brouillard photochimique oxydant* and *brouillard photo-oxydant*).

As for their scope, TermiumPlus has a broader vision on term variation, as the field *semantic parameters* (generic, pejorative, specific) implies. For example, many entries include as term variants a hypernym, as those of *environmental pollution* or *water pollution*, which also include *pollution* as a variant (although it is not accompanied by the label *generic*).

In terms of location and prominence, variants appear at the term level and are described in the same detail as main entry terms, with the exception of those included as key terms, as previously mentioned. The preferred term is displayed first followed by variants, including abbreviations.

The description of variants is scattered through different data fields. This is only natural if we take into account the fact that there are different variation parameters and that the description of term variation sometimes happens at the term level and some others by comparison to others. For instance, information related to the geographic origin of a term only needs a geographic label assigned to the term. In contrast, semantic parameters or certain usage constraints depend on the comparison of several variants in different contexts.

Variation parameters at the term level include status (acceptability and official parameters), source (origin), geography (geographic parameters), time (temporal labels), term formation devices (linguistic parameters), frequency, diaphasic parameters (sociolinguistic parameters), and semantic distance (semantic parameters).

However, as was the case in IATE, not all of these fields are filled in each entry. Moreover, there seems to be a certain overlap among these fields. Acceptability is related to the status of terms (correct, avoid, unofficial), but official status parameters seem to point in the same direction (standardized, officially approved). The same happens with geographic parameters and origin. For instance, within geographic parameters Termium Plus includes *intergovernmental* or *international*, which could be considered origin parameters. Likewise, within origin parameters we can find *proposal*, which would actually be related to acceptability or source.

The most interesting field disambiguating the use of term variants is that of observation, at the language level. For instance, in the *pollution* entry, two observations are found: *(1) Pollution is very often used in the more general sense of "environmental pollution". (2) In water pollution, the term contamination is often used as a synonym (...) However, a distinction should be made between "pollution" and "contamination": the latter implies a health danger (particularly to humans).* However, this information is not included in the *water pollution* entry.

Nevertheless, this field is not always included nor it always deals with variation. The problem lies in the fact that the information contained in this field varies in nature and is not systematized. Sometimes it includes conceptual information and some others information related to the terms. For example, in the French entry of *pollution abatement*, two of these types of observations are made: *(1) Le terme français "abatement", dans le domaine de la pollution, est extrêmement douteux: il n'est attesté dans aucun ouvrage spécialisé que nous possédons* (term usage)*; (2) Les nombreux moyens de lutte contre la pollution visent comme objectifs quantitatifs soit la suppression totale des polluants (dépollution), soit leur réduction. D'autres mesures antipollution visent plutôt des objectifs qualitatifs comme l'élimination sélective des pollutants, par exemple, les plus dangereux ou les plus cancérogènes* (conceptual expansion).

When this information concerns the use of one particular term variant, the label *see observation* is included at the term level, but several observations may be included in the entry. The user must then guess which observation was referred to.

We must acknowledge the considerable effort made by this resource to represent term variation in terms of coverage and thoroughness. It is an extremely valuable TKB, but the main drawbacks regarding term variation are related to the lack of information (not all fields are filled) and systematicity (some fields overlap and different types of information are included within a single field, such as that of observation).

3.3 MuLex

MuLex (http://mulex.altervista.org/) is a TKB that describes English and Italian terms related to the legal area of victims of crime. The TKB contains a total of 149 English terms and 197 Italian terms. It was developed in Peruzzo (2012), a study that pays special attention to term variation. MuLex is also concept-oriented, although when non main terms are selected, language-level information is not shown, so it can be viewed as a hybrid approach. Moreover, the searches are based on terms rather than concepts. At the entry level, data categories are: i) subject; ii) subfield; iii) concept field; iv) concept map. At the language level, data categories include i) definition; ii) term variants; iii) notes on term variation; iv) equivalent term; v) note. Finally, the following fields are included at the term level: i) part of speech; ii) usage label (main term, uncommon); iii) category (graphic variant, short form, full form, graphic variant, initials); iv) regional label (UK, EU, CoE); v) style label (official, official EU); vi) lexica (found in IATE); vii) legal system (UK, US); viii) synonymy degree (~, <), only for entries which are not main terms; ix) phraseology (e.g. to claim, to obtain compensation from the offender); x) grammar (e.g. the term was only found in its plural form); xi) context (i.e. an excerpt).

The fact that notes on term variation are at the language level, and only shown when the main term is searched for, implies that these notes are included when different variants are compared. Information characterizing variants individually are included in the fields of usage label, category, regional label, style label, legal system and context. The synonymy degree, when included, is always used with regards to the main term, but not to the rest of the variants. Therefore, the main term has a privileged role in the TKB over the term variants. In contrast, the number of variants included in MuLex is high and their scope is broad. For instance, the entry of *victim* includes variants such as v*ictim of criminal conduct, victim of a crime, victim of the offence* and *crime victim*. They are accompanied by the following note: *When used in national texts, the terms "victim", "victim of a crime", "victim of the offence" and "crime victim" refer to national legislation and can therefore be considered to have a narrower meaning compared to the meaning they have in EU texts. The same holds for "victim of criminal conduct", which is only used in national texts and has a narrower meaning compared to the main term "victim", since the mentioned criminal conduct constitutes an offence specifically regulated by British law*.

These notes are extremely useful, especially considering the diverse nature of legal entities depending on the legal system and/or the institutions producing the texts. They always deal with issues related to conceptual asymmetries, vagueness, and regional differences, which, in this domain, are of paramount importance when disambiguating the use of variants, both in comprehension or production tasks. The problem is that these notes are not found in all entries where several variants coexist. For instance, the entry of *mediation in criminal cases* has nine different variants but no notes on variation are found. In those cases, only usage, regional, style and lexica labels may help to discriminate the use of each variant.

3.4 VariMed

VariMed (http://varimed.ugr.es/) is a TKB developed in the University of Granada and other institutions (Tercedor-Sánchez et al. 2014), which includes terms on the medical domain in English and Spanish, focusing on term variation as a cognitive and communicative phenomenon. VariMed is also concept-oriented but with a particular focus on the description of terms. At the entry level data categories are: i) definition; ii) concept type; iii) related concepts; iv) organs affected; v) images. There is no language-specific information and at the term level data categories are: i) part-of-speech; ii) language; iii) register (formal, informal, jargon, neutral, children); iv) conceptual dimension highlighted (affects, agent/result, visual attribute, non-visual attribute, composed of, time, intensity, discoverer, body part, geography, metaphor/metonymy), v) geographical use (UK, US), vi) familiarity degree (a Likert scale); related variants (abbreviated form of, full form of, often confused with) and vii) other labels (abbreviation, short form, French calque, English calque, eponym, false friend, misspelling, neologism, ICD-10 nomenclature, MeSH nomenclature, greek/latin origin, English borrowing, borrowing from other languages, acronym, culture-specific term, frequent term, non-recommended term, obsolete term, orthographic variant).

Most fields at the term level are aimed at the description of term variants (i.e. register, conceptual dimension, geographical use, familiarity, and other labels). As opposed to Termium Plus, where overlapping fields were found, in VariMed different types and causes of variation converge in an all-purpose field, that of other labels. In this field term formation devices coexist with other parameters such as frequency, time, origin or reliability.

VariMed includes a myriad of variants in both English and Spanish. For instance, type 2 diabetes has up to 7 terms in English (*adult-onset diabetes, diabetes mellitus type 2, NIDDM, non-insulin dependent diabetes mellitus, non-insulin-dependent diabetes mellitus, T2DM, type 2 diabetes*) and 9 terms in Spanish (*diabetes insulinorresistente, diabetes mellitus de inicio adulto, diabetes mellitus de inicio lento, diabetes mellitus estable, diabetes mellitus no insulinodependiente, diabetes mellitus resistente a la cetosis, diabetes mellitus tipo 2, DM-2, DMNID*).

It follows a "strict" vision of variation (i.e. no hypernyms are included), but it shows a wide range of variants in terms of register (i.e. many informal variants and jargon), and morphosyntactic variants are included as full-fledged term entries. They are thus all described with the same level of detail although, as what seems to be a trend, not all fields are always filled. For example, among all variants of type 2 diabetes, *adult-onset diabetes* is categorized as neutral from the register viewpoint, *diabetes mellitus type 2* is categorized as formal and pertaining to the MeSH nomenclature, *NIDDM* is categorized as a formal acronym, *non-insulin dependent diabetes mellitus* as a formal orthographic variant, *non-insulin-dependent diabetes mellitus* as a formal variant pertaining to the MeSH and ICD-10 nomenclatures, *T2DM* as a formal acronym and *type 2 diabetes* as formal. Therefore, more information would still be needed for the disambiguation of all seven variants for text production purposes.

3.5 Summary

Despite the prevalence of term variation, terminographic resources do not usually describe the different possibilities and/or the criteria guiding their selection. In contrast, those which attempt to add pragmatic information do not show this kind of data in a parameterized way. For instance, IATE does not always include this information consistently (i.e. the same kind of data are shown in different fields). TERMIUM Plus shows a rich selection of term variants in many entries. Furthermore, in the field Observations it includes useful information regarding the use of term variants. However, it would benefit from a more systematic approach, since the data provided are not structured around a predetermined set of data categories, they are included at the language level and different types of information (i.e. conceptual and usage-related) are included in the same field. Something similar happens with MuLex, where the most interesting feature disambiguating term variants is placed at the language level but cannot always be found. VariMed features multiple labels to characterize term variants and includes a wide range of them, from specialized nomenclatures to colloquial forms, but they are not always systematically described.

Commonly found data categories are those related to register, reliability, time, origin, geographical use, formation device or status, but the most useful information in terms of usage disambiguation is often found in the form of notes.

4 Representing Term Variation in EcoLexicon: a Proposal

To improve the current description of term variation in EcoLexicon (a comprehensive list of variants but only occasionally described in the form of notes), we designed a preliminary, enhanced model of representation of term variants in this TKB (León-Araúz et al. 2020). This proposal incorporates and improves the tendencies found in other terminographic resources, as well as new approaches, and presents them in a consistent, systematic way. It was tested on multiword terms (e.g. *doubly-fed induction generator*), which are frequent combinations in specialized discourse that are especially prone to variation (Cabezas-García *in press*).

The internal organization of EcoLexicon is essential to understand this model of representation. Based on the TBX standard, EcoLexicon is also structured in three levels: (1) entry level, (2) language level, and (3) term level. When representing term variation in a TKB, terminographers need to decide at which level they will record each type of information. Previous classifications of term variation 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 León-Araúz & Cabezas-García (in press) and the data categories found in the resources analyzed in this paper, a set of descriptive fields was devised. Some of them are included in the description of individual terms (i.e. term level) and some others are a set of criteria according to which all term variants of a concept could be grouped and compared (thus, at the language level).

4.1 Variation Fields at the Term Level: Term Entries

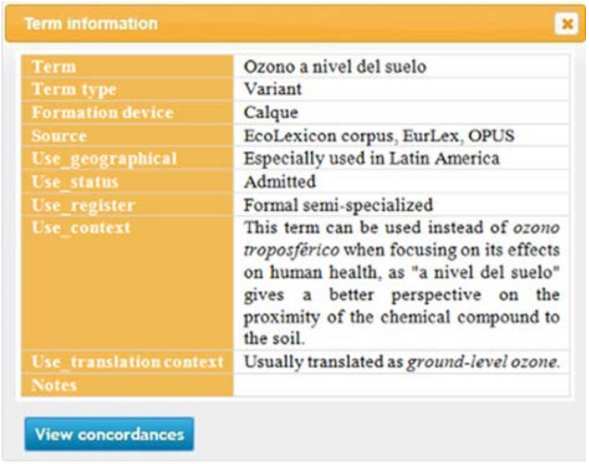
Term entries in EcoLexicon contain the following fields so far: language, term type (main entry term, variant, acronym), part-of-speech, gender, and note. However, for an accurate representation of term variation, other values and fields needed to be added.

Table 1 shows the structure of a new term entry proposal for the TKB, including data categories and their values (their type and possible options, whether they are 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

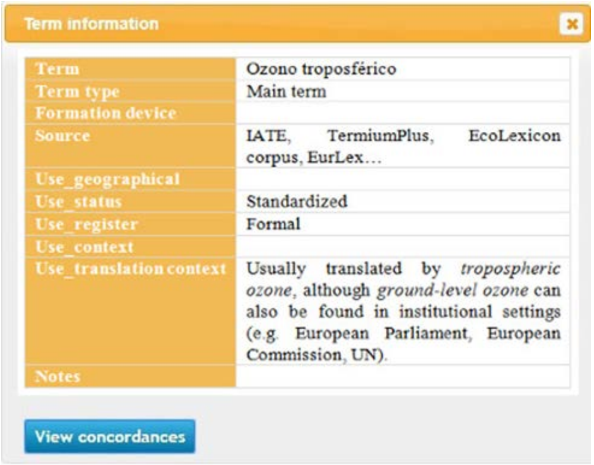
Table 1: Data fields at the term level.

Figures 1 and 2 illustrate how the new fields would describe at the term level the Spanish terms *ozono a nivel del suelo* and *ozono troposférico*, two variants of *tropospheric ozone*, also known as *ground-level ozone*, *surface ozone*, and *low level ozone*.



Term information ☒

Term	Ozono a nivel del suelo
Term type	Variant
Formation device	Calque
Source	EcoLexicon corpus, EurLex, OPUS
Use_geographical	Especially used in Latin America
Use_status	Admitted
Use_register	Formal semi-specialized
Use_context	This term can be used instead of *ozono troposférico* when focusing on its effects on human health, as "a nivel del suelo" gives a better perspective on the proximity of the chemical compound to the soil.
Use_translation context	Usually translated as *ground-level ozone*.
Notes	



Term information ☒

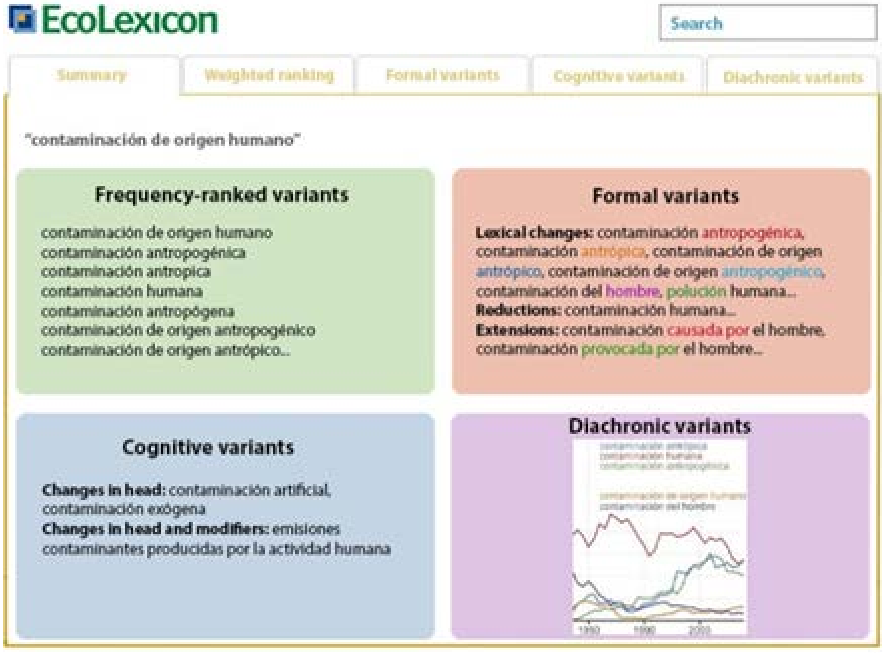
Term	Ozono troposférico
Term type	Main term
Formation device	
Source	IATE, TermiumPlus, EcoLexicon corpus, EurLex…
Use_geographical	
Use_status	Standardized
Use_register	Formal
Use_context	
Use_translation context	Usually translated by *tropospheric ozone*, although *ground-level ozone* can also be found in institutional settings (e.g. European Parliament, European Commission, UN).
Notes	

Figures 1 and 2: Term entries for *ozono a nivel del suelo* and *ozono troposférico*.

The Use_context includes any information about the nuances that a particular variant may have in comparable corpora (i.e. a context of original production rather than translation), whereas the Use_translation context is filled when clear patterns are found regarding the asymmetries of equivalence in parallel corpora (i.e. translations).
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.

4.2 Variation Fields at the Language Level: Contrastive Views

Regarding the criteria for the grouping and comparison of all term variants of a concept, which is essential for proper language production, a new module in EcoLexicon was devised (Figures 3-7). This module enhances the representation of term variants by grouping them together and highlighting their differences based on frequency, meaning, form, and usage trends over time. For this reason, it belongs to the language level, since information is not term-centered. It is divided into five tabs.



EcoLexicon

Search

Summary | Weighted ranking | Formal variants | Cognitive variants | Diachronic variants

"contaminación de origen humano"

Frequency-ranked variants

contaminación de origen humano
contaminación antropogénica
contaminación antropica
contaminación humana
contaminación antropógena
contaminación de origen antropogénico
contaminación de origen antrópico...

Formal variants

Lexical changes: contaminación antropogénica, contaminación antrópica, contaminación de origen antrópico, contaminación de origen antropogénico, contaminación del hombre, polución humana...
Reductions: contaminación humana...
Extensions: contaminación causada por el hombre, contaminación provocada por el hombre...

Cognitive variants

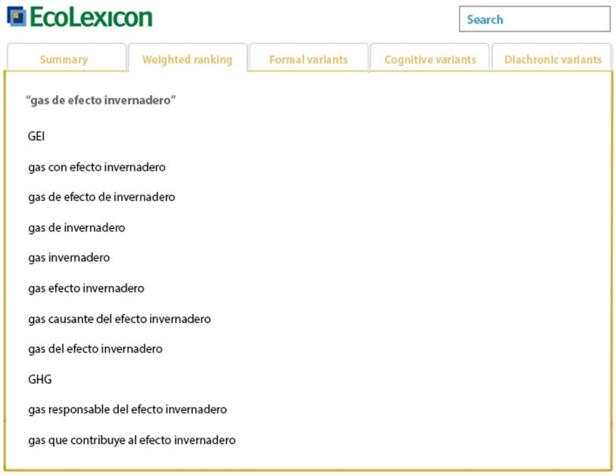
Changes in head: contaminación artificial, contaminación exógena
Changes in head and modifiers: emisiones contaminantes producidas por la actividad humana

Diachronic variants

Figure 3: Summary view for *anthropogenic pollution* Spanish variants.

The first tab contains a summary of the comparison with regards to a previously established main term. Figure 3 summarizes the variants of *contaminación de origen humano* (*anthropogenic pollution*). Figures 4-7 show different types of variant classification: frequency-ranked variants, formal variants, cognitive variants, and diachronic variants. All examples are illustrated with different groups of term variants that best exemplify the approach.

Figure 4 ranks the most representative term variants of *gas de efecto invernadero* (*greenhouse gas*), according to a procedure developed in León-Araúz et al. (2020), which identifies the most established variants of a term. Figure 5 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 6 shows the term variants of *smog fotoquímico* (*photochemical smog*), in regard to the conceptual categories and semantic relations codified in every term. Finally, Figure 7 depicts the term variants of *agotamiento del ozono* (*ozone depletion*) in a diachronic graph drawn from the Google N-gram viewer.



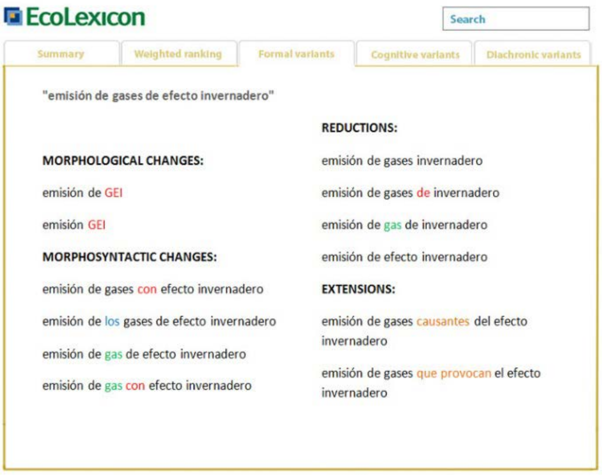
EcoLexicon

Search

Summary | Weighted ranking | Formal variants | Cognitive variants | Diachronic variants

"gas de efecto invernadero"

GEI
gas con efecto invernadero
gas de efecto de invernadero
gas de invernadero
gas invernadero
gas efecto invernadero
gas causante del efecto invernadero
gas del efecto invernadero
GHG
gas responsable del efecto invernadero
gas que contribuye al efecto invernadero



EcoLexicon

Search

Summary | Weighted ranking | Formal variants | Cognitive variants | Diachronic variants

"emisión de gases de efecto invernadero"

MORPHOLOGICAL CHANGES:
emisión de GEI
emisión GEI

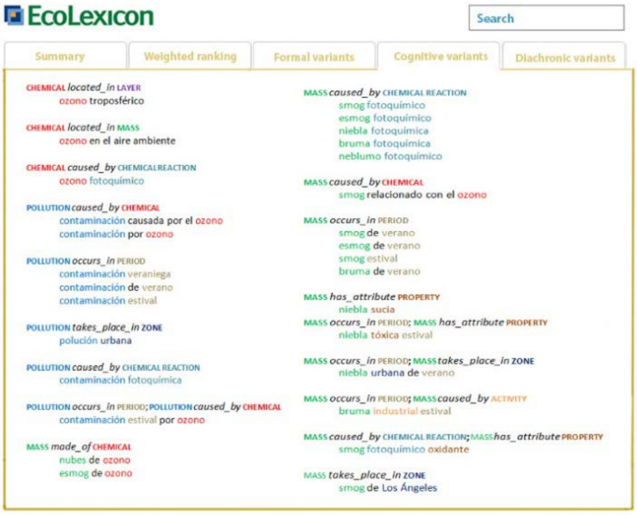
MORPHOSYNTACTIC CHANGES:
emisión de gases con efecto invernadero
emisión de los gases de efecto invernadero
emisión de gas de efecto invernadero
emisión de gas con efecto invernadero

REDUCTIONS:
emisión de gases invernadero
emisión de gases de invernadero
emisión de gas de invernadero
emisión de efecto invernadero

EXTENSIONS:
emisión de gases causantes del efecto invernadero
emisión de gases que provocan el efecto invernadero

Figure 4: Frequency-ranked *greenhouse gas* Spanish variants.

Figure 5: Formal *greenhouse gas emission* Spanish variants.



CHEMICAL *located_in* LAYER
ozono troposférico

CHEMICAL *located_in* MASS
ozono en el aire ambiente

CHEMICAL *caused_by* CHEMICALREACTION
ozono fotoquímico

POLLUTION *caused_by* CHEMICAL
contaminación causada por el ozono
contaminación por ozono

POLLUTION *occurs_in* PERIOD
contaminación veraniega
contaminación de verano
contaminación estival

POLLUTION *takes_place_in* ZONE
polución urbana

POLLUTION *caused_by* CHEMICAL REACTION
contaminación fotoquímica

POLLUTION *occurs_in* PERIOD; POLLUTION *caused_by* CHEMICAL
contaminación estival por ozono

MASS *made_of* CHEMICAL
nubes de ozono
esmog de ozono

MASS *caused_by* CHEMICAL REACTION
smog fotoquímico
esmog fotoquímico
niebla fotoquímica
bruma fotoquímica
neblumo fotoquímico

MASS *caused_by* CHEMICAL
smog relacionado con el ozono

MASS *occurs_in* PERIOD
smog de verano
esmog de verano
smog estival
bruma de verano

MASS *has_attribute* PROPERTY
niebla sucia
MASS *occurs_in* PERIOD; MASS *has_attribute* PROPERTY
niebla tóxica estival

MASS *occurs_in* PERIOD; MASS *takes_place_in* ZONE
niebla urbana de verano

MASS *occurs_in* PERIOD; MASS *caused_by* ACTIVITY
bruma industrial estival

MASS *caused_by* CHEMICAL REACTION; MASS *has_attribute* PROPERTY
smog fotoquímico oxidante

MASS *takes_place_in* ZONE
smog de Los Ángeles

Figure 6: Cognitive *photochemical smog* Spanish variants.

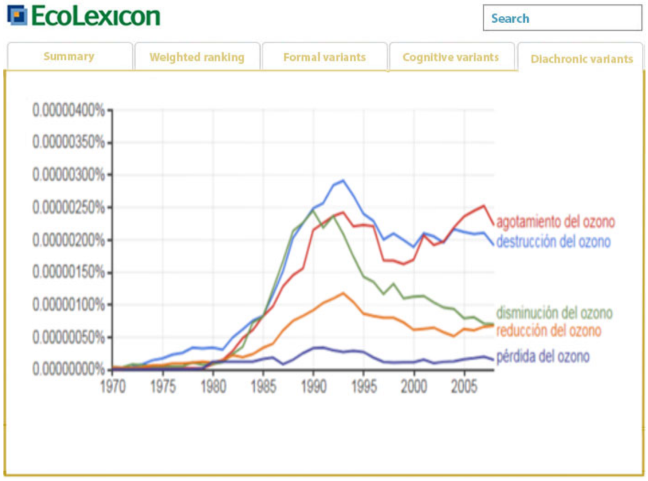


Figure 7: Diachronic view for *ozone depletion* Spanish variants.

5 Conclusions

Properly handling term variation is essential in order to understand and produce quality texts. Terminographic resources play a major role in this respect, as the linguistic resources most used for these tasks. Nevertheless, when resorting to them, users often find, at best, various synonyms with no indication (or unstructured data) on which term should be used in a particular context; or at worst, a lack of coverage of the different forms of naming the same concept. The resources analyzed (i.e. IATE, TERMIUM Plus, MuLex, and VariMed) were not among the worst-case scenarios, since they satisfactorily cover term variants and often provide indications on their usage. However, an in-depth analysis revealed some inconsistencies, which shows that there is still room for improvement regarding the representation of term variation.

Apart from the different views of term variation (some are broad, considering variants terms that do not convey exactly the same concept; other are stricter, acknowledging variation only when the same concept was evoked), a different coverage of term variants was observed (which is in line with the variation scope adopted). Since these resources were mostly concept-oriented, term variants were included in concept entries, and were usually described with the same detail as main entry terms. Nevertheless, it is the information provided for each variant, as well as the data categories chosen, that differs most among these resources. Their inclusion of pragmatic information about variants is helpful, although this does not appear for every single variant contained in the resource. Different fields and data categories are obviously used in every resource. In addition, data are not always systematically presented (sometimes the same information is presented in distinct fields), which could hinder access to information. Finally, a new model of representing term variation was devised in the TKB EcoLexicon (León-Araúz et al. 2020), which provides users with more usage-related, consistent information (the weakness of most terminographic resources). What remains to be done is to fill the new fields for all variants in the TKB, which will undoubtedly be a time-consuming task.

6 References

Abekawa, T., Kageura, K. (2008). QRcep: a term variation and context explorer incorporated in a translation aid system on the web. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress, Euralex 2008, 15-19 July 2008*. Barcelona: Universitat Pompeu Fabra, pp. 915-922.

Alves Costa, L., Fernández-Silva, S. (2018). A variação denominativa explícita na Lexicografia no Brasil: pressupostos para a organização microestrutural do Dicionário de Lexicografia Brasileira. In *Ibérica*, 36, pp. 93-118.

Bowker, L., Hawkins, S. (2006). Variation in the organization of medical terms. Exploring some motivations for term choice. In *Terminology*, 12(1), pp. 79-110.

Cabezas-García, M. (In press). *Los términos compuestos desde la Terminología y la Traducción*. Berlin: Peter Lang.

Candel-Mora, M.A., Carrió-Pastor, M.L. (2012). Corpus analysis: a pragmatic perspective on term variation. In *RESLA (Revista Española de Lingüística Aplicada)*, 2012, pp. 33-50.

Daille, B. (2017). *Term Variation in Specialised Corpora: Characterisation, Automatic Discovery and Applications*. Amsterdam: John Benjamins.

Faber, P., León-Araúz, P. (2016). Specialized knowledge representation and the parameterization of context. In *Frontiers in Psychology*, 7(196), pp. 1-20.

Fernández-Silva, S. (2016). The Cognitive and Rhetorical Role of Term Variation and its Contribution to Knowledge Construction in Research Articles. In *Terminology*, 22(1), pp. 52-79.

Fernández-Silva, S. (2018). The cognitive and communicative functions of term variation in research articles: a comparative study in Psychology and Geology. In *Applied Linguistics*, 2018, pp. 1-23.

Fernández-Silva, S., Freixa, J., Cabré, M.T. (2009). The multiple motivation in the denomination of concepts. In *Journal* *of Terminology Science and Research*, 20, pp. 1-24.

Fontenelle, T., Rummel, D. (2014). Term Banks. In P. Hanks, G.-M. De Schriver (eds.) International Handbook of Lexis and Lexicography. Berlin-Heidelberg: Springer, pp. 1-12.

Freixa, J. (2006). Causes of Denominative Variation in Terminology: A typology proposal. In *Terminology*, 12(1), pp. 51-77.

Giacomini, L. (2018) Frame-based Lexicography: Presenting Multiword Terms in a Technical E-dictionary. In N. Calzolari et al. (eds.) *Proceedings of the XVIII EURALEX International Congress*, *Euralex 2018, 17-21 July 2018*. Ljubljana: EURALEX, pp. 309-318.

Hazem, A., Daille, B. (2018). Word Embedding Approach for Synonym Extraction of Multi-Word Terms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, 7-12 May 2018.* Miyazaki, Japan: ELRA, pp. 297-303.

Janssen, M. (2006). Orthographic variation in lexical databases. In E. Corino, C. Marello, C. Onesti (eds.) *Proceedings of the XII EURALEX International Congress*, *Euralex 2006, 6-9 September 2006*. Turin: Alessandria, Edizioni dell'Orso, pp. 167-172.

Kerremans, K. (2015). Managing terminological and translational diversity in parallel corpora: a case study in institutional translation. In *InTRAlinea: Online Translation Journal,* 2015.

Kerremans, K. (2017). Towards a resource of semantically and contextually structured term variants and their translations. In P. Drouin, A. Francœur, J. Humbley, A. Picton (eds.) Multiple Perspectives on Terminological Variation. Amsterdam: John Benjamins, pp. 83-108.

Kruse, T., Giacomini, L. (2019). Planning a Domain-specific Electronic Dictionary for the Mathematical Field of Graph Theory: Definitional Patterns and Term Variation. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference, 1-3 October 2019, Sintra*. Brno: Lexical Computing CZ, s.r.o., 676-693.

L'Homme, M.C. (2020). *Lexical Semantics for Terminology: An Introduction*. John Benjamins.

León-Araúz, P. (2015). Term variation in the psychiatric domain: transparency and multidimensionality. In P. ten Hacken, R. Panocová (eds.) Word Formation and Transparency in Medical English. Newcastle-upon-Tyne: Cambridge Scholars Publishing, pp. 33-54.

León-Araúz, P. (2017). Term and Concept Variation in Specialized Knowledge Dynamics. In P. Drouin, A. Francœur, J. Humbley, A. Picton (eds.) Multiple Perspectives on Terminological Variation. Amsterdam: John Benjamins, pp. 213-258.

León-Araúz, P., Cabezas-García, M. (In press). Term and translation variation of multi-word terms. In *MonTI: Monografías de Traducción e Interpretación*.

León-Araúz, P., Cabezas-García, M., Reimerink, A. 2020. Representing Multiword Term Variation in a Terminological Knowledge Base: a Corpus-Based Study. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation, LREC 2020, 11-16 May 2020.* Marseille, France: ELRA, pp. 2351-2360.

Louw, P. A. (1998). Synonymy in the Translation Equivalent Paradigms of a Standard Translation Dictionary. In *Lexikos*, 8, pp. 173-182.

Marí, I. (2017). Notes on the treatment of variation in the Diccionari català-valencià-balear (DCVB). In *Dialectologia,* special issue VII, pp. 113-131.

Peruzzo, K. (2012). Terminological Equivalence and Variation in the EU Multi-level Jurisdiction: A Case Study on Victims of Crime. PhD thesis. Università degli Studi di Trieste, Trieste, Italy.

Sanz-Vicente, L. (2011). Análisis contrastivo de la terminología de la teledetección. La traducción de compuestos sintagmáticos nominales del inglés al español. PhD thesis. Universidad de Salamanca, Salamanca, Spain.

Tercedor-Sánchez, M., López-Rodríguez, C.I., Prieto Velasco, J.A. (2014). También los pacientes hacen terminología: retos del proyecto VariMed. In *Panace@: revista de Medicina, Lenguaje y Traducción*, 25(39), pp. 95-103.

Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness.