UNIVERSIDAD DE GRANADA
FACULTAD DE TRADUCCIÓN E INTERPRETACIÓN

TRABAJO FIN DE GRADO

THE TRANSLATION OF COMPLEX NOMINALS IN THE FIELD OF AIR QUALITY TREATMENT

Presentado por:
D.ª Sandra Gutiérrez Bullón

Tutora:
Prof.ª Dra. Pilar León Araúz

Curso académico 2019/2020

Acknowledgements

First and foremost, I would like to thank Professor Pilar León for her guidance, encouragement and enthusiasm throughout the process. Her brilliant mind and her never-ending passion for Terminology have helped shaped this project into what it is now.

I am also grateful to Professor Silvia Montero, for offering me the excellent opportunity to work for her, and take my first steps in the world of Terminography while simultaneously writing this dissertation.

Finally, I would like to extend my words of gratitude to my mother, for her endless love and support, and for always believing in myself no matter what.

To all of you, thank you. You inspire me.

Contents

1. Introduction 1
2. Theoretical framework 3
2.1. Scientific language and specialized translation 3
2.1.1 Multi-word terms 3
2.1.2. Term variation in specialized discourse 7
2.2. Theory and practice of Terminology 12
2.2.1 The cognitive approach of Frame-Based Terminology 13
2.2.2 Terminological knowledge bases: EcoLexicon 14
2.3. Translation and Computational Linguistics 15
2.3.1. Corpus linguistics 15
2.3.2. Machine translation 16
3. Materials and methods 18
3.1. Corpus design and compilation 18
3.2. Extraction of 3- and 4-word CNs within the English corpus 19
3.3 Structural disambiguation of CNs 23
3.4 Corpus-based semantic analysis of CNs 24
3.5 Identification of target language equivalents 25
3.5.1 Extraction of translation variants within terminographic resources 25
3.5.2 Extraction of Spanish variants within the comparable corpus 27
3.5.3 Extraction of Spanish variants within the web and other corpora 28
3.6 Evaluation of machine translation output 29
4. Results and discussion 31
4.1 Analysis of source term variation 31
4.2 Analysis of target term variation 32
4.3 Analysis of machine translation output 34
4.4. Towards a protocol for translating CNs of more than 3 constituents 38
5. Conclusions 46
References 48
Annex

1. Introduction

Complex nominals (CNs), such as *total ozone column*, are an integral part of specialized communication. These multi-word units constitute one of the main term formation mechanisms, as they allow multiple possibilities of conceptual combinations. Whereas CNs have attracted the interest of researchers across various disciplines, such as Natural Language Processing, in the present study we focus on the challenges faced by translators when rendering these units in the English-to-Spanish language pair—from source term decoding to target term production.

The structural ambiguity of CNs, together with the fact that the semantic relation between the constituents cannot be inferred by the head and modifiers (Ó Séaghdha & Copestake, 2013), makes the analysis of these units a difficult task. Aside from their cognitive and structural complexity, the proliferation of different forms further hinders the translation process. As a result, our study delves into not only the decoding process of these units, but also the analysis of term variation in the two languages.

Moreover, with machine translation gaining ground in the translation industry, we set out to assess the performance of these engines in the translation of English CNs into Spanish—a two-fold approach which aims to unveil the ins and outs of the translation of CNs at the present time.

Based on the difficulties these units pose for human and machine translators alike, our research is ultimately oriented towards the much-needed development of a series of guidelines to translate English CNs into Spanish, where the notions of neology and secondary term formation inevitably enter the picture.

Instead of limiting to the study of 2, 3, or 4-word CNs—which have been researched in some depth (Nakov, 2013; Sanz-Vicente, 2012a; Cabezas-García, 2019; Cabezas-García & León-Araúz, 2019; *inter alia*)—the scope of our study is widened to include up to 7-word combinations.

The specialized domain analyzed is air quality treatment, which is explored by means of two manually-compiled comparable corpora (EN, ES) in combination with other online corpora and terminographic resources, as well as the output of some of the main MT engines available in the web. In the past decade, the issue of air quality has been in the international agenda, as a main cause of concern across different countries, especially since the implementation of the 2008/50/EC Directive on Ambient Air Quality and Cleaner Air for Europe[1]. The conversation around climate change is more pressing than ever, which is reflected in the increasing number of scientific journals centered around this issue, such as *Atmospheric Pollution Research* or *Atmospheric Environment*.

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1486474738782&uri=CELEX:02008L0050-20150918

The main goal of this work was, thus, to study the translation of CNs in the field of air quality treatment through the lens of both human and machine translation. To this end, a set of specific objectives were established:

(1) to address the fundamental characteristics of CNs and review preceding postulates on variation in specialized discourse;
(2) to investigate term and translation variation of CNs by means of comparable corpora and other terminographic resources;
(3) to evaluate the machine translation output of CNs of 3 or more constituents in the English>Spanish language pair and
(4) to design a procedure for translating correspondences for CNs of four or more constituents, for which no pre-established Spanish equivalents could be found.

The rest of this work is organized as follows. Section 2 presents the theoretical framework upon which the research is based, focusing on three broad topics: (i) scientific language and specialized translation; (ii) the theory and practice of Terminology and (iii) Translation and Computational Linguistics. More specifically, Section 2.1 addresses multi-word terms and explores the main aspects of term variation in specialized discourse. Section 2.2 describes the main premises of Frame-Based Terminology and presents EcoLexicon, the terminological knowledge base resulting from its application, whereas Section 2.3 focuses on corpus linguistics and the various approaches to machine translation. Section 3 explains the materials used and the methods followed in this research. In section 4, the preliminary results of this research are presented and discussed. Finally, Section 5 sums up the conclusions that can be derived from this study and outlines plans for future research.

2. Theoretical framework

2.1. Scientific language and specialized translation

Scientific language exhibits a series of features which poses significant challenges in specialized translation—both in text understanding and text production. Aside from presenting information in a synthetical way, scientific discourse exhibits a high level of precision and information packing. While this also has consequences in pragmatics and syntax, it primarily affects the lexis of the texts, which feature a high level of terminological density. In the following sections we will focus on two issues central to specialized translation: multi-word terms and term variation.

2.1.1 Multi-word terms

In scientific and technical communication, MWTs are the most frequent type of lexical unit (Meyer & Mackintosh, 1996; Ramisch, 2015). Therefore, specialized translation implies dealing with a vast quantity of MWTs. In the following sections we will focus on CNs, which are one of the main types of MWTs and one of the most frequent term formation mechanisms, due to their multiple possibilities of conceptual combination. (León-Araúz & Cabezas-García, in press).

2.1.1.1 Definition and linguistic properties of complex nominals

MWTs are sequences of two or more elements that designate a specialized concept (Cabezas-García & León-Araúz, 2019). Since these terms usually have a nominal head, they are known as complex nominals, noun compounds or nominal compounds. These have been defined in various ways. In the present study we adopt the definition in Levi (1978), "syntactic construction dominated by a N node and composed (in its simplest form) of a head noun preceded by a modifier which is either another noun or a nominal adjective" (Levi, 1978: 39).

These compounds can be endocentric or exocentric (Levi, 1978; Štekauer, 1998; Nakov, 2013). In an endocentric complex nominal, "one member functions as the head and the other as its modifier, attributing a property to the head" (Nakov, 2013: 299). In contrast, exocentric CNs appear to lack a head (Bauer, 2008), and thus are not subtypes of any of their constituents. The most frequent MWTs in specialized texts are endocentric CN, "which are the specification of a hypernym" (Cabezas-García & León-Araúz, 2019).

Nakov (2013) identified a series of linguistic properties of CNs, which included (1) headedness, (2) transparency (ranging from completely idiomatic CNs such as *honeymoon* to transparent ones such as *water bottle*), (3) syntactic ambiguity, and (4) language dependency (i.e. a combination of words may be regarded as a compound in a language but not in others, due to language-internal reasons).

English and Spanish are distinguished by different linguistic features, partly due to their roots (Germanic for English, Romance for Spanish). For instance, the preferred terminogenesic processes in English do not match those in Spanish. English CNs are most commonly the result of pre-modification formation patterns (Kim & Baldwin, 2013; Levi, 1978; Sager et al., 1980), with noun heads being modified by other nouns or adjectives (e.g. *black carbon aerosol*). Conversely, post-modification is the preferred formation process in Spanish. Whereas adjective and verb compounds are right-headed, Spanish noun compounds are largely left-headed, with noun heads modified by an adjective or a prepositional phrase (e.g. *aerosol de carbono negro*).

Some authors (Zuluaga, 1975; García-Page, 2008) contend that MWTs fall outside the scope of phraseology. Nonetheless, we agree with those who assume them to be phraseological units (PUs) (Benson et al., 1986; Ramisch, 2015; Cabezas-García, 2019; *inter alia*), because they share the defining features of PUs—they are multiword expressions whose constituents co-occur frequently and function as a whole, thereby showing a certain degree of lexicalization. They differ from idioms in that they are more transparent (i.e. less lexicalized and idiomatic), and they convey concepts.

2.1.1.2. Structural disambiguation of MWTs

Two-term combinations have traditionally been the main focus in MWTs research. Longer sequences, despite posing a much bigger translation problem, have received less specific attention. Indeed, from three components onwards, the interpretation of MWTs poses a new challenge: bracketing. In order to tackle the bracketing of three-term MWTs, Natural Language Processing (NLP) has proposed two models: the adjacency model and dependency model.

The adjacency model (Marcus, 1980; Pustejovsky et al., 1993) takes an MWT p1p2p3 and compares if p2 is more related to p1 or p3. For that purpose, the number of occurrences of p1p2 and p2p3 are compared. For instance, in *air quality treatment* there are more occurrences of *air quality* than of *air treatment* in any corpus. Thus, a left-bracketing structure is inferred *[air quality] treatment*.

The dependency model (Lauer, 1995) also takes an MWT p1p2p3 and compares whether p1 is more strongly associated with p2 or p3. Therefore, the analysis does not start from the central term, as in the adjacency model, but rather from the first one to the left. When p1 is more strongly associated with p2 than to p3, there is a left bracketing (*[passive air] sampler*). However, following both models the two possible combinations often show similar frequencies, especially in combinations of more than three terms. By the same token, adjacency and dependency frequencies cannot be the single disambiguating criterion when working with small or unbalanced corpora. With this in mind, Nakov and Hearst (2005: 19-21) point out other indicators that can clarify the dependencies in English MWTs. These include the identification of term variants on the web. If they have the following characteristics (see Figure 1), they point to an internal group.

CNs in need of structural disambiguation	Denominative variant	Bracketing indicators	Bracketing
cell cycle análisis	*cell-cycle análisis*	hyphen	[*cell cycle*] *analysis*
brain stem cell	*brain's stem cell*	possessive marker	*brain* [*stem cell*]
plasmodium vivax malaria	*Plasmodium vivax Malaria*	internal capitalization	[*plasmodium vivax*] *malaria*
leukemia lymphoma cell	*leukemia/lymphoma cell*	embedded slash	*leukemia* [*lymphoma cell*]
growth factor beta	*growth factor (beta)*	brackets	[growth factor] beta
tumor necrosis factor	*tumor necrosis factor (NF)*	abbreviation	*tumor* [*necrosis factor*]
health care reform	*healthcare reform*	concatenation	[*health care*] *reform*
adult male rat	*male adult rat*	change in order	*adult* [*male rat*]
tyrosine kinase activation	*tyrosine kinases activation*	internal inflection	[*tyrosine kinase*] *activation*

Figure 1. Bracketing indicators proposed in Nakov and Hearst 2005: 19-21, as seen in Cabezas-García 2019: 98

Nakov and Hearst (2005) also suggest that paraphrases are useful for identifying internal dependencies in MWTs. For instance, *health care reform* is left-bracketed because paraphrases separating those groups can be found, as in "reform in health care." The bracketing indicators in Nakov and Hearst (2005) are very useful for the disambiguation of English MWTs. However, they may not apply to other languages, namely those which do not have markers such as the possessive genitive or internal inflection (Cabezas-García & León Araúz, 2019).

Additional clues to the structure of MWTs are offered in Barrière and Ménard (2014). The authors argue that internal dependencies are based on relational, coordinating or lexical links. To initially determine that certain constituents are linked by a semantic relation, Barrière and Ménard (2014) rely on the use of prepositions. For instance, they search for n1 for n2 in the corpus. If occurrences are found, n1 and n2 are said to encode a semantic relation and are thus bracketed. The main shortfall of this criterion lies in the fact that it cannot be applied to specialized discourse, "where all MWT constituents usually belong to a concept system, and thus encode different semantic relations" (Cabezas-García and León Araúz 2019). So much so that in MWTs such as *desert dust aerosol*, there are semantic relations between all of its constituents: *aerosol* located_in *desert*, *aerosol* made_of *dust*, and *dust* part_of *desert*.

In short, more than twenty years after the development of bracketing models, structural disambiguation still remains problematic. This is especially true for MWTs of more than three constituents. To address the disambiguation of the structural dependencies in 3- and 4-word MWTs, Cabezas-García and León-Araúz (2019) devised a set of indicators and steps based, in turn, on other bracketing models described in the literature, such as the adjacency, dependency or shortening model. These are discussed in more detail in Section 3.3, as part of the methodology for the present study.

2.1.1.3 Semantic interpretation of MWTs

The semantic content of MWTs is hardly ever transparent. Structural and semantic analysis of MWTs go hand in hand, both being essential steps towards the understanding of these units. In structurally disambiguating MWTs, bracketing paves the way for their conceptual analysis (e.g. an error in the bracketing of *boundary layer aerosol* would eventually mislead the decoding of the semantics—instead of an aerosol found in the *boundary layer*, it could be interpreted as an aerosol presented *in* layers *at* a given border).

At the same time, understanding the meaning of the formants is also crucial to bracketing (see Section 2.1.1.2). Thus, the interpretation of the conceptual content in MWTs must naturally begin with the study of the semantic features of their formants. Only then is it possible to express the underlying conceptual relations within these units.

To this end, linguists have traditionally used inventories of semantic relations, ranging from coarse-grained classifications (e.g. Vanderwende's 1994) to fine-grained groupings (e.g. Nastase & Szpakowicz, 2003) to domain-specific inventories (e.g. Rosario et al., 2002). Though semantic relations have drawn the attention of several disciplines since ancient times, they have most recently become a major theme of interest of Computational Linguistics, as they present a "convenient and natural way to organize huge amounts of lexical data in ontologies, wordnets and other machine-readable lexical sources" (Nastase et al., 2013).

Semantic relations have advantages, such as parsimony and generalization (Hendrickx et al., 2013). However, the shortcomings of this approach have been discussed by authors like Nakov (2013), who finds a limitation in the need to choose from a large number of inventories, as well as in the fact that certain combination of nominals can map onto more than one relation (e.g. *desert aerosol* could be understood in terms of aerosol *part_of* or *located_at* desert). Besides this lack of mutual exclusivity, it has been observed that an exhaustive list of relations to enable the description between any combination of nominals does not exist (Downing, 1977; Jespersen, 1942).

Still, decoding the semantic relations is not a step which should not be skipped when interpreting the conceptual information in a CN, especially considering how sometimes the external form of two hyponymic CNs is identical (e.g. Adj+N in *urban aerosol* and *sulfate aerosol*), but they express two different semantic relations (aerosol *located_in* urban; aerosol *made_of* sulfate). These are particularly helpful in translation from a language which presents a high degree of noun-packing, such as English, towards a much

more heavily inflected one, such as Spanish (e.g. *acid rain pollution* > *contaminación derivada de la lluvia ácida*).

2.1.2. Term variation in specialized discourse

Variation is a key element in all languages, and specialized discourse is no exception. However, as León-Araúz (2017) attests in her study, variation is a relatively new area of study in Terminology, "as traditional approaches initially relied on the univocity principle and represented concepts in static universal structures." In other words, the prescriptive approach taken by Wüster's General Theory of Terminology (1968) established that a term was said to allude to only one concept, and a concept was named by only one term. Problematic aspects such as context, phraseology, and variation were perceived as obstacles for effective expert communication (León-Araúz & Cabezas-García, in press). This means that, for a long time, these phenomena were downplayed and largely ignored for the sake of precision, even though the emergence of variants are often motivated by the search for new ways of conveying new meaning (see Section 2.1.2.1 & Section 2.1.2.2)

As Cabezas-García and León-Araúz (2019) expound, variation did not become a focus until the advent of the new theories of terminology, which formulated communicative and cognitive approaches, and acknowledged the variable nature of both terms and concepts (Cabré, 1993; Temmerman, 2000; Freixa, 2006; Faber, 2009; León-Araúz, 2017).

According to Freixa (2006) and Sanz-Vicente (2011), variation in general language is much greater than in specialized discourse. Nonetheless, the latter still exhibits a considerable degree of variation, as specialized domains are dynamic, and dynamism is an inevitable source of variation (León-Araúz 2017).

Variation can affect meanings (i.e. conceptual variation), with a lexical unit being used to name different concepts. This is the case of *aerosol*, which, as defined in Merriam-Webster online, can allude to either a suspension of fine solid or liquid particles in gas, a substance dispensed from a pressurized container as an aerosol, or even the container for the latter. Conversely, one and the same concept can have different denominations, i.e. lexicalized forms. The latter and most frequent one is usually referred to as term or denominative variation (e.g. *black carbon* and *elemental carbon*), and will constitute the focus of our study. MWTs, henceforth CNs, are especially inclined to denominative variation, as they are more lexicalized than other phraseological units (León-Araúz & Cabezas-García, 2019).

As expounded in the following subsections, variation may happen with a specific purpose (Bowker, 1998; Kerremans, 2017; Freixa & Fernández-Silva, 2017; *inter alia*) or it may reveal the novelty of concepts (i.e. neologisms) (Cabré, 1993; Picton, 2011; Carrió-Pastor & Candel-Mora, 2013).

2.1.2.1. Causes of term variation

For decades, the General Theory of Terminology (GTG) prevailed in terminology studies. This meant that the richness of variation was artificially obscured in the interest of the bi-univocal comprehension of terms. There were therefore no grounds to analyze the causes of a phenomenon which only existed as an exception. Nevertheless, in the past two decades term variation has attracted the attention of several scholars (Freixa, 2006; Tercedor-Sánchez, 2011; Daille, 2017; León-Araúz, 2017; *inter alia*)—as has its causes.

Indeed, discovering the causes or types of variation is important for both theoretical and practical reasons (Candel-Mora & Carrió-Pastor, 2012). From a theoretical perspective, it may reflect the mental processes involved in the selection of one specific term. On a practical level, this information could help terminologists or translators in production tasks since they need to know in what context and why they are expected to use one specific variant instead of another.

Traditionally, the reasons for denominative variation have been established within user-based (geographic, temporal and social parameters) and usage-based (tenor, field and mode) frameworks (Gregory & Carroll 1978). However, this broad division only provides a partial representation of this complex phenomenon.

Freixa (2006) envisages a more comprehensive classification of the causes and sub-causes of denominative variation in terminology, ranging from (1) dialectal, (2) functional, (3) discursive and (4) interlinguistic causes to (5) cognitive ones.

Dialectal reasons are based on the geographical, chronological, or social origin of the authors. Functional reasons are linked to field, tenor, and channel, motivated by the author's desire to avoid repetition and/or their search for more creative, emphatic or expressive variants. Interlinguistic variation, on the other hand, is caused by contact between languages, namely in translation contexts. This is the case of CN variants in the English-to-Spanish language pair, where the proliferation of different forms in source language, together with their unsystematic representation in terminographic resources, often results in a broad spectrum of translation for these terms in Spanish.

According to Freixa (2002), cognitive term variants are not only formally different, but also semantically diverse, as they give a particular vision of the concept. This is often referred to as multidimensionality, a phenomenon which explains how the categorization of concepts varies depending on its features. As explained in Cabezas-García (2019), each of those features constitutes a dimension, and when a concept can be organized according a number of dimensions, it is said to be multidimensional.

Dimension	Term variant
+Discoverer	Korsakoff's psychosis
+Symptom	burning-mouth syndrome
+Cause	alcohol-induced amnestic disorder
+Body_part	teeth grinding
+Patient	boxer's dementia
+Result	bedwetting
+Intensity	mild cognitive impairment
+Time	short-term insomnia

+Location	prison psychosis

Figure 2. Conceptual dimensions highlighted in different term variants (León-Araúz 2017: 223)

As described in León-Araúz (2017), term variants which emerge as a result of this phenomenon go to highlight specific facets of one same concept, and are often found to be CNs. Therefore, multidimensionality finds in CNs an ideal basis for its analysis (Meyer & Mackintosh, 1996).

2.1.2.2. Term variation taxonomy

Term variation can acquire a wide range of forms, hence the multiple typologies proposed in the literature (Daille, 2005; Aguado de Cea & Montiel-Ponsoda, 2012; Faber & León-Araúz, 2016; *inter alia*).

Faber and León-Araúz (2016) classify term variants in four broad groups, specifying whether semantics or communicative situations are affected:

(A) Orthographic variants, such as *motor-vehicle pollution* and *motor vehicle pollution*, which do not have geographic causes and do not alter semantics or the communicative situation.

(B) Diatopic variants:

(i) Orthographic variants that do not affect semantics, e.g. *pediatric, paediatric*.

(ii) Dialectal variants, which have the potential to affect semantics if culture-bound factors are involved, e.g. *elevator, lift*.

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

(iv) Calques, which can alter both semantics and the communicative situation, e.g. *moss bag technique > técnica de moss bag*.

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

(i) Abbreviation, e.g. *secondary organic aerosol, SOA.*

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

(D) Diaphasic variants:

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

-Scientific names, e.g. *Passer domesticus, sparrow*.

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

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

-Formulas, e.g. *methane, NH_4.*

-Symbols, e.g. *$, dollar*.

(ii) Informal variants, which affect the communicative situation and possibly 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 alter semantics or the communicative situation if the specialized domains have different term preferences, e.g. *ultrafine particles* and *nanoparticles* represent the same concept except that the first one is used in Toxicology and the latter in Engineering.

(E) Dimensional variants, which are often CNs and affect semantics because they activate different dimensions of the same concept, e.g. *esmog fotoquímico [photochemical smog], niebla tóxica estival [summer smog]*.

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

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

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

(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]*.

Nonetheless, as advanced by Daille (2005), term variation typologies are ultimately dependent on the final application for which they have been established.

To study term variation in translation contexts, Cabezas-García and León-Araúz (2020) expanded the variation categories described in Faber and León-Araúz (2016) in a typology specifically conceived to characterize translation equivalents of CNs (see Section 4.2 for further discussion).

2.1.2.3. Consequences of term variation

As can be deduced from the previous section, term variation often has conceptual and communicative implications. That is to say that the use of one term or another may affect the semantics of a concept or the communicative situation in which the concept is activated (León-Araúz & Reimerink, 2016). While some authors argue that changes in the form sometimes fail to affect meaning (Fernández-Silva et al., 2009), the change in form often brings along a shift in perception. As a result, several classifications based on the semantic distance of term variants on a monolingual level have emerged. Aguado de Cea and Montiel-Ponsoda (2012) and Fernández-Silva (2018) distinguish three main groups: variants with (1) minimum, (2) medium, and (3) maximum semantic distance.

The first set encompasses terms which are conceptually equivalent. For Aguado de Cea and Montiel-Ponsoda (2012) these include synonyms, such as graphical and orthographical variants, inflectional variants, and morphosyntactic variants. Fernández-Silva (2018) adds morphological variants and specifies that, in MWTs, synonymy can affect just one of the constituents.

Variants with a medium semantic distance differ from these in that they are partial synonyms or terminological units that highlight different aspects of the same concept, such as stylistic or connotative variants (*insect, bug*), diachronic variants (*tuberculosis,*

phthisis), dialectal variants (*sidewalk, pavement,*), pragmatic or register variants (*horripilation, goosebumps*), and explanatory variants (*immigration law, law for regulating and controlling immigration*) (Aguado de Cea & Montiel Ponsoda, 2012). As explained in León-Araúz and Cabezas-García (in press), maximum distance in CNs, on the other hand, refers to conceptual changes which are reflected in the modifiers, whether they are subject to reductions (*diesel exhaust pollution*, *diesel pollution*), additions or deletions of non-defining characteristics (*anthropogenic emission, anthropogenic air emission*), or the use of a different defining feature (*wintertime aerosol, ultrafine aerosol*) (Fernández-Silva, 2018).

Variation in translation contexts also has consequences. Ignoring term variation when translating specialized texts has been documented to be equally problematic (see e.g. Resche, 2004). It can result in translators over-standardizing, "creating consistency in places where the use of variants was deliberate and well-reasoned" (Bowker & Hawkins, 2006: 80).

2.1.2.4. Neology and secondary term formation

Terminology and neology are closely related, as documented in the existing literature (Kageura, 2002; Sanz-Vicente, 2011, 2012a, 2012b; Humbley & García-Palacios, 2012; Cabré et al., 2012; Pecman, 2012, 2014; Fernández-Domínguez, 2016). The constant evolution of knowledge in specialized fields entails the parallel development of terminology. New realities and designation requirements lead to new terms, i.e. terminological neologisms, also called neonyms (cf. Rondeau, 1984: 121-122).

Cabré et al. (2012) outlines different situations which can prompt the creation of new terms: (a) the naming of a new discovery or invention; (b) in a translation context, the need to propose an equivalent for a term in the source text which had so far only been named in the language that created it; or (c) the establishment of appropriate terms in language planning. While the last one escapes the scope of our study, the two first are most likely to arise.

Neologisms have been categorized in various ways. As Cabré et al. (2012) sketch in their study, some important contributions to classify neologisms in general are the distinction between general neology and specialized neology (Rondeau, 1984; Cabré, 1993; Humbley, 2006) as well as that between denominative neology—also called referential neology—and stylistic neology (Guilbert, 1975), or called expressive neology by Cabré (1993). The first is specially related to terminology, stemming from the need to name a new concept, whereas the second is associated with communication at discourse level. On the other hand, Boulanger (1989) proposes a distinction between spontaneous neology and planned neology.

For the purpose of our study, we will now delve into specialized neology (hereafter referred to as terminological neology), as it plays an important role in English and Spanish term formation.

Following the dichotomy *néonymie d'origine* and *néonymie d'appoint* previously established by Rondeau (1984), Sager (1990: 80) draws a distinction between primary and secondary term formation. This distinction classifies specialized neologisms into two large groups based on the context: those appearing in languages together with knowledge production, and those appearing in processes of knowledge transfer between different speakers' communities (Cabré et al., 2012).

The second is the case of the English-Spanish language pair. English being the *lingua franca* of scientific and technological communication, terminological neology proves to be "a one-way process, from English to the rest of languages" (Sanz-Vicente, 2012). Thus, English constitutes the language of primary term formation while the rest are based on secondary term formation, i.e. importing and adapting terms from English.

This phenomenon is also explained through the notion of terminological dependency, defined as "the subordinating relationship established between two languages in a specific terminological field" (Humbley & García-Palacios, 2012). This phenomenon is particularly evident in scientific contexts—the more innovative the research, the more marked it proves to be.

In Sanz-Vicente (2012), secondary term formation in expert to expert communication is tackled through the corpus-based analysis of CNs (referred to as terminological syntagmatic compounds (TSCs)). In translating English TSCs into Spanish, results showed an underlying preference for calques (either through direct loans or a literal translation), even where their semantic relations were not transparent. Achieving morphosemantic similarity arises as the priority when transferring TSCs into Spanish. With regards to this, Humbley and García-Palacios (2012) advance the idea that scientists whose native language is not English simply aim to "convey the specialized concept without regard to the linguistic elements or the terminology used for this purpose." They have interiorized the hegemony of English, subordinating their native naming process to criteria imposed by the *lingua franca.*

As a result, CNs feature a propensity for term variation, which only adds to the complexity of translating these units, especially considering the unsystematic treatment they receive in terminographic resources.

2.2. Theory and practice of Terminology

In the past decade, the study of terminology and specialized translation has undergone a cognitive shift. Terminology theories have evolved[2] from the Wüster's prescriptive model towards more descriptive ones, such as Socioterminology, the Communicative Theory of Terminology (CTT), and Sociocognitive Terminology Theory (STT), ultimately leading to the one our study is largely based on: Frame-Based Terminology

[2] For an analysis of the different approaches to Terminology from Wüster's General Theory of Terminology onwards, see León-Araúz, P. (2009). Representación multidimensional del conocimiento especializado: el uso de marcos desde la estructura hasta la microestructura (PhD). Granada: Universidad de Granada.

(FBT). Accordingly, the following sections explain the basic principles of this theory and describe EcoLexicon, the practical application of FBT.

2.2.1 The cognitive approach of Frame-Based Terminology

Frame-Based Terminology (FBT), henceforth FBT, is a very recent cognitive approach to terminology proposed by Faber (2009, 2012, 2015). It operates on the premise that, in scientific and technical communication, "specialized knowledge units activate domain-specific semantic frames that are in consonance with the users' background knowledge" (Faber et al., 2016: 73).

FBT directly links specialized knowledge representation to cognitive linguistics and semantics (Faber, 2012). As such, it integrates some of the premises of Communicative Theory of Terminology (Cabré, 1993) and Sociocognitive Theory of Terminology (Temmerman, 2000, 2001), which also study terms by analyzing their behavior in texts. It also maintains that trying to find a distinction between terms and words is no longer fruitful or even viable, since they both show the same behavior but in different contexts. However, FBT differs from these approaches in that its methodology combines premises from psychological and linguistic models and theories such as the Lexical Grammar Model (Faber & Mairal, 1999; Martín-Mingorance, 1989: 227–253), Frame Semantics (Fillmore, 1985: 222–254), the Generative Lexicon (Pustejovsky, 1995) and Situated Cognition (Barsalou, 2003).

As its name implies, FBT applies the notion of 'frame', defined as a cognitive structuring device based on experience and held in long-term memory, which provides the background knowledge and motivation for the existence of words in a language as well as the way those words are used in discourse.

The FBT approach to terminology and terminology management sets the theoretical framework for applications such as the multilingual and multimodal terminological resource EcoLexicon (Faber et al., 2014; León-Araúz et al. 2016; San Martín et al., 2017; *inter alia*).

FBT focuses on (1) conceptual organization, based on frames or events; (2) the multidimensional nature of terminological units, by accounting for both hierarchical and non-hierarchical relations; and (3) the extraction of semantic and syntactic information through the use of multilingual corpora (Faber, 2009; Faber et al., 2016; Buendía-Castro, 2013; *inter alia*). One of the basic premises of this approach is that conceptual networks are based on an underlying domain event, which generates templates for the actions and processes that take place in the specialized field as well as the entities that participate in them (Faber, 2009), as illustrated in Figure 3.

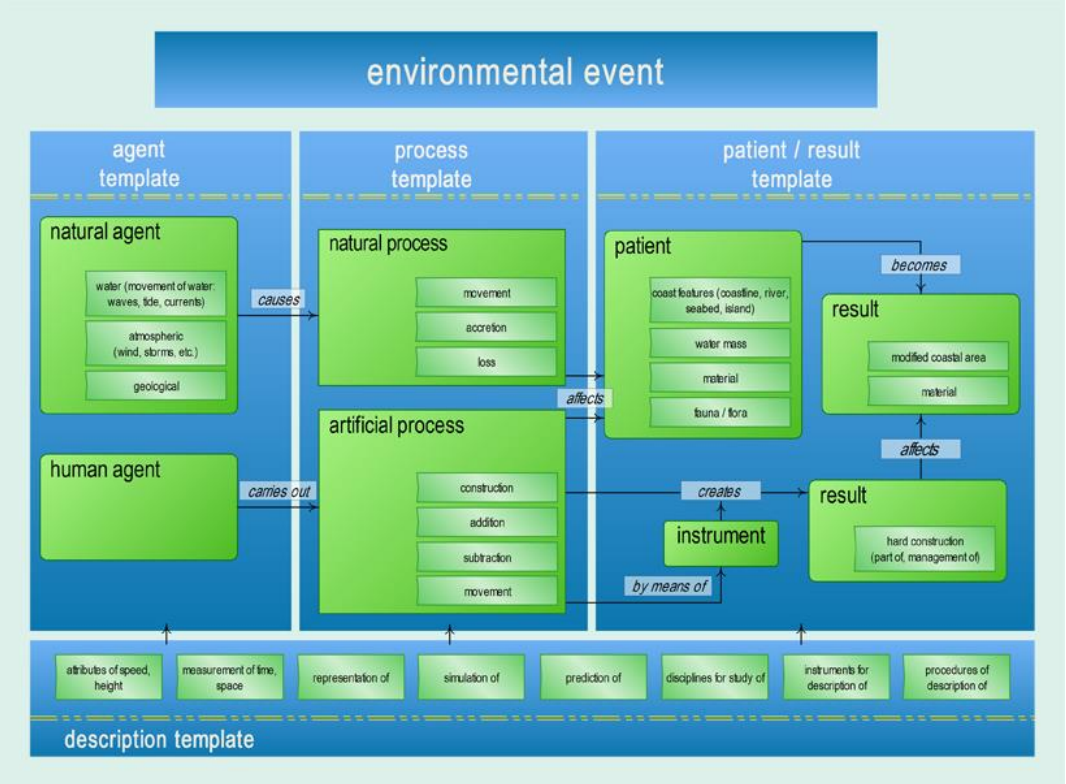

Figure 3. The environmental event, as seen in <http://lexicon.ugr.es/fbt>

Generic categories are thus organized in a domain event (Barsalou, 2003: 513; Faber et al., 2005), which provides a frame for the configuration of more specific concepts. The specific concepts within each category are structured in a network where they are linked by both vertical (hierarchical) and horizontal (non-hierarchical) relations.

2.2.2 Terminological knowledge bases: EcoLexicon

EcoLexicon (<https://ecolexicon.ugr.es/>) is the practical application of FBT. It is a terminological knowledge base (TKB) of environmental science with terms in six languages: English, French, German, Modern Greek, Russian, and Spanish. TKBs are cognitive terminological resources that represent the specialized knowledge of a certain field through related concepts and the terms that designate them in one or various languages (Gil-Berrozpe, 2017).

The FBT methodology used to design EcoLexicon is two-parted. The underlying conceptual framework of a knowledge-domain event is specified by means of an integrated top-down and bottom-up approach. On the one hand, information is extracted from a corpus of texts in various languages, specifically related to the domain (bottom-up). The top-down approach, on the other hand, gathers information provided by specialized dictionaries and other reference material, complemented by the help of experts in the field (Faber, 2009: 124).

A notable example of CN representation in TKBs can be found in EcoLexicon's phraseology module. In Cabezas-García (2019) a proposal specifically oriented towards the representation of these units is designed, including the following views: (i) CN formation from one or more terms; (ii) equivalents in the English-Spanish language pair; (iii) morphosyntactic combinations of CN constituents; (iv) semantic combinations (categories, roles and relations present in CNs), and (v) summary. This terminographic resource provides a wide range of information showed by means of an enriched structure (e.g. grouping term variants in concepts and specifying the relation of hyponyms with their hypernyms, as well as the different dimensions emphasized by terms), thus constituting an outstanding example of how the design of TKBs should always seek to find the needs of its myriad of users.

2.3. Translation and Computational Linguistics

Due to the properties discussed in section 2.1, most compound terms pose a challenge for the computational processing of natural languages—notably complex nominals. Research on computational linguistics contributes to the development of not only human and machine translation studies, but also a myriad of other Natural Language Processing applications, such as building ontologies (Venkatsubramanyan & Perez-Carballo, 2004), and machine translation (Baldwin & Tanaka, 2004).

In this section we will provide an overview of the latest advances and trends in the field of machine translation, as well as an introduction to the use of corpora in translation studies.

2.3.1. Corpus linguistics

The contribution of corpus linguistics to translation studies has been significant, as corpora "have provided a basis for descriptive research and allowed for the empirical testing of theoretical hypotheses" (Zanetti, 2014). The first corpus-based translation studies date back to the 1980s. Though there is no single agreed upon definition of corpus linguistics (cf. Taylor, 2008), in this study we share the vision of McEnery and Wilson (1996) and Meyer (2002), who define it as "an approach or a methodology for studying language use."

Traditionally, a distinction is made between parallel corpora and comparable corpora. A parallel corpus can be defined as a collection of source texts and their aligned translations in one (bilingual corpora) or more languages (multilingual corpora). These can be uni-directional (e.g. from English into Spanish or from Spanish into English alone), bi-directional (e.g. containing both English source texts with their Spanish translations as well as Spanish source texts with their English translations), or multi-directional (e.g. the same piece of writing in English, Spanish, and Chinese) (McEnery & Xiao, 2007: 2). Notably, texts simultaneously produced in multiple languages, such as

EU legistation, are also categorized as parallel corpora (cf. Hunston, 2002: 15). Parallel corpora facilitate the identification of translation equivalences. However, they are ultimately scarce, which has led to a greater presence of comparable corpora in translations studies in the latest years (Lapshinova-Koltunski, 2013; Jiménez-Crespo & Tercedor-Sánchez, 2016; León-Araúz et al., 2020; *inter alia*). Comparable corpora are generally defined as corpora containing components that are collected using the same sampling frame and similar balance and representativeness, i.e. components which are not translations of each other but still match in terms of proportion, genre, domain and sampling period (McEnery & Xiao, 2007: 3). However, it should be borne in mind that the terminology is somewhat unstable (Zanettin, 2012: 149), since the distinction between the two types of corpora is not always clear cut (Fantinuoli & Zanettin, 2015: 3).

The applications of corpus linguistics to translation research are both theoretical and practical (Hunston, 2002: 123). From a theoretical perspective, corpora allow us to study the translation process by exploring how an idea in one language is conveyed in another language and by comparing the textual and linguistic features. From a practical perspective, corpora provide a "workbench for training translators and a basis for developing applications like machine translation (MT) and computer-assisted translation (CAT) systems" (Xiao, 2007).

Altogether, corpora provide a window into real-life language use, be it through original texts or translations, which makes them an invaluable documentation source for translators. Later in this study, we will delve into the use of corpora as a reliable tool to solve different translation problems, namely the analysis of CNs.

2.3.2. Machine translation

Machine translation (MT) investigates the approaches to automatically convert text in one natural language into another, producing fluent text in the output language without altering the meaning of the input text. It is a subfield of computational linguistics that draws ideas from linguistics, computer science, information theory, artificial intelligence, and statistics.

As explained in Maučec and Donaj (2019), the first approaches for MT were based on linguistic rules that were used to "parse the source sentence and create the intermediate representation, from which the target language sentence was created." These rule-based translation methods include dictionary-based MT, transfer-based MT, and interlingual MT.

While rule-based approaches are useful to translate between closely related languages, these are costly and time-consuming to implement and maintain, as they require linguistic experts to apply language rules to the system. As rules are added and updated, there is also the potential of generating "ambiguity and translation degradation." (Maučec & Donaj, 2019).

Statistical MT, based on statistical methods (Koehn et al., 2003), was a dominant approach over the past 20 years. However, it faces many obstacles, such as the difficult

processing of highly inflectional languages (especially as target languages). These days, with statistical MT almost reaching the limits of its capacity, the deep learning-based approach of neural MT is rapidly becoming the technology of the future (Maučec & Donaj, 2019). However, as Way (2018) explains, "[Neural] MT output can be deceptively fluent; sometimes perfect target-language sentences are output, and less thorough translators and proofreaders may be seduced into accepting such translations, despite the fact that such translations may not be an actual translation of the source sentence at hand at all."

Though some translators fail to acknowledge the capability of MT, there is no real doubt that MT is currently being deployed by millions of people on a daily basis (Way, 2018). In 2016 Google stated that the average daily volume of its MT system was about 143 billion words a day across 100 language combinations. If all the translation requests that DeepL and other online systems respond to on a daily basis are added in, it is hard to deny the "utility of online MT across a wide range of use-cases and language pairs to millions of distinct users" (Way, 2018: 162).

Recent research (see e.g. Nunes-Vieira, 2018) suggests that MT opens new opportunities for translators, as they can benefit from integrating post-editing in the translation process, significantly reducing the time and therefore the costs. Nevertheless, there is still the question of whether the quality of translation will remain at the same level. With MT usage increasing exponentially, human evaluation of MT output remains crucial. While the integration of human and machine translation can result in a promising workflow in certain contexts, the features characterizing specialized translation (see Section 2) are likely to make it difficult for MT to provide translation at the expected quality level. Furthermore, the task of automatically translating specialized texts becomes even more challenging for these MT systems when faced with CNs.

3. Materials and methods

The sections below describe the different materials used to conduct our study, which include an English-Spanish comparable corpus on air pollution and air quality treatment, a selection of terminographic resources and other open-access online corpora, as well as the use of various MT engines. It also explains the methodology followed in (1) the compilation of the corpus, (2) the extraction and analysis of the CNs, (3) the identification of target language variants, and (4) the criteria established for the evaluation of MT output.

3.1. Corpus design and compilation

For the purpose of this study two corpora were manually compiled: an English corpus and a Spanish corpus of some 1,500,000 words each. The corpora are made up of specialized texts selected on the basis of the following criteria: (1) topic, (2) language, (3) subject field and (4) text type.

Though all texts deal with the topic of air pollution and air quality treatment, the English corpus is specifically composed of scientific articles published in high impact journals such as *Atmospheric Environment*, *Environmental Pollution* and *Atmospheric Pollution Research*.

Air pollution and air quality treatment are the object of study of several disciplines ranging from Medicine to Engineering. However, the present study focuses on the scientific production of the subject fields of Physics, Chemistry, Meteorology and Chemical and Environmental Engineering.

The Spanish corpus is slightly different for various reasons. Scientific research is mainly written and published in English nowadays, as it constitutes the *lingua franca* of the sciences. The dominance of English in scientific research is further linked with the threat of domain loss, which can be defined as the situation in which scientists eventually lose the ability to communicate in their native language on all levels of a specialized domain in favor of the preferred language—English (Ferguson, 2007). As Humbley and García Palacios (2012) point out, texts written in Spanish and French on scientifically advanced subjects are thus currently "less numerous, less well disseminated and have less prestige than their equivalents in English, the only and important exception in the academic context being the doctoral thesis." This has a series of implications on translation and term variation, as could be read in section 2.1.2.4.

In other words, the cutting-edge research found in the numerous scientific articles written in English cannot be found in Spanish articles, which led us to consider other text types presenting the high level of specialization needed for this study. As a consequence, the Spanish corpus is mainly composed of doctoral theses, all dealing with the same subject field as the texts in the English corpus. It also contains other forms of academic

research such as bachelor's and master's dissertations, as well as legislation, book chapters, reports and a limited number of articles.

After collecting the texts, both corpora were compiled in the well-known corpus query system Sketch Engine (Kilgarriff et al., 2004).

3.2. Extraction of 3- and 4-word CNs within the English corpus

Multi-word term extraction is a difficult process for many of the same reasons as CN identification, namely syntactic flexibility and ambiguity. Though recent research has pointed out the multiple advantages of automatic CN extraction, especially via association measures (Baldwin & Kim, 2010), the CNs that constitute the object of our study were extracted manually using Sketch Engine. This approach, however time-consuming, guaranteed the relevance of such terms for the purpose of our study.

As pointed out in previous sections, scientific research is dominantly published in English. The use and formation of new CNs is thus particularly prolific in such language, which is why we extracted the CNs for our study from the English corpus compiled for that purpose.

In order to avoid selecting terms that were only the product of a certain author's use of language, we extracted only those used in at least three different texts, similarly to Sanz-Vicente (2011) and Cabezas-García (2019). Bowker (1998: 493) suggests that terms should only be considered if present in a minimum of 12 texts. However, we deem three a more reasonable number for a corpus of our size, as we might otherwise be leaving out some interesting CNs worth studying. This is especially likely to happen in such a rapidly developing field of study as air pollution and air quality treatment research, with less widespread CNs being the product of state-of-the-art innovation.

The present study focuses on complex nominals, i.e. expressions with a head noun. Thus, the first step in the extraction process was the search for the nouns that make up the corpus, which would later be found acting as heads or modifiers of CNs. While adjectives such as *anthropogenic* or *atmospheric* were also rather frequent, nouns were more productive for the purposes of our study. We used the following CQL (Corpus Query Language) expression to extract these nouns, ordering the results by frequency and lemma: [tag="N.*"]. Once the resulting list was thoroughly examined, six nouns were selected as the main participants in the EVENT OF AIR POLLUTION: *air*, *quality*, *emission, aerosol, particle, pollutant*.

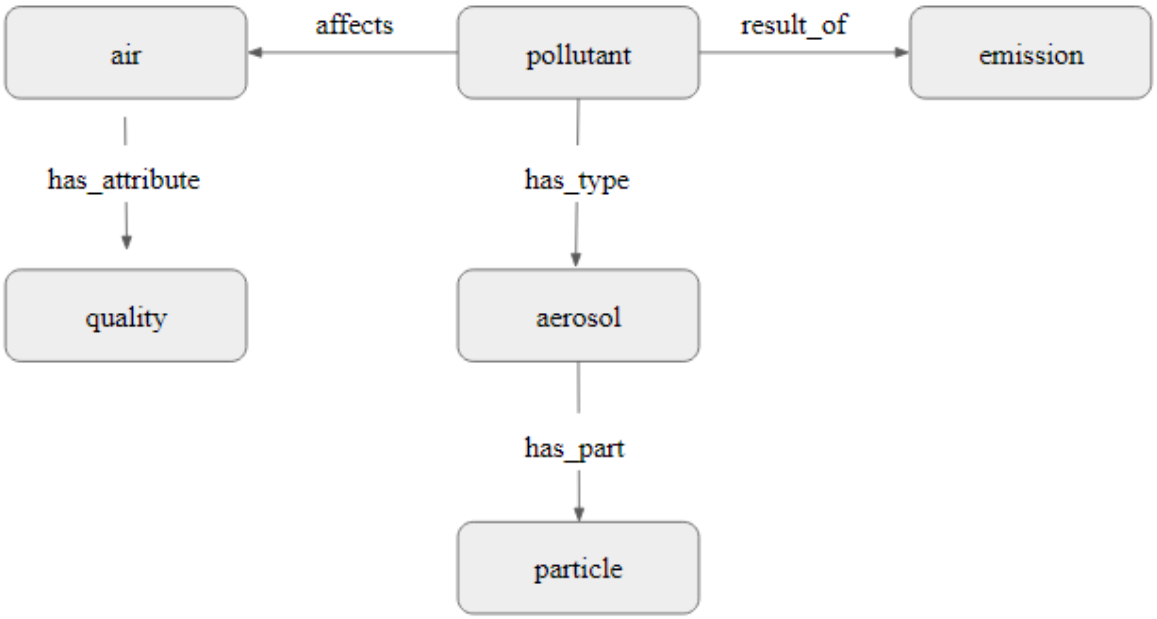

Figure 4. Simplified version of the conceptual system underlying the AIR POLLUTION EVENT

One of the aims of our study was to examine term and translation variation in the air quality treatment domain. For that purpose, we needed to extract terms widespread in the scientific community, which would allow us to find target language equivalents (1) in our Spanish comparable corpus, (2) in various terminographic resources, and/or (3) in other online corpora. Dynamism is a key feature of specialized domains such as the one in hand, which is an inevitable source of variation (León-Araúz, 2017). Thus, the dynamic, unstable translation equivalents for these frequently used CNs will serve to characterize translation variation in this subdomain.

To this end, the previously extracted nouns were then used as CN heads in CQL queries, which also allow to search for specific morphosyntactic patterns. The CQL expression in Figure 5 searches for the lemma *aerosol* ([lemma="aerosol"]) (or *pollutant*, *emission*, *air*, *quality*, *particle* in the following queries) preceded by nouns, adjectives, adverbs, past participles, or present participles ([tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]) appearing two or more times ({2,}). These are all elements which have been found to pre-modify CNs in the literature. Therefore, we search for right-headed CNs such as *fine mode aerosol*, as pre-modification has been documented to be the most frequent structure for these compounds. Nouns or adjectives are excluded on the rightmost part of the query to avoid extracting longer terms in which *aerosol* is not the head ([tag!="N.*|JJ.*"]).

```
[tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]{2,}[lemma="aerosol"][tag!="N.*|JJ.*"]
```

Figure 5. CQL query to extract CNs whose head is *aerosol*

We were also interested in studying variation in CNs where these nouns act not as the head but as the modifier. To this end, we performed the CQL query in Figure 6, which elicits pre-modified CNs such as *low molecular weight aerosol proteinaceous matter*. More specifically, the query below searches for the lemma *aerosol* ([lemma="aerosol"]) which may be followed or preceded by nouns, adjectives, adverbs, past participles or present participles ([tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]{0,}). Because we search for complex nominals, the head is necessarily a noun ([tag="N.*"]) which cannot be followed by other adjectives or nouns ([tag!="JJ.*|N.*"].

```
[tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]{0,}[lemma="aerosol"][tag="N.*|JJ.*|RB.*|VV
N.*|VVG.*"]{0,}[tag="N.*"][tag!="JJ.*|N.*"]
```

Figure 6. CQL query to extract CNs where *aerosol* acts as modifier

The terms extracted through these queries will also allow us to study how nouns describing a key concept in a specific domain are sometimes progressively turned into highly-productive noun heads by linking them to other lexical elements, thereby conveying more characteristics of the concepts and resulting in the formation of long-strung hyponymic CNs (e.g. *emission > volatile organic compound emission > non-methane volatile organic compound emission > industrial non-methane volatile organic compound emission*).

Another focus of the study was to address the more obscure CNs—which may or may not include the nouns which make up the EVENT OF AIR POLLUTION—from both a structural and a cognitive point of view. Complex nominals such as *portable light-scattering aerosol monitor* pose many challenges, not only for terminologists but also for human translators and MT systems alike. These challenges stem from the linguistic properties discussed in section 2.1.1, namely their lack of transparency, and the omission of elements, which significantly hinders the task of interpreting them both semantically and structurally. In contrast with the less syntactically and semantically complex terms extracted in the first phase, pre-established translation equivalents for these CNs are scarce or simply non-existent, as they are the product of the latest research innovations, documented only in English. Unsurprisingly, these longer terms are seldom found in parallel and comparable corpora, and/or terminographic resources, which only adds to the greater difficulty they present. Thus, the terms extracted next will serve as a start point to develop a series of guidelines for the translation of these CNs from English into Spanish. For this purpose, we performed the following CQL queries to extract three- and four-word CNs, respectively:

```
[tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]{2}[tag="N.*"]
[tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]{3}[tag="N.*"]
```

Figure 7. CQL query to extract 3- and 4-word CNs

Both CQL queries elicit CNs with a nominal head [tag="N.*"], which can be preceded by two ({2}; first CQL query) and three ({3}; second CQL query) nouns, adjectives, adverbs, past participles or present participles. We did not search for CNs of more than four constituents in anticipation of a low number of occurrences, which would complicate the extraction of results. Nonetheless, a close observation of the list resulting from these last queries revealed that some of those 4-word CNs were in fact 5- and 6-word combinations (e.g. *PUF disk passive air sampler > polyurethane foam disk passive air sampler*). Similarly, the use of acronyms within some of the extracted CNs suggests that the conceptual information conveyed is greater than expected (e.g. *anthropogenic aerosol ERF > anthropogenic aerosol effective radiative forcing*).

These longer terms are often the product of cutting-edge research published only in English. The lack of documented translations for such 3-, 4- and 5-word terms will drive us to carry out an analysis of their semantic relations and their syntactic structure, which will result in the implementation of a protocol to translate them into Spanish. In doing so we would be taking a step further in complex nominal research, which up until now has focused mainly in shorter compounds. The intrinsic complexity of these terms is particularly relevant to our study, as it will also allow us to test the limits of some of the most frequently used MT systems worldwide.

Though small in number, the extracted CNs (see sample in Figure 8) were the product of a thorough extraction process which will serve as starting point for a cross-sectional analysis of such terms, allowing us to study them from very different angles.

Three-term CNs	Freq.	Four-, five-, six- and seven-term CNs	Freq.
aerosol optical depth	193	PUF disk sampler	60
passive air sampler	141	age toluene SOA particle	38
dry deposition flux	77	diurnal FRP cycle	37
black carbon aerosol	64	industrial NMVOCs emission	29
air pollutant concentration	49	anthropogenic aerosol ERF	28
fine particulate matter	46	air quality monitoring station	23
sea salt aerosol	24	hybrid single-particle Lagrangian integrated trajectory	10
point source emission	22	particle number size distribution	10

Figure 8. Sample of the CNs extracted as a result of the described queries

Finally, after verifying the CNs in concordance lines, we performed the following complementary CQL queries in search for denominative variants of the extracted CNs, some of which had already been spotted during the compilation of the corpus. Once

identified, these would guide us in the structural disambiguation and semantic interpretation tasks.

[tag="RB.*"]{0,}[word="referred"][tag="RB.*"]{0,}[word="to"]
{0,}[word="as"]{0,}[tag="N.*|JJ.*|RB.*|VVG.*|VVN.*"]{0,}[tag="N.*"]

[tag="RB.*"]{1,}[word="known"][tag="RB.*"]{0,}[word="as"]{1,}[tag="N.*|JJ.*|R
B.*|VVG.*|VVN.*"]{0,}[tag="N.*"]

Figure 7. CQL queries to extract synonyms through KPs

3.3 Structural disambiguation of CNs

Following the extraction of terms, a series of disambiguation tasks were performed in order to access the internal structure of CNs.

While the bracketing models proposed in the literature are oriented towards the disambiguation of 3- and 4-word CNs, they were still considered for the 5-, 6- and 7-word CNs in our sample. Instead of comparing the results of all the possible combinations (which would be time-consuming and highly confusing), we performed the queries needed only to confirm the suspected bracketing grouping—as a working translator would do during a given translation commission. The bracketing signs described in Section 2.1.1.2 (cf. Nakov & Hearst, 2005), together with our observations while extracting the data, were what ultimately led us to suspect a specific bracketing. Indeed, the denominative variants grouped during the extraction process clarified the dependencies to a certain extent, which were subsequently confirmed through queries in the corpus. We checked that the candidate CN complied with at least two of the following indicators (cf. León-Araúz & Cabezas-García, 2019):

(1) Adjacent groupings within the CN appeared as independent terms
(2) The most frequent adjacent grouping was still more frequent than other dependencies
(3) Bracketing groupings did not allow the insertion of external elements modifying their meaning
(4) Bracketing groupings were found combined with other elements
(5) Bracketing groupings had synonyms or antonyms.

For instance, in the case of *polyurethane foam disk passive air sampler*, the coexistence of the denominative variants *polyurethane foam (PUF) disk passive air sampler, PUF-based passive air samplers PUF-PAS, PUF disk sampler* and *PUF disk passive air sampler* suggested the following groupings: *[polyurethane foam disk] [passive air sampler]*. Next, we applied the first indicator, which appeared to confirm our suspicions. We sought further confirmation with the second and fourth criterion, and the

high number of occurrences of *passive air sampler*, both independently and in combination with other elements (*XAD passive air sampler, indoor passive air sampler*) appeared to be conclusive. Finally, the multiple occurrences of the antonym of one of the possible groupings (*active air sampler*) served as final confirmation.

3.4 Corpus-based semantic analysis of CNs

The proposal of translation correspondences for CNs required prior decoding of the semantics within these terms. More specifically, we focused on the analysis of the internal relations linking the constituents of a CN, as the correct interpretation of these is what will ultimately give the translator the freedom to produce target language oriented, non-calqued equivalents.

The process of decoding the semantics of the term was done by means of paraphrases. Nonetheless, because the extracted sample of terms included CNs of up to 7 constituents, finding a paraphrase which perfectly clarified the relations of each constituent with the rest of the components was unlikely. Instead, we performed queries which considered up to three of their formants at a time, guided by the findings of the structural disambiguation process.

```
[lemma="emission"][]{0,2}[tag="IN" &
lemma!="like"][]{0,2}[lemma="point"][lemma="source"][lemma!="emission"]
within <s/>
```

Figure 8. CQL query to extract prepositional paraphrases of *point source emission*

The CQL query in Figure 8, for instance, elicited paraphrases where p3 was linked to p1p2 through a preposition but not the lemma *like* (which would spoil the results), all within the same sentence, such as *emission **assigned to** any individual point source* or *emissions **from** point sources*.

We also searched for verb paraphrases, using queries such as the one in Figure 9.

```
[lemma="size"][lemma="distribution"][lemma!="number"][]{0,10}[tag="VV.*"]
[]{0,10}[lemma="number"] within <s/>
```

Figure 9. CQL query to extract verb paraphrases of *particle number size distribution*

When looking for verb paraphrases, we allowed a span of up to 10 words in between to obtain better results, while still having all three propositions in the same sentence. The CQL query above elicited verb paraphrases of p2p3 V p1. Similarly, we performed others which helped us find paraphrases for other combinations (see Section 4.4 for further illustration).

3.5 Identification of target language equivalents

Following the analysis of the source terms, various resources were used for the identification of their translation equivalents. Firstly, we searched for Spanish equivalents in a selection of multilingual terminographic resources. Subsequently, we made use of the Spanish comparable corpus to find original production of CNs. Finally, we resorted to parallel corpora and the web.

3.5.1 Extraction of translation variants within terminographic resources

When faced with a term that escapes their knowledge, translators' first instinct is to turn to terminographic resources of their choosing in hopes to find the solution to their translation problem. In the case of CNs, the search for the exact equivalent of a given term is often unproductive, as their representation in these resources is usually rather unsystematic. However, when dealing with such multi-word units, terminographic resources still constitute an ideal place to start the search for the right equivalent. Indeed, before they start looking for target language (TL) correspondences in comparable corpora, translators first need to know which shorter terms the CN is expected to contain in the TL in order to perform the right queries.

Accordingly, since one of the main interests of our study was the analysis of variation in English-to-Spanish translation contexts, we began our search for variants in two multilingual terminographic resources: IATE and TERMIUM Plus. IATE (<https://iate.europa.eu>), which stands for Interactive Terminology for Europe, is the shared terminology database of the institutions and agencies of the European Union. The database is fed by EU translators and terminologists *ad hoc*, i.e. according to their needs. As such, it describes concepts from a wide range of specialized domains—including the environment—in the official languages of the EU. TERMIUM Plus (<https://www.btb.termiumplus.fgc.ca/>) is a terminology database created and maintained by the Government of Canada, which contains millions of terms from various specialized fields in four languages: English, French, Spanish and Portuguese.

Firstly, each CN was looked up in full (see Figures 10 & 11) filtering the searches by domain (environment).

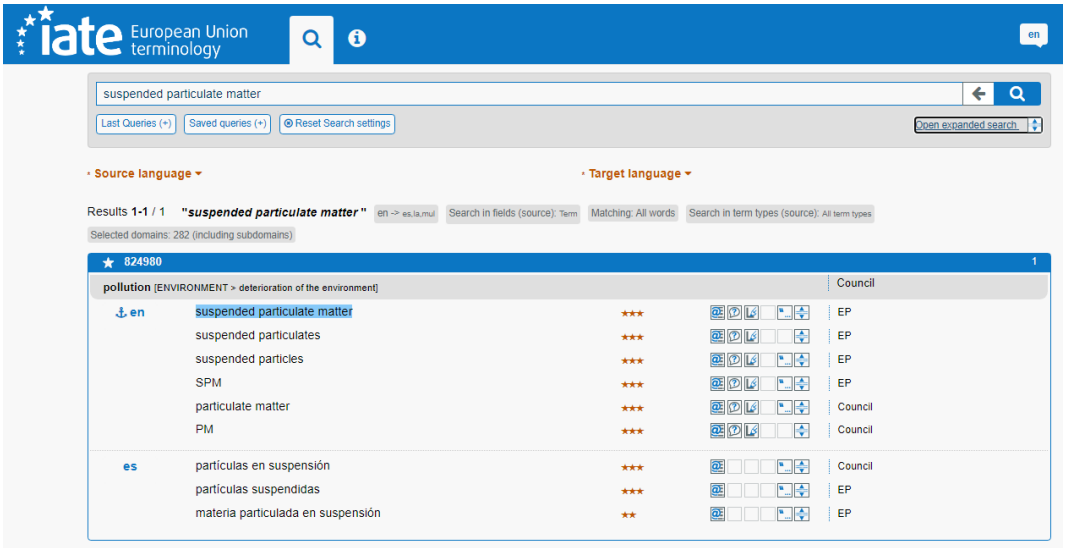

Figure 10. English variants for *suspended particulate matter* and their respective Spanish correspondences in IATE (last accessed on June 11th 2020)

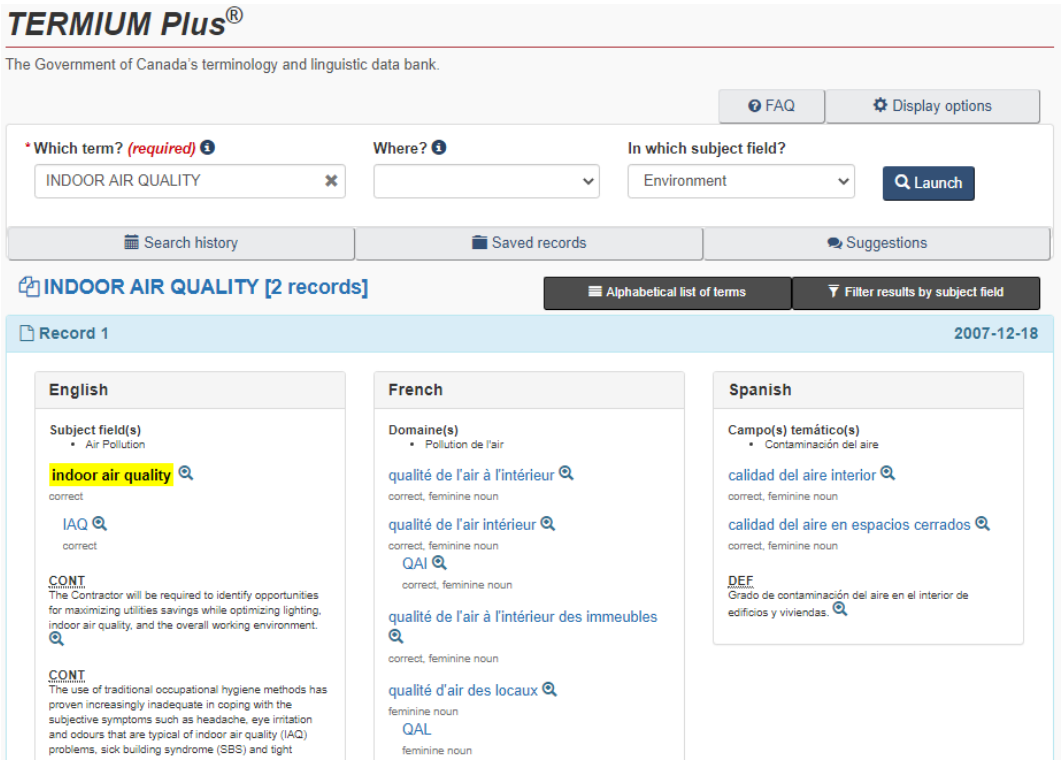

Figure 11. English, French and Spanish variants for *indoor air quality* found in TERMIUM Plus (last accessed on June 11th 2020)

When no results were retrieved, we checked whether other shorter terms encompassed by those same CNs could be found. In order to bridge this gap, we

continued our search for Spanish variants by means of corpora, as described in the following section.

3.5.2 Extraction of Spanish variants within the comparable corpus

After checking the two chosen multilingual databases in search for Spanish equivalents of the CNs, we performed a series of queries in Sketch Engine with the aim of extracting variants originally produced in Spanish. These queries were guided by (1) the denominative variants of the source CNs identified during the extraction of these (Section 3.2), and (2) the information gathered while checking the terminographic resources (Section 3.5.1). In a similar way to Cabezas-García (2019), we made use of the filter context in Sketch Engine to find target language equivalents of a given CN based on the constituents it was expected to have after examining the data collected in (1) and (2).

For instance, in the case of *fine mode particulate matter*, a series of source term variants had been identified in the extraction process, including *fine particles*, *fine mode aerosol* and *PM2.5*. The equivalents for some of their constituents had subsequently been found in the multilingual terminographic resources.

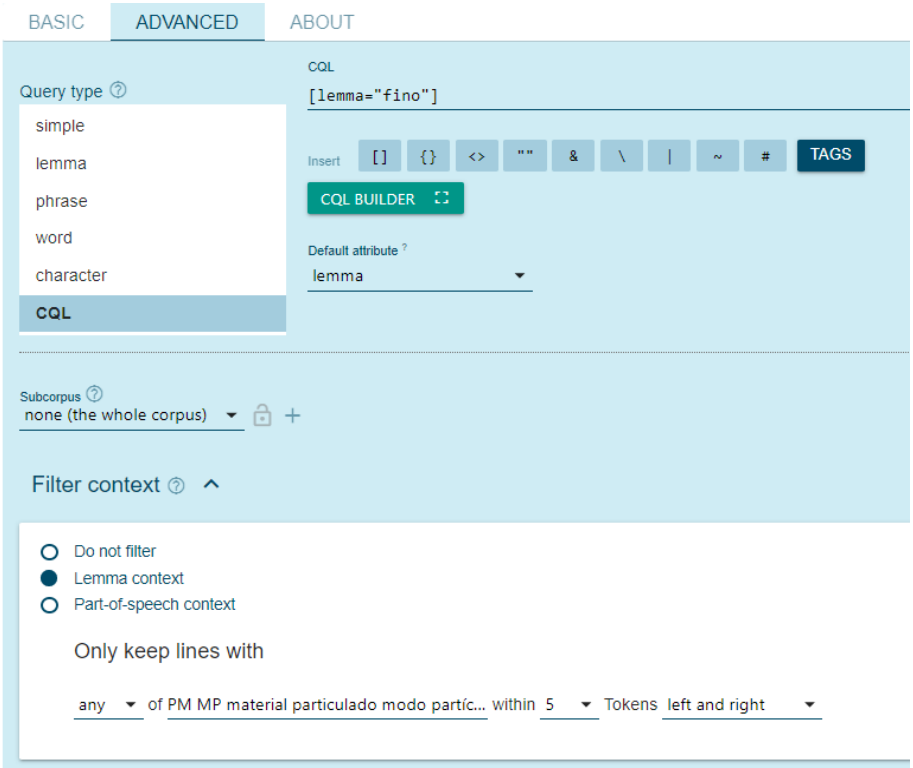

Figure 12. CQL query which elicited the lemma *fino* filtered by context

As a result, we were able to perform the query in Figure 12, which elicited the lemma *fino* found only in lines containing any of the following lemmas *PM*, *MP*, *material*,

particulado, modo, partícula, aerosol within a span of 5 words, either on the left or on the right. In the figure below, some examples of the retrieved concordances can be seen.

	Details	Left context	KWIC	Right context
1	doc#0	cuantas horas a unos pocos días. </s><s> La razón entre el **modo** de partículas	**finas**	y gruesas puede variar en función de la región de la atmósfera y de la estación d
2	doc#0	estándar del modo i. </s><s> Para el aerosol atmosférico, se definen los **modos**	**finos**	y gruesos (Whitby y Cantrell, 1976; Seinfeld y Pandis 1998; Dubovik et al., 2002
3	doc#0	kajima et al. (1996). </s><s> En el caso donde predomine el **modo** de partículas	**finas**	, el número de puntos para r > 0.5 µm para el método de King et al. (1978) es inc
4	doc#0	s, desviaciones estándar y concentración volumétrica de partículas para el **modo**	**fino**	de partículas con radios inferiores a 0.5 µm (ecuaciones 5.5, 5.12, 5.13 y 5.14). <
5	doc#0	el día como durante la noche hay un aporte importante en el **modo** de partículas	**finas**	. </s><s> Esto puede explicarse debido a los incendios detectados en Andalucía
6	doc#1	DA fine mode fraction Algoritmo de deconvolución espectral para fracción **modo**	**fino**	. </s><s> máximos espesores ópticos del aerosol en la longitud de onda de 500
7	doc#1	ura 4, con diferencias significativas entre los tamaños de las partículas en **modo**	**fino**	, que se asocia con el tiempo de vida de las partículas, y los procesos de conver
8	doc#1	1.4 a 2.1, como resultado de diferencias en el tamaño de las partículas de **modo**	**fino**	e índices de refracción. </s><s> La caracterización del aerosol es una primera ap
9	doc#1	primera aproximación de la distribución del aerosol en la atmósfera, con **modos**	**finos**	del aerosol, generado por quema de biomasa cercana a los sitios de medición. <
10	doc#1	erial mineral insoluble, sales marina y **material** biológico, en contraste la fracción	**fina**	y ultrafina está principalmente constituida por agregados carbonáceos con metal
11	doc#1	or de 0.5 indica que el 50% del PM10 está constituido por partículas en el **modo**	**fino**	, PM1). </s><s> La Figura 4.4 muestra la evolución del cociente PM1/PM10 para
12	doc#1	Particulado (PM): también llamado **material** particulado atmosférico o partículas	**finas**	. </s><s> Son pequeñas partículas de sólido o líquido suspendidas en un gas. </
13	doc#3	DA fine mode fraction Algoritmo de deconvolución espectral para fracción **modo**	**fino**	. 152 Cy aí gol once T iencias del Agua, vol. V, núm. 1, enero-febrero de 2014 Cc
14	doc#3	ura 4, con diferencias significativas entre los tamaños de las partículas en **modo**	**fino**	, que se asocia con el tiempo de vida de las partículas, y los procesos de conver
15	doc#3	1.4 a 2.1, como resultado de diferencias en el tamaño de las partículas de **modo**	**fino**	e índices de refracción. </s><s> La caracterización del aerosol es una primera ap
16	doc#3	erosol permitió identificar la distribución del aerosol en la atmósfera, con **modos**	**finos**	del aerosol, asociado alto contenido de vapor de agua y con baja absorción con
17	doc#5	el material **particulado** . </s><s> Dentro del material particulado, la fracción más	**fina**	ha sido asociada a eventos de mortalidad y morbilidad en la población (Dockery
18	doc#5	lesafíos vendrán con la entrada en vigencia de la Norma de **Material** Particulado	**Fino**	(MP2,5) a partir del 2012 (D.S Nº 12/2001). </s><s> Si bien, solo a contar de tal
19	doc#5	rmación del MP2.5 y determinar la composición química del **material** particulado	**fino**	en la Región Metropolitana, de manera de apoyar el seguimiento de la evolución
20	doc#5	yar el seguimiento de la evolución de la calidad del aire por **material** particulado	**fino**	y evaluar los impactos de la implementación del PPDA. 9 2 ACTIVIDADES REAL

Figure 13. Concordance results for the query [lemma="fino"] filtered by context (see Figure 12)

Each query was ultimately tailored to each specific CN, with some requiring more filtering of the context than others.

3.5.3 Extraction of Spanish variants within the web and other corpora

While the search for Spanish equivalents in the comparable corpus proved productive in most instances, its limited size capped the number of retrieved variants. As a consequence, we turned to the web for other Spanish original variants. We believe the web to be a useful resource for translators, one they often rely on for documentation. However, the results need to be filtered, namely in specialized translation. Therefore, we only used Google Scholar, which is limited to research works (<https://scholar.google.com/>). This was more time-consuming because as opposed to Sketch Engine, it did not allow searches using CQL queries. However, the occurrences there further confirmed or dismantled what had been observed in the comparable corpus.

Finally, we widened the scope of materials used to include other parallel corpora, which would give us some insights into variation in translation contexts. More specifically, we used the EurLex English-Spanish corpus (Vaisa et al. 2016)—available in Sketch Engine—and Linguee, an online bilingual concordancer frequently used among language learners and translation trainees.

The queries performed in the first one shared the same features as the ones described in Section 3.5.2., whereas searching for equivalents in Linguee was similar to doing so in Google Scholar—most variants needed to be looked up several times to take into account Spanish number and gender inflection.

3.6 Evaluation of machine translation output

For the purpose of carrying out a nuanced analysis of machine translation output of English-to-Spanish CNs, we used four different MT engines to perform the translation of a sample of 30 CNs, which included rule-based (Apertium), statistical-neural (Google, Systran) and neural (DeepL) systems. These CNs, which featured different morphosyntactic structures (N+N+N+N, Adj+N+N+N, N+N+Adj+N, etc.) were then organized in three 10-term groups, based on the number of constituents: (1) three-word, (2) four-word, and (3) five-, six-, seven-word CNs. This would allow us to study and compare to what extent each of the different approaches succeeded in translating English-to-Spanish CNs of different length and structure, and where their main limitations lied.

With the rise of MT in recent years came the need to develop metrics for the evaluation of the output quality. Different initiatives have been undertaken to design metrics for the evaluation of MT output, namely the Multidimensional Quality Metrics (MQM) framework, developed as part of the EU-funded QTLaunchPad project (Lommel et al., 2014), and the Dynamic Quality Framework (DQF), developed by TAUS (Görög, 2014). Both standards have now harmonized their error typologies as part of the EU-funded QT21 project, covering an extensive typology of issues, as well as a scoring mechanism to quantify translation quality. While the set of issues comprises 11 general categories (*accuracy, compatibility, design, fluency, internationalization, locale convention, style, terminology, verity, compatibility,* and *other*), annotators are encouraged to devise their metrics according to (1) the type of text being evaluated, and (2) the purpose of the evaluation, selecting issue types which are granular enough to address the research questions posed and restricted enough to be kept in mind by annotators (Burchardt & Lommel, 2014).

With this in mind, we adapted the MQM-DQF harmonized error typology specifically for the purposes of our study. To begin with, while the above-mentioned metrics are designed to assess translation errors in texts (i.e. multiple segments), our study required metrics which were term-oriented. This automatically discarded the inclusion of categories such as *design* and *compatibility*. Secondly, because the evaluated CNs are used almost exclusively in scientific and technical discourse, matters of *style* or *locale convention* were not the focus either. Instead, we aimed to measure the performance of MT systems based on their ability to respond to the challenges which CNs such as *aerosol single-scattering albedometer* pose. For this reason, the parameters used for the evaluation of MT output were three-fold:

(1) Accurate provision of translation equivalents for each of the constituents

To score the first point out of the three, MT output was expected to provide an accurate translation for each constituent in the CN. This implies a complete lack of (i) omissions; (ii) untranslated content; (iii) mistranslations and (iv) additions.

The accuracy of the equivalents was assessed in terms of the domain expectations, i.e. the point was scored provided that the chosen term was not contrary to the domain expectations with regard to terminology.

(2) Accurate identification of the bracketing groupings

Secondly, the output could score another point if the internal dependencies within the CNs were accurately identified. As discussed throughout this work, this aspect is crucial in the translation of these multi-word units, since inaccuracies in the bracketing can significantly alter the meaning of the term—even if each constituent has been adequately rendered in the TL, e.g. *aggregate anthropogenic carbon dioxide equivalent emissions* > **emisiones equivalentes al dióxido con carbono antropogénico agregado* (*emisiones antropógenas agregadas expresadas en el dióxido de carbono equivalente*).

(3) Fluent translation of the CN as a whole

Conversely, fluency was measured looking only at the target term and is based on criteria such as grammar, idiomaticity, and style. In other words, these issues had to do with the linguistic "well-formedness" of the term, assessed without regard to whether it was a translation or not.

The output (candidate or hypotheses term) of each MT system was measured by comparison with human translations of the input, which served as a baseline. When no documented translations could be found—notably in longer terms of recent creation such as *in vitro inhalation bioaccessibility procedure*—the source term was used to determine how much of the content had been rendered into the target language (accuracy), in combination with our own observations when analysing the data as native speakers of Spanish (fluency).

Whereas the MQM-DQF metrics scored translation output according to the number and severity of the errors, we envisioned an *ad hoc* point system which measured output based "on their hits instead of their misses". In other words, we assigned a point for each of the three criteria the translation output fulfilled, so as to measure the degree to which the translated term meets quality requirements, with an emphasis in accuracy and fluency. The criteria (further dissected in Section 4.4) were clear-cut, and the score absolute, with the aim to ensure the transparency of the evaluation and tackle the subjectivity inherent to the human translation of MT.

4. Results and discussion

As part of this research, we studied the translation of a sample of CNs in the field of air pollution and air quality treatment through the analysis of term variants both in the source language and in translation contexts. This was followed by the discussion resulting from the evaluation of the MT output of these units. Based on the results obtained throughout the different analyses, we proposed a protocol for the translation of CNs into Spanish.

4.1 Analysis of source term variation

While the focus of our study lied on the translation of CNs, prior examination of term variation in the source language was key to understand its origin. Though most of the categories described in Faber and León Araúz (2016: 12-13) occurred, we will now delve into the ones that proved most significant in the studied sample of CNs.

The use of orthographic variants, such as *air-quality management* and *air quality management*, was present in most CNs. However, far from being influenced by geographic origin, these variants appeared to be used by authors in an attempt to make the internal dependencies clearer. This can also be appreciated in variants such as *coarse mode sea salt aerosol*, which coexists which *coarse mode sea-salt aerosol*. As pointed out in Nakov and Hearst (2005), hyphens serve to disclose the underlying bracketing mechanism, as do acronyms.

Indeed, acronyms constituted another primary source of variation, especially due to their high level of instability. Short form variants sometimes encompassed the whole CN (e.g. *two-step laser mass spectrometry > L2MS*), whereas in other instances only part of the term was shortened, which is the case of *aged toluene SOA particles*. Far from being standardised, these internal acronyms seem to be motivated by the author's wish to clarify the bracketing of a given CN. The reason behind the instability of these short form variants (e.g. *direct radiative effect of aerosols > DRE of aerosols; DREA*) is none other than the fact that the priority of scientists is getting the message across, while the linguistic aspects are usually overlooked for the sake of synthesis.

Although the variant types identified so far were undoubtedly relevant in our analysis, cognitive variants deserve special mention for the challenges they pose, both from the point of view of translation and knowledge representation. The analysed cognitive variants were the product of multidimensionality, i.e. they expressed different dimensions of a single concept. As such, they had an effect on the semantics of the CN, which both translators and terminologists need to be aware of before choosing or proposing the respective target language equivalents. This can be illustrated through the example of *brown carbon*, which coexists with the variant *light-absorbing organic carbon*. When both are included in a given text, the translator may be tempted to translate both terms for the one they are familiar with, or the one included in their go-to terminographic resource.

However, as stated in previous sections, the decision of the author to use each of them is most likely a conscious one, in an effort to highlight a specific dimension of the concept for communicative purposes. In these cases, standardising would not be the right move, as it has a detrimental effect in communication.

The topic of air pollution and air quality treatment has been addressed by numerous disciplines. As such, the terminology was also particularly prone to domain-specific variation. While climatologists appear to label air pollutants based on their chemical composition (e.g. *mineral dust aerosol*), toxicologists make use of denominative variants which reflect the shape of these pollutants in terms of *coarse*, *fine* or *ultrafine* particles (e.g. *fine particle aerosol*). Meteorologists, on the other hand, measure particles based on their size, usually through the use of formulas (e.g. *particulate matter 2.5 micrometres or less in diameter* > $PM_{2.5}$). Nevertheless, these designations are often used as synonyms when they are not. This is the case of *black carbon, soot aerosol* and *elemental carbon*, which despite highlighting different dimensions of the concept, are used interchangeably in the literature. Representation of these concepts in TKBs and other terminographic resources needs to aspire to reflect these nuances, in order to prevent domain loss as well as potential translation errors derived from a lack of understanding of the concepts and their inherent dynamicity.

4.2 Analysis of target term variation

Multiple variants of the studied concepts also emerged in Spanish, some in original production (still influenced by the contact with the *lingua franca*, as the product of secondary term formation), others as a direct result of translation errors. We will now analyse these using the typology in León-Araúz and Cabezas-García (in press) as a framework, adapting it to the study at hand.

To begin with, omissions were present in most forms, ranging from the omission of articles (*distribución de tamaño de aerosol, distribución de tamaño del aerosol*) to the omission of formants, either of the head or the modifiers. Omission of the head often implied transposition, as seen in *aerosol de polvo desértico* > *polvo desértico* or *aerosol de sal marina* > *sal marina*. Furthermore, when omissions affected the modifiers, this sometimes implied the activation of hypernyms (e.g. *estaciones de muestreo de la calidad del aire* > *estaciones de muestreo*), though this was not always the case (e.g. *efecto radiativo directo de los aerosoles* > *efecto directo de los aerosoles*).

Notably, highly specific English CNs were hardly ever rendered fully in Spanish (e.g. *PUF disk passive air sampler* > *muestreador pasivo de espuma de poliuretano*). While English noun-packing helps string concepts together smoothly, Spanish is not so open to such long terms, especially when the general tendency is to link all the constituents with the preposition *de* (English: *of*): *muestreador pasivo* ***de*** *aire* ***de*** *disco* ***de*** *espuma* ***de*** *poliuretano*. It has been observed that, when the process of secondary term formation is not given the attention it deserves, sloppy translations (such as the one above) emerge.

As a result, experts tend to omit the expendable constituents, sacrificing accuracy for the sake of fluency, an issue which is later addressed in Section 4.4. This is also a reflection of the already discussed phenomenon of domain loss, where the same experts who work with the entities these terms designate can only name them in the *lingua franca*, often through the use of the acronym. So much so that it leads to another type of variation which was repeatedly observed—the use of hybrid-term forms. Indeed, half-native/half-borrowing variants such as *mapas TOMS de índice de aerosol* (where TOMS stands for Total Ozone Mapping Spectrometer) emerged frequently.

Noun packing in the source language usually called for expansions in the target language. Terms such as *aerosol volcánico* coexisted with forms such as *aerosol de origen volcánico*, thereby making the agent explicit. In other instances, the conceptual information which needed explicitation was the patient, e.g. *contaminación por partículas > contaminación atmosférica por partículas*. The high number of variants where information was made explicit continues to suggest how overusing the preposition *de* does not comply with the internal rules of Spanish, a language known to be highly inflectional. Instead, the more fluent, target-language oriented variants rejected the use of adjectives (influenced by the noun packing in the *lingua franca*) and substituted these for explicitations such as the one that follows: *efecto radiativo directo > efecto directo sobre el balance radiativo terrestre*. Incidentally, permutations in the prepositions linking the constituents of CNs—especially in the expression of cause—were especially frequent, ranging from less to more specific ones (e.g. *contaminación de material particulado, contaminación con material particulado, contaminación por material particulado*). In other instances, permutations affected not only prepositions but the entire CN, as seen in *columna total de ozono > ozono total en columna*, where both were widespread to a similar extent.

Transpositions of adjectives by "of + noun" (e.g. *techo atmosférico > techo de la atmósfera*) and by periphrasis (e.g. *aerosoles por quema de biomasa > aerosoles generados/producidos/emitidos por la quema de biomasa*) also occurred. These two structures appeared to be among the preferred term formation mechanisms in original production contexts, which is later taken into account in the protocol (Section 4.4).

Structural shifts were also central to our analysis, whether they affected the head or the modifiers. Nouns were either shifted by a synonym (e.g. *estaciones de monitorización/vigilancia/control/monitoreo de la calidad del aire*), or by modulation, namely by near-synonym (*espesor óptico de aerosol > profundidad óptica de aerosol*) or by metonym (*aerosol de carbono elemental > partículas de carbono elemental*). Whereas, in some cases, the shifts merely affected the modifying adjectives without altering the communicative situation (*e.g. efecto radiativo directo > efecto radiativo instantáneo*; *aerosol marítimo/marino/oceánico/del mar*), in others these were actually a reflection of the multidimensionality in the domain. In other words, shifts by modulation attempted to highlight a specific dimension of the concept, such as in *aerosol de polvo mineral* (composition) and *aerosol de origen desértico* or *aerosol sahariano* (location).

At different points throughout this work we have circled back to the importance of bracketing. Coincidentally, due to errors in the internal dependence analysis (or the lack thereof) several examples of inaccuracies were found, such as *liberación de calor por fuentes antropogénicas* > **liberación de calor antropogénico*, where the permutation of the modifier changes the meaning. Similarly, other inaccuracies occurred as a consequence of failing to pinpoint the semantic relation between the formants (e.g. *espesor óptico del aerosol* > **espesor óptico por aerosoles*, where patient and cause were confused. Far from being intentional, these variants constitute translation errors, derived from an insufficient analysis of the structure and semantics of the CN at hand.

This stresses the need to develop a series of guidelines which, in a protocol-like manner, lead translators in the translation of these units, emphasizing the need to not only take the necessary steps towards the full understanding of the concept (structural and semantic decoding), but also produce equivalents which meet the expectations of the target reader in terms of both accuracy and fluency.

4.3 Analysis of machine translation output

Our study next looked at the translation of CNs in the field at hand through the lens of MT software. According to the set of criteria defined in Section 3.6, each output was scored 1-3, where 1 was the lowest and 3 was the highest score. Accuracy, also referred to as 'adequacy' by some scholars (e.g. White & O'Connell, 1994) was measured in terms of how much of the meaning expressed in the source term was rendered in the translation. The decision to assign two thirds of the score to accuracy was deliberate, as accurate translations are what ultimately guarantee expert-to-expert communication in specialized discourse.

The first parameter assessed whether the MT engine succeeded in providing a domain-oriented equivalent for each of the constituents. Both Google Translate and DeepL scored this first point in 77% of the cases (in 10/10 of 3-word CNs, 8/10 in 4-word CNs, and 5/10 of the remaining group of 5-, 6- and 7-word CNs). Systran only met this criterion in 60% of the cases (7/10, 8/10, and 3/10 in each group, respectively), whereas Apertium's rate dropped to 26% overall (5/10, 3/10, and 0/0).

The main issue found in all four engines was the presence of untranslated content in the CNs, mostly in the form of acronyms. Output included translations such as **aerosol antropogénico ERF* for *anthropogenic aerosol ERF* (*FRE de aerosoles antropogénicos*). These generally failed to translate acronyms, regardless of their position within the CN (initial, central, final) or their length. However, both Google and DeepL did accurately identify some exceptions, such as *COVNM* as the Spanish equivalent for *NMVOCs* in *industrial NMVOCs emission*, which set them apart from the other two.

While no omissions of content occurred, Apertium's output left most constituents untranslated, sometimes even rendering them in another language, e.g. *portable light-scattering aerosol monitor* > **portátil monitor de aerosol que esparci ligero.* In other

instances, the Spanish equivalents provided were simply not accurate because they were not consistent with the domain's expectations (e.g. **modo tosco* for *coarse mode*, instead of *modo grueso*). Systran also incurred in mistranslations, such as **efecto radiactivo de albedo superficial* for *surface albedo radiative effect*, thereby significantly changing the intended meaning.

With regards to the second parameter, Google Translate and DeepL stood out once again, accurately identifying the bracketing groupings in 67% and 63% of the CNs, respectively. Both scored a point in 8/10 of 3-word CNs and 9/10 of 4-word CNs, with only a slight difference in CNs of 5 or more constituents (3/10 and 2/10, separately). Systran only identified the dependencies in about half of the cases (53%), in 6, 7 and 3 CNs out of the 10 in each group, respectively. Apertium, on its part, barely met the criterion in 33% of the cases, with no CNs of over 5-constituents accurately bracketed, and only 5 out of 10 in each of the other two groups.

The underlying morphosyntactic structures of the CNs also had an influence in the performance of MT software. In particular, where an adjective preceded the noun on the rightmost part of the term (e.g. N+N+Adj+N), internal dependencies were accurately pinpointed, such as in the abovementioned example of *surface albedo radiative effect* (*efecto radiativo de albedo de superficie*). This can be associated with the fact that, statistically, such an adjective is quite likely to modify the noun that follows. Conversely, adjectives in initial position often led to inaccurate bracketing groupings (e.g.**emisión de COVNM industriales* for *industrial NMVOCs emission*), as did participles. This can be further illustrated through the example of **partícula SOA de toluene envejecido*. The internal dependencies (*aged [toluene SOA] particle*) were not accurately identified, which also points directly to the failure of MT software to access the semantic relations within the constituents (e.g. SOA *made_of* toluene; particle *has_attribute* aged).

While the bracketing issues commented so far had a significant impact in translation quality, the most glaring errors were found in CNs of more than 5 constituents. An example of this is the translation of *matrix-assisted laser desorption/ionization time-of-flight mass spectrometry* (Figure 14), where none of the MT engines succeeded in identifying bracketing groupings. As a result, the output—despite having its constituents accurately rendered in Spanish—failed to get the source content across, widely differing from reference human translations, such as *espectrometría de masas por tiempo de vuelo con desorción/ionización asistida por una matriz*.

Apertium	Google Translate
Matricial-Ionización de Desorción de Láser/Asistida Tiempo-de-Espectrometría de Masa del Vuelo	Espectrometría de masa de tiempo de vuelo / desorción láser asistida por matriz
DeepL	Systran

Desorción e ionización láser asistida por matriz Espectrometría de masas en tiempo de vuelo	Espectrometría de masa de tiempo de vuelo asistida por matriz/ionización Tiempo de espectrometría de masa de vuelo

Figure 14. English-to-Spanish MT output for *matrix-assisted laser desorption/ionization mass spectrometry*

On the other hand, major errors in fluency were often the product of a chain reaction—a poor translation of each of the constituents, followed by inaccuracies in the bracketing, led to largelyunintelligible translations, such as **masa de aire respalda trajectory* for *air mass back trajectory*. Nonetheless, in a number of instances where the translation of each constituent was broadly accurate and the bracketing groupings were adequately identified, certain fluency issues still surfaced, thus preventing the output from scoring that third and final point. The identified issues had to do with matters of grammar, idiomaticity, or style.

In absolute numbers, the performance of each MT engine can be seen in the following graph (Figure 15).

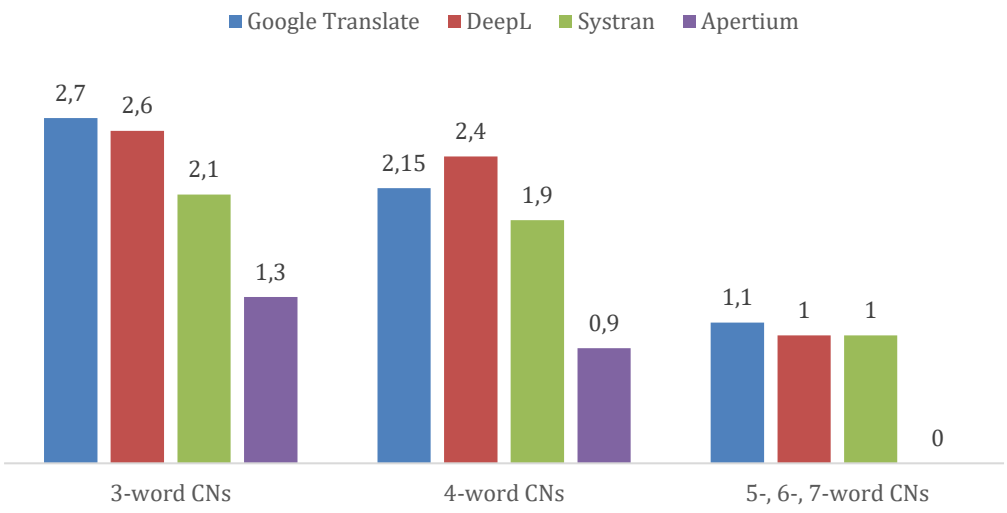

Figure 15. Evaluation of the four MT engines according to the established parameters

However, if we take into account the high expectations of the readers of specialized translation, as well as the fact that a translation can be considered fluent despite not being accurate, MT engines should ideally be assessed as to whether they meet or not all three criteria (Figure 16). This was seldom the case, except for the more widespread CNs[3], such

[3] The sample of terms analyzed included some which we deemed more complex to decode than others, with a view to test the limits of the different engines.

as *organic carbon aerosol*, which was accurately and fluently translated by all four MT systems.

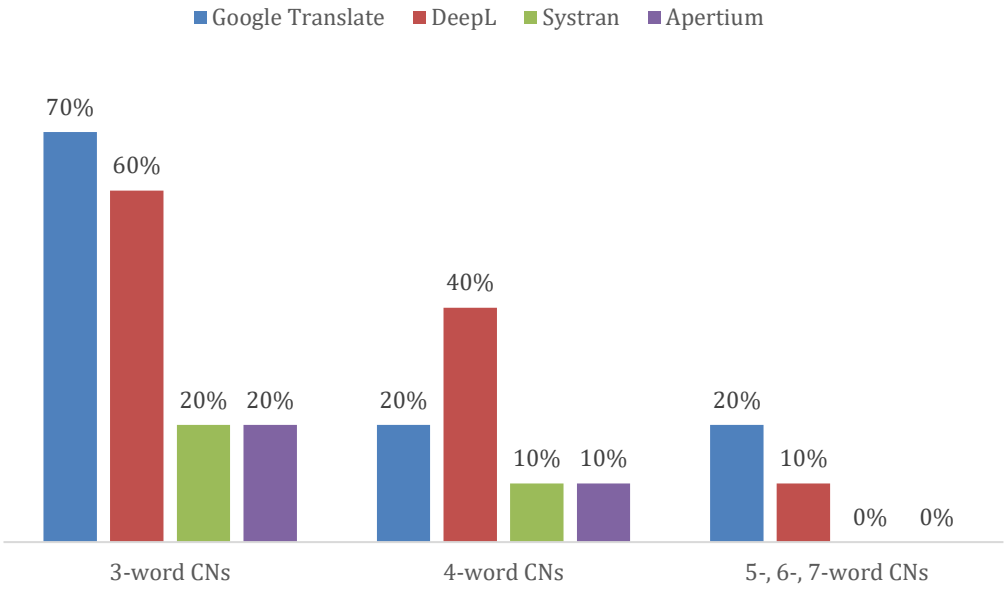

Figure 16. Rate of success in scoring all 3 points, i.e. being fully accurate and fluent, by group

From a more functional, user-oriented perspective, the results evidence some initial points:

(1) Both Google Translate and DeepL stood out in the translation of 3-word CNs, which suggests that they can be of use in the translation of these units, provided that they are followed by post-editing (errors are still expected to occur in 30-40% of cases, either in terms of accuracy or fluency).

(2) While DeepL appears to be a step ahead of the rest of the engines, its performance in the translation of 4-word CNs is still lacking. As a result, using machine software for the translation of these longer CNs is hardly an option for the time being. Similarly, the translations of CNs of 5-, 6- or 7-word constituents provided by the machine translation systems did not meet minimum quality requirements either, and consequently appear to be of little use in specialized translation.

(3) The performance of MT systems was observed to be influenced by the complexity of the CNs, with regards to not only length (the longer the CN is, the higher the number of issues that appear to arise), but also structure. Incidentally, having adjectives precede the noun head appears to point these engines to the correct bracketing. Along these lines, when CNs feature a right bracketing, neural and statistical engines seem to be more likely to accurately identify the head (since pre-modification is known to be the most common term formation process in English), even if issues still arise when disambiguating the relations between the modifiers.

4.4. Towards a protocol for translating CNs of more than 3 constituents

As evidenced by the analysis carried out in previous sections, the translation of CNs poses a major challenge even for human translators. While the day may come when MT engines overcome such a challenge smoothly, the results in Section 4.3 reveal that these still have a long way to go—hence the need to develop a protocol which guides human translators in the translation of these multi-word units.

In the process of translating a CN into another language, the translator encounters a myriad of issues, ranging from the search for documentation—essential for the understanding of these terms—to the production of term equivalents which meet the domain's expectations. For this reason, we have envisioned the following 5-step protocol, which addresses the difficulties associated with the translation of CNs of 3 or more constituents:

(1) Identifying the bracketing groupings by means of a corpus
(2) Verifying the semantic relations holding the constituents together
(3) Looking up the constituents in terminographic resources[4]
(4) Querying comparable or parallel corpora to see domain terminology in use
(5) Producing target language oriented equivalents of the CN

Naturally, when translators come across a CN whose Spanish equivalent escapes their knowledge, their first instinct is to look it up in their go-to multilingual term base. While this search is at times unproductive, this does not mean that these resources should be discarded—they can still be of use in later steps of the translation process.

Though post-editing continues to gain ground in the industry, proficiency in the target language is not enough to assess MT output (see Section 4.3). The complexity of these multi-word units requires prior disambiguation of their morphosyntactic structure and decoding of their semantic content. Only then does the linguist have the knowledge to question the output in full detail and make any necessary changes to improve such results. In other words, the translation process of highly specialized CNs as the ones in our sample must begin by understanding the unit in full. To this end, the translator should take the following two steps:

(1) Identifying the bracketing groupings by means of a corpus

Internal dependencies are hardly ever transparent in CNs of more than 3 constituents. For this reason, the process of rendering them into another language needs to start by disambiguating the syntactic structure. Otherwise, translators could find themselves looking up the wrong concepts in the terminographic resources. This also paves the way for the subsequent decoding of the semantic relations within the CN.

[4] This step can be complemented with the use of MT engines (see following paragraphs).

Ideally, the bracketing would be done by means of corpus. By making use of corpus query engines such as Sketch Engine, which gives user the option to perform lemmatized searches and CQL queries, the translator could save some precious time in the translation process.

For instance, in the scenario of a specific translation commission, the translator could easily devote some time to the compilation of his/her own personal comparable corpus from the references cited in the scientific article they have been asked to translate. Using queries such as the ones exemplified in Section 3.2 would guide him/her in the analysis of the CNs at hand.

However, not all translators have access to the same resources or the same training. If they cannot get hold of a collection of comparable texts or a subscription to a corpus query engine, they can still make use of the web as a corpus. When duly filtered, e.g. using only Google Scholar, web queries can still give some valuable insights into the internal dependencies or semantic relations within the CN.

The already discussed indicators (cf. Cabezas-García 2019) have proven to be useful in this first step. Translators, as linguists, should ultimately follow their instincts, and confirm the compliance of as many of these indicators as they deem necessary to confirm their suspicions.

(2) Verifying the semantic relations holding the constituents together

Where the structural disambiguation is performed smoothly, this next step serves to confirm the suspected morphosyntactic structure and provide clarification of the semantics of the CN—a basis necessary for the production of target language equivalents. Conversely, where issues are encountered in the bracketing process (step 1), the semantic analysis helps elucidate the dependencies. These two steps are, thus, complementary in nature.

While sometimes it is the same constituents in the CN which shed light on the semantic relations (e.g. *dust-dominated aerosol*), in most cases the translator has to perform specific corpus queries to make these explicit, namely by means of paraphrases. As explained in Section 3, the more highly specialized a CN is, the harder the search for verb paraphrases is, which is where free paraphrases, in conjunction with KPs, come into play (see Section 3.2 and Section 3.3).

Again, if the translator's resources are limited to the web, he/she can still make use of the advanced search in Google Scholar (e.g. *"aerosol * desert"*; *dust "known as" *aerosol*") which simply goes to stress the importance of fully grasping the semantics of a CN before attempting to produce a target language equivalent. In other words, failing to have access to more sophisticated tools should not be an excuse for skipping this step—what a translator lacks in resources, he/she has to make up for in resourcefulness, especially when translating such cognitively complex units.

Once full, nuanced understanding of both the structure and the semantics of the CN has been achieved, the translator can begin his/her quest to render it in the target language:

(3) Looking up the constituents in terminographic resources

As explored in previous sections, the data found in terminographic resources, no matter how little, guides the search for translation equivalents (e.g. if the translator has no background knowledge of the meaning of *depletion*, he/she can hardly make corpus queries which search for Spanish variants of *ozone layer depletion*).

At this point, the translator is expected to have a rough idea of what form the term is going to adopt in the target language. As a consequence, he/she could choose to complement their documentation process with the aid of this software. However, having MT engines perform automatic translations of the CNs should always be done with a view to postediting, as this can by no means substitute the use of terminographic resources, duly designed and maintained by linguists. Furthermore, the translator should resist the temptation to skip the first to steps and turn straight to MT, given not only the poor performance documented in our analysis (Section 4.3), but also the fact that—due to the complexity of these units—it is only after analyzing the structure and semantics of the CN that a translator is in a position to assess the adequacy and fluency of the output.

(4) Querying comparable or parallel corpora to see domain terminology in use

The next logical step towards the production of target language equivalents of the CNs is to find potential variants in texts originally written in the TL, i.e. in Spanish. While parallel corpora have been documented to show more translation variants than parallel corpora (Sanz-Vicente, 2011; Miyata & Kageura, 2016; León-Araúz & Cabezas-García [in press]), comparable corpora are undoubtedly of use as well, where queries such as the one introduced in Section 3.4 can guide the production of equivalents, in order to make sure the proposed Spanish CNs meet the domain's expectations.[5]

(5) Producing target language oriented equivalents of the CN

Based on our observations while analyzing the data, the production of Spanish equivalents for the sample of CNs in the field of air pollution and air quality treatment lacked fluency, i.e. they were not target language oriented. While accuracy—due to a severe lack of understanding of the internal dependencies and semantic relations within the CNs—was also an issue at times, this should not be a problem provided

[5] An open-access alternative can be found in the use of bilingual concordancers, such as Linguee.

that the steps outlined so far are followed. Even if the equivalents provided by the multilingual databases contained errors, these could easily be spotted in the results of the comparable corpora.

However, it is fluency (paired with accuracy) what will ultimately differentiate translators from MT software. To achieve this goal, we will now proceed to describe a series of guidelines—derived from our analysis of term variation—which we believe need to be taken into account when translating CNs in the English-Spanish language pair.

To begin with, English noun-packing is not well-received in Spanish. That is to say, equivalents in Spanish do NOT need to feature the same morphosyntactic structure. While unexperienced translators, especially non-native speakers of Spanish, may have certain reservations about producing functional equivalents of these terms, translation quality ultimately depends on it. In our analysis of target term variants (Section 4.2) we observed several trends, such as the following:

(a) If the CN strings three or more nouns together, the constituents should not be linked using the preposition *de*. The existing pattern in Spanish is to offer more nuanced linkers between the constituents.
(b) The expression of cause should therefore be made explicit either using the preposition *por* or a more specific periphrastic structure, depending on the CN, e.g. **aerosol de quema de biomasa* > aerosol *por* quema de biomasa, aerosol *generado por* combustion de biomasa, aerosol *emitido por* quema de biomasa.
(c) Location is not only expressed by means of adjectives. While that is the exact tendency in English, Spanish tends to express location in more elaborated structures, a consequence of post-modification being the preferred term formation structure in this language. An example of this can be found in *indoor air pollution*, where *contaminación del aire interior* coexists with a myriad of variants, such as *contaminación del aire en especios cerrados, contaminación del aire en las viviendas, contaminación del aire en ambientes cerrados* or *contaminación del aire reinante en el interior del hogar*.

These multi-word units are usually neologisms, created by the need to designate latest innovations in the ever-changing field of air pollution. The translator, as a linguist, is in a privileged position to render those same neologisms in Spanish, instead of leaving untranslated content within the CN. As the identification of these concepts in the international scientific community is also key for Spanish-speaking experts, the recommended translation strategy ensures both identification and comprehension of these units. For instance, instead of using *AOD* as the acronym for *profundidad óptica de aerosoles*, the translation solution would include both in the first mention: *POA (AOD, en inglés)*. In doing so, the translator would be combatting the phenomenon of domain loss, ensuring that neither accuracy nor fluency take a hit.

Based on our observations, we believe term variation to be a key element of fluency in the Spanish language, especially for discursive and cognitive reasons. On

the one hand, repetition is an indicator of poor written expression in Spanish, whereas in English it is not. As such, discursive variants will enrich scientific discourse, provided that they are the product of informed decisions. For instance, structural shifts by hypernyms, as well as metonyms (e.g *contaminación del aire ambiente urbano por partículas ultrafinas* > *contaminación urbana*), seem to be among the most frequent types of variation in this language pair. Moreover, translators are encouraged to embrace multidimensionality also in the target language, using the dynamicity of these concepts to their advantage, i.e. to reflect the author's term choices in a fluent, nuanced way, further meeting the domain's expectations.

4.4.1 Case study: particle number size distribution

We will now illustrate the use of the above-featured protocol in the English-to-Spanish translation process of a CN in the subdomain of air pollution and air quality treatment: *particle number size distribution*.

First and foremost, the lack of transparency of this unit calls for the disambiguation of the syntactic structure (N+N+N+N). The following five possible combinations were considered:

[particle number] [size distribution]
particle [number size] distribution
[particle number size] distribution
particle [number size distribution]
particle [number] size distribution

We proceeded to check and compare the frequency of the bracketing groupings within the CN according to the adjacency, dependency and shortening models, making use of CQL queries similar to the one below.

[tag!="JJ.*|N.*|RB.*|VVG.*|VVN.*"][lemma="particle"][lemma="size"]
[lemma="distribution"][tag!="N.*|JJ.*"]

The results of the queries shed some light on the internal dependencies within the CN, and helped us discard some of the initially considered combinations (e.g. no results were found for the adjacent grouping of *number size* as an independent term in the corpus). The high number of occurrences of *particle size distribution*, together with our observations while compiling the data, led us to suspect that the correct bracketing was either *particle [number] size distribution* or *particle [number size distribution]*. Nonetheless, we still needed to seek further clarification, as not only the structure but also the semantics of the CN at hand were significantly obscure.

Particle number size distribution	Freq.
size distribution	150
particle size	66
particle size distribution	43
particle number	41
particle distribution	8
number distribution	6
number size distribution	3
particle number size	1
number size	0
particle number distribution	0

Figure 17. Frequencies of possible bracketing groupings in *particle number size distribution*

Therefore, we moved on to perform a series of additional queries which would help us decode the CN by means of verb and free paraphrases, such as the following:

([lemma="size"][]{0,10}[tag="V.*"][]{0,10}[lemma="distribution"]within <s/>)
|([lemma="distribution"][]{0,10}[tag="V.*"][]{0,10}[lemma="size"]within <s/>)

([lemma="distribution"][]{0,10}[tag="V.*"][]{0,10}[lemma="number"]
[lemma="size"] within <s/>)|
([lemma="number"][lemma="size"][]{0,10}[tag="V.*"][]{0,10}[lemma="distribution"]within <s/>)

([lemma="distribution"][lemma!="size"][]{0,10}[lemma!="distribution"]
[lemma="size"] within <s/>)
|([lemma!="distribution"][lemma="size"][]{0,10}[lemma="distribution"][lemma!="size"] within <s/>)

To begin with, term variants such as *size distribution of particles* (transposition by "of + noun") appeared to confirm the correct bracketing grouping being *particle [number] size distribution*.

and by giving some examples of particle measurements illustrative of specific phenomena in the atmosphere. 2. **Size distribution of particles** in the atmosphere There are three distinct modes into which airborne particles can typically be divided.

Concordances such as the ones below also guided us in the identification of constituents omitted in the CN, the full CN being *aerosol particle number size distribution* (*aerosol* made_of *particles*), and *number* actually referring to *number concentration* or *number density*, which undoubtedly shed some light on the semantics of the CN:

In general, it can be seen that average **aerosol number size distribution** has a successive distribution for smaller particles (<0.3 μm) and a sharp decrease at 0.3 μm, since burning emissions can elevate **number concentrations of aerosols** at 0.11

in air pollution and climate change. We performed a systematic airport study to characterize real-time size and **number density distribution** , chemical composition and morphology of the aerosols (∼10nm–10μm) using complementary cutting-edge and novel

We apply a four-modal log-normal size distribution to fit **aerosol size distribution**. • Meteorological conditions are found to exert a major influence on the **aerosol distributions**

number size distribution of ED cases at different altitudes and different seasons. Total **number concentrations of particles** of the four modes and the parameters that characterized the **number size distributions** at different altitudes of three seasons are listed in Table 3. Note that the three altitude ranges are 0–

Other concordances helped elucidate the semantic relations between the rest of the constituents (*particle* has_attribute *size*), while also pointing us to context lemmas which would be of use in the corpus-based search for equivalents (e.g. *diameter*):

in abundance at ca. 1{2 m m, there is a subsequent growth in particle abundance (in terms of mass but not number) for **particles which extend in size up to** ca. 100 m m, although above 10 m m **diameter** their atmospheric lifetime becomes rather short. These coarse mode

airborne particles by ltration and weighing the lter before and after particle collection. In order to **restrict the particles to a given size range**, such as PM10, PM2:5 or PM1:0, size- selective inlets are available which restrict the particles allowed access

a minor (if any) impact on CCN activation kinetics. Differentiating and normalizing Eq. (4) gives the probability distribution of κ, p(κ), **for particles of constant size** from which one can compute the variance of the distribution function

The suspected bracketing also complied with the indicator of finding the bracketing groupings in combination with other elements, as can be seen below:

mode (<2.1 μm) but increased in the coarse mode between 2.1 and 4.7 μm during the dissipation stage. This difference in **particle NH4+ size distribution** was likely caused by the enhanced RH that facilitated the hygroscopic growth of coarse particles (>2.1 μm), causing

by Schripp et al. (2013) showed that exhaled e-cig aerosols from a single e-cig user have a real-time **bimodal particle size distribution** at 30 and 100nm, compared to a bimodal particle size distribution that varied at 11–25 and 96–175nm for mainstream e-cig particles (Mikheev et al., 2016).

Coincidentally, concordances resulting from queries such as the exemplified in Section 3.3 seemed to reveal the what the semantic relation holding linking *number* to the rest of the constituents was: *particle size distribution* weighted_by *number* (more specifically, *number density*).

m) Figure 1. A measured **particle size distribution from suburban Birmingham, weighted by (a) number**, (b) surface area, and (c) volume.

and a moderately elevated mass concentration at the Marylebone Road site, with the traf c increment in the **particle size distribution having a mode in the number-weighted distribution at ca. 20 nm diameter**, well below that recorded using conventional dilution tunnel methods in many of the studies of

Though to grasp the full meaning of these concepts one would need a solid background in Mathematics, the performed corpus queries elicited concordances which helped us,

translators, access the semantic relations within the CN to the extent where we can render them in the target language with minimal risk of inaccuracies.

We subsequently made use of the two terminographic resources already introduced during our work to search for potential equivalents of the constituents of the CN. IATE provided an equivalent for one of the groupings within the CN, *particle size distribution* > *distribución del tamaño de partícula*, while TERMIUM Plus also suggested *granulometría* as equivalent for this 3-word combination.

Machine translation output of this CN, however, did not provide further clarification, as the constituents were all linked by the preposition *de* in an opaque way: **distribución del tamaño del número de partícula*.

Thus, using the data we had collected so far, we performed some final queries in the Spanish comparable corpus, as well as in the EurLex Spanish corpus, alternatively searching for the lemmas *distribución* and *granulometría*, making use of the context filter (including only lines which contained any of the following lemmas within 7 tokens, left or right: *número, numérica, tamaño, diámetro, partícula, aerosol*). We also checked Google Scholar for equivalents of the groupings by making using of operators, namely inverted commas.

distribución del tamaño de número
distribución de tamaño del número de partículas
distribución numérica de tamaños de aerosoles
distribución de numérica de tamaños de los aerosoles
distribución de tamaño numérica de aerosoles

While some of them showed inaccuracies (**distribución de numérica de tamaños de los aerosoles*), others did guide us in the production of the Spanish equivalent. Taking into account that it should not include the preposition *de* more than 3 times, and that we should detach our translation from the noun-packed source term, we finally produced the following variants:

distribución numérica del tamaño de los aerosoles
distribución del tamaño de los aerosoles por densidad numérica
distribución numérica del tamaño de las partículas de aerosol

As can be seen in this case study, there is no universal solution to decode all CNs and render them in the target language. Instead, our protocol presents a series of steps that translators are encouraged to follow in their quest to render these units in Spanish. It relies heavily on the translator's own needs, as each CN is a world of its own (depending on its semantic opacity, morphosyntactic structure, among other factors), and will ultimately require its own tailored documentation and production process.

5. Conclusions

The present study focused on the translation of CNs in the field of air pollution and air quality treatment. We explored how both human and machine translators approach the decoding of these units, the challenges they face, and where their main limitations lie.

The field of air pollution and air quality treatment was mainly explored by means of two comparable corpora (one in English, another in Spanish) of specialized texts within the hard sciences, as well as other terminographic resources and online corpora. Across the different domains (Physics, Chemistry, Engineering, etc.), multi-word units were found to be particularly prone to both denominative and cognitive variation. Indeed, multidimensionality proved to be one of the most significant causes underlying the proliferation of forms to designate a single concept. In order to produce quality translations of these CNs, translators therefore need to be aware of the different dimensions being highlighted in each case to ensure communication in the target language, Spanish, is both accurate and fluent. However, this was not always an easy task, considering the cognitive complexity of these units.

While previous research had focused on the study of shorter combinations, we set out to analyze CNs of up to 7 constituents. The lack of systematic representation of these long, highly specific multi-word terms, together with the difficulties posed by their structural and semantic ambiguity, was documented to result in inaccuracies and other issues in the translation variants that emerged.

Along these lines, though MT software has undoubtedly been advancing in leaps and bounds for the past five decades, it still suffers badly from incorrect translations of CNs. Our analysis of MT output of CNs within this field demonstrated that the automatic decoding of these units cannot yet be done successfully. Furthermore, the output not only failed to accurately disambiguate the internal dependencies of CNs of more than 3 constituents, but also to fluently render these units according to the internal rules of Spanish.

These findings prompted us to propose a protocol which translators can benefit from when translating English CNs into Spanish, one which offers guidelines for the different steps of the translation process—from text understanding to text production. To rise above the shortcomings of machine translation approaches, human translators are encouraged to first (1) analyze the CN in full, both structurally and semantically, as understanding the conceptual nuances is what will ultimately give them the freedom to produce target language oriented equivalents; next (2) find translations for the constituents of the CN in terminographic resources, if needed, and (3) make use of corpora to look for *in vivo* variants, and finally, (4) based on the data collected in the previous steps, produce a Spanish CN which is not only an accurate rendition of the source term, but also a fluent one, i.e. one which complies with the internal rules and preferences of the Spanish language in scientific discourse.

While it can be argued that multi-word terms of over 3 formants should not be the norm, the reality is somewhat different. The priority in expert-to-expert communication

is getting the message across, while linguistic sensitivity is relegated to the background. Thus, it is up to translators to take on the challenge to produce equivalents which not only meet the reader's expectations in terms of accuracy and fluency, but also prevent the spread of domain loss. In order to do so, translators are encouraged to embrace term variation and the phenomenon of multidimensionality when rendering English CNs in Spanish, as this is what make the quality of their translations stand out from those of machine engines, given the dominant position these units take up in specialized translation.

These preliminary results open the door to new lines of research, which have not as yet been addressed because of the characteristics of the study. The main lines for future research include the following:

(a) Analyzing term variation of CNs in the subject field of air pollution, not just from the perspective of the hard sciences, but also from the field of Medicine, with a view to dissecting multidimensionality in greater depth.

(b) Studying the translation of CNs in other language pairs, namely the Chinese-to-English one, in the subject field of air pollution and air quality treatment. Due to the relevance this environmental issue has in China, research is constantly being published in the country, which suggests the possibility of noun-packing in English CNs of this field being influenced by the characteristics of Mandarin Chinese.

(c) Exploring MT output of CNs in more detail. It is our aim to incorporate other rising engines to our analysis (e.g. Amazon's), as well as to propose a more granular typology for the errors encountered.

(d) Evaluating the use of our protocol in the classroom, with a view to enhance the training of translators in scientific language and specialized translation.

In conclusion, this undergraduate dissertation addressed the study of CNs specifically from the point of view of translation, further confirming the challenges they pose for both human and machine translators. While the development of the latter falls within the scope of Natural Language Processing, it is our belief that the study of these units should be given more specific attention in specialized translation courses. Research of CNs in translation has come a long way, but it still has a long way to go. Furthermore, the fact that they still prove problematic in parallel and comparable texts goes to show the importance of extrapolating the findings in the literature to the training of translators and interpreters. Research innovation should be coupled with innovation in teaching, in order to provide future professionals with the skills and tools needed to ensure translation quality of these multi-word terms.

References

Aguado de Cea, G. & Montiel-Ponsoda, E. (2012). *Term variants in ontologies*. In *XXX Congreso Internacional de AESLA 2012* (436-443). Lleida: AESLA.

Baldwin, T. & Kim, S. N. (2010). Multiword Expressions. In N. Indurkhya & F. J. Damerau (eds.), *Handbook of Natural Language Processing (2nd Edition)* (267-292). Boca Ratón: CRC Press.

Baldwin, T. & Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. In T. Tanaka et al. (eds.), *Second ACL Workshop on Multiword Expressions: Integrating Processing, Barcelona, Spain* (24-31). Morristown: ACL.

Barrière, C. & Ménard, P. A. (2014). Multiword noun compound bracketing using Wikipedia. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA)* (72-80). Dublín: ACL, Dublin City University.

Barsalou, L. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes 5*(6), 513–562.

Bauer, L. (2008). Les composés exocentriques de l'anglais. In D. Amiot (ed.), *La composition dans une perspective typologique* (35-47). Arras: Artois Presses Université.

Benson, M., Benson, E. & Ilson, R. F. (1986). Lexicographic Description of English. *Studies in Language Companion Series 14*. Amsterdam & Philadelphia: John Benjamins.

Boulanger, J.C. (1989). L'évolution du concept de néologie de la linguistique aux industries de la langue. In C. De Schaetzen (ed.), *Proceedings of Terminologie diachronique*. Paris & Brussels: CILF and Ministère de la Communauté française de Belgique.

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

Bowker, L. (1998). Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study. *Meta 43*(4), 631-651.

Buendía-Castro, M. (2013). *Phraseology in specialized language and its representation in environmental knowledge resources*. Granada: Universidad de Granada.

Burchardt, A. & Lommel, A. (2014). *Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality*. Accessed June 16, 2020. <http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf>

Cabezas García, M. (2019). *Los compuestos nominales en terminología: formación, traducción y representación*. Granada: Universidad de Granada.

Cabezas-García, M. & León-Araúz, P. (2019). On the structural disambiguation of multi-word terms. In G. Corpas-Pastor & R. Mitkov (eds.), *Computational and Corpus-Based Phraseology* (46-60). Cham: Springer.

Cabré, M. T. (1993). *La terminología. Teoría, metodología, aplicaciones*. Barcelona: Antártida, Empúries.

Cabré, M. T., Estopà-Bagot, R. & Vargas-Sierra, C. (2012). Neology in specialized communication. *Neology in Specialized Communication. Special issue of Terminology, 18*(1), 1-8.

Carrió Pastor, M. L. & Candel Mora, M. A. (2013). Variation in the translation patterns of English complex noun phrases into Spanish in a specific domain. *Languages in Contrast, 13*(1), 28-45.

Daille, B. (2005). Variations and application-oriented terminology engineering. *Terminology, 11*(1), 181-197.

Daille, B. (2017). *Term Variation in Specialised Corpora: Characterisation, automatic discovery and applications*. Amsterdam & Philadelphia: John Benjamins.

Downing, P. (1977). On the creation and use of English compound nouns. *Language, 53*, 810-842.

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

Faber, P. & Mairal, R. (1999). *Constructing a lexicon of English verbs*. Berlin: Mouton de Gruyter.

Faber, P. (2009). The cognitive shift in terminology and specialized translation. *MonTi: Monografías de Traducción e Interpretación, 1*(1), 107–134.

Faber, P. (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlín & Boston: De Gruyter Mouton.

Faber, P. (2015). Frames as a framework for terminology. In H. Kockaert & F. Steurs (eds.) *Handbook of Terminology* (14-33). Amsterdam & Philadelpia: John Benjamins.

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.

Faber, P., León-Araúz, P. & Reimerink, A. (2016). EcoLexicon: new features and challenges. In I. Kernerman et al. (eds.) *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference* (73-80).

Fantinuoli, C. & Zanettin, F. (2015). Creating and Using Multilingual Corpora in Translation Studies. In C. Fantinuoli & F. Zanettin (eds.), *New Directions in Corpus-based Translation Studies* (1-11). Berlin: Language Science Press.

Ferguson, G. (2007). The global spread of English, scientific communication and ESP: Questions of equity, access and domain loss. *Ibérica, 13*, 7-38.

Fernández-Domínguez, J. (2016). A morphosemantic investigation of term formation processes in English and Spanish. *Languages in Contrast, 16*(1), 54–83.

Fernández-Silva, S. (2018). The Cognitive and Communicative Functions of Term Variation in Research Articles: A Comparative Study in Psychology and Geology. *Applied Linguistics, 40*(4), 624–645.

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

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica 6*(2), 222–254.

Freixa, J. & Fernández Silva, S. (2017). Terminological variation and the unsaturability of concepts. In P. Drouin et al. (eds.), *Multiple Perspectives on Terminological Variation*, (155-181). Amsterdam: John Benjamins.

Freixa, J. (2002). *La variació terminològica. Anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient*. Barcelona: Universitat de Barcelona.

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

García-Page, M. (2008). *Introducción a la fraseología española*. Barcelona: Anthropos.

Gil-Berrozpe (2017). *Corpus-based identification of hyponymy subtypes and knowledge patterns in the environmental domain* (unpublished master's thesis). Universidad de Granada, Granada.

Görög, A. (2014). Quality Evaluation Today: The Dynamic Quality Framework. *Translating and the Computer, 36*.

Gregory, M. & Carroll, S. (1978). *Language and Situation: Language Varieties and their Social Contexts*. Londres, Henley & Boston: Routledge & Kegan Paul.

Guilbert, L. (1975). *La neologie lexicale*. Paris: Larousse.

Hendrickx, I., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Szpakowicz, S. & Veale, T. (2013). SemEval-2013 task 4: Free paraphrases of noun compounds. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (138-143). Atlanta: ACL.

Humbley, J. & García-Palacios, J. (2012). Neology and Terminological Dependency. *Terminology 18*(1), 59-85.

Humbley, J. (2006). La néologie: interface entre ancien et nouveau. In R. Greenstein (ed.), *Langues et cultures: une histoire d'interface*. Paris: Publications de la Sorbonne.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Jespersen, O. (1942). *A modern English grammar: On historical principles.* Copenhagen: Munksgaard.

Jiménez Crespo, M.A. & Tercedor Sánchez, M. (2016). 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. John Benjamins.

Kageura, K. (2002). *The Dynamics of Terminology. A Descriptive Theory of Term Formation and Terminological Growth.* Amsterdam: John Benjamins.

Kageura, K. (2015). Terminology and lexicography. In H. J. Kockaert & F. Steurs (eds.), *Handbook of Terminology, 1* (45-59). Amsterdam & Philadelphia: John Benjamins.

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* (83-108). Amsterdam: John Benjamins.

Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.), *Proceedings of the 11th EURALEX International Congress* (105-115). Lorient: EURALEX.

Kim, S. N. & Baldwin, T. (2013). A lexical semantic approach to interpreting and bracketing English noun compounds. *Natural Language Engineering 19*(3), 385-407.

Koehn, P. Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (127–133):

Lapshinova-Koltunski, E. (2013). VARTRA: A Comparable Corpus for Analysis of Translation Variation. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora* (77-86). Sofia: Association for Computational Linguistics.

Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Noun Compounds*. Sydney: Macquarie University.

León-Araúz, P. & Cabezas-García, M. I. (in press). Term and translation variation of multi-word terms.

León-Araúz, P. & Reimerink, A. (2016). Evaluation of EcoLexicon Images. In F. Khan et al. (eds.), *Joint Second Workshop on Language and Ontology & Terminology and Knowledge Structures (LangOnto2 + TermiKS) in conjunction with the 10th edition of the Language Resources and Evaluation Conference* (16-22).

León-Araúz, P. (2017). Term and concept variation in specialized knowledge dynamics. In P. Drouin et al. (eds.), *Multiple Perspectives on Terminological Variation* (213-258). Amsterdam & Philadelphia: John Benjamins.

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 12th Conference on Language Resources and Evaluation (LREC 2020)* (2351-2360). Marseille: ELRA.

Levi, J. N. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.

Lommel, A., Burchardt, A. & Uszkoreit, H. (2015). Quality Translation 21 D3.1: Harmonised Metric. *QT 21 Consortium*. <http://www.qt21.eu/wp-content/uploads/2015/11/QT21-D3-1.pdf> Accessed May 18, 2020.

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica:* Tecnologies *de La Traducció*, *12*(12), 455.

Marcus, M. P. (1980). *A theory of syntactic recognition for natural language*. Cambridge: Mit Press.

Martín-Mingorance, L. (1989). Functional grammar and lexematics. In J. Tomaszczyk & B. Lewandowska (eds.), *Meaning and lexicography* (227–253). Amsterdam: John Benjamins.

Maučec, M.S. & Donaj, G. (2020). Machine Translation and the Evaluation of Its Quality. In A. Sadollah & T. Sinha (eds.), *Recent Trends in Computational Intelligence*.

McEnery, T. & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., & Xiao, R. (2007). Parallel and comparable corpora: What is happening? In G. Anderman, & M. Rogers, eds. *Incorporating Corpora: The Linguist and the Translator* (18-31). Clevedon: Multilingual Matters.

Meyer, C. F. (2002). *English corpus linguistics: An introduction.* Cambridge: Cambridge University Press.

Meyer, I. & Mackintosh, K. (1996). Refining the terminographer's concept-analysis methods: How can phraseology help? *Terminology, 3*(1), 1-26.

Miyata, R. & Kageura, K. (2016). Constructing and Evaluating Controlled Bilingual Terminologies. In *Proceedings of the Fifth International Workshop on Computational Terminology (Computerm2016)* (88-93). Osaka: ACL.

Nakov, P. & Hearst, M. (2005). Search engine statistics beyond then-gram: application to noun compound bracketing. In *Proceedings of the Ninth Conference on Computational Natural Language Learning, CoNLL '05* (17-24). Ann Arbor: ACL.

Nakov, P. (2013). On the interpretation of noun compounds: syntax, semantics, and entailment. *Natural Language Engineering 19*, 291-330.

Nastase, V. & Szpakowicz, S. (2003). Exploring noun-modifier semantic relations. In J. Geertzen et al. (eds.), *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)* (285-301).

Nastase, V., Nakov, P., Ó Séaghdha, D. & Szpakowicz, S. (2013). *Semantic relations between nominals*. San Rafael: Morgan & Claypool.

Nunes-Vieira, L. (2020). Automation anxiety and translators. *Translation Studies, 13*(1), 1-21.

Ó Séaghdha, D. & Copestake, A. (2013). Interpreting compound nouns with kernel methods. *Natural Language Engineering, 19*(3), 331-356

Pecman, M. (2012). Tentativeness in Term Formation. A Study of Neology as a Rhetorical Device in Scientific Papers. *Terminology*, *18*(1), 27-58.

Pecman, M. (2014). Variation as a cognitive device: how scientists construct knowledge through term formation. *Terminology, 20*(1), 1-24.

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.

Pustejovsky, J., Anick, P. & Bergler, S. (1993). Lexical semantic techniques for corpus analysis. *Computational Linguistics, 19*(2), 331-358.

Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework.* Cham: Springer.

Resche, C. (2004). Investigating 'Greenspanese': From Hedging to 'Fuzzy Transparency'. *Discourse and Society*, *15*(6), 723-744.

Rondeau, G. (1984). *Introduction à la terminologie*. Chicoutimi: Gaëtan Morin.

Rosario, B., Hearst, M. & Fillmore, C. J. (2002). «The descent of hierarchy, and selection in Relational Semantics». En Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, P. Isabelle (ed.), 247-254. Filadelfia: ACL.

Sager, J. C., Dungworth, D. & McDonald, P. F. (1980). *English special languages. Principles and practice in science and technology*. Wiesbaden: Brandstetter Verlag.

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.

Sanz Vicente, L. (2012a). Approaching Secondary Term Formation through the Analysis of Multiword units: An English–Spanish Contrastive Study. *Terminology, 18*(1), 105-127.

Sanz Vicente, L. (2012b). Searching for patterns in the transfer of multiword units: a corpus-based contrastive study on secondary term formation. In T. Gornostav (ed.), *Proceedings of CHAT 2012. The 2nd Workshop on the Creation, Harmonization and Application of Terminology Resources. Co-located with TKE 2012* (11-18). Linköping: Linköping University Electronic Press.

Štekauer, P. (1998). *An onomasiological theory of English word-formation*. Amsterdam: John Benjamins.

Taylor, C. (2008). What is corpus linguistics? What the data says. *ICAME Journal, 32*, 179-200.

Temmerman, R. (2000). *Towards new ways of terminology description: the sociocognitive-approach*. Amsterdam & Philadelphia: John Benjamins.

Tercedor-Sánchez, M. (2011). The cognitive dynamics of terminological variation. *Terminology, 17*(2), 181-197.

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

Vanderwende, L. (1994). Algorithm for automatic interpretation of noun sequences. In M. Nagao & Y. Wilks (eds.), *Proceedings of the 15th Conference on Computational Linguistics (COLING 1994)* (782-788). Kyoto: ACL.

Venkatsubramanyan, S. & Perez-Carballo, J. (2004). Multiword Expression Filtering for Building Knowledge. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing* (40-47). Barcelona: Association for Computational Linguistics.

Way, A. (2018). Quality expectations of machine translation. In J. Moorkens et al. (eds.), *Translation Quality Assessment: From Principles to Practice* (159-178). Cham: Springer International Publishing.

White, J. S. & O'Connell, T. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas.*

Wüster, E. (1968). *The Machine Tool. An Interlingual Dictionary of Basic Concepts.* London: Technical Press.

Zanettin, F. (2014). Corpora in Translation. In J. House (ed.), *Translation: A Multidisciplinary Approach* (178-199). Basingstoke: Palgrave Macmillan.

Zuluaga, A. (1975). La fijación fraseológica. *Thesaurus, XXX*(2), 225-248.

Annex

		Google Translate	DeepL	Apertium	Systran
3-word CNs	biomass burning aerosol	aerosol de quema de biomasa	aerosol de quema de biomasa	Biomasa aerosol en llamas	aerosoles de combustión de biomasa
	passive air sampler	muestreador de aire pasivo	muestreador de aire pasivo	aire pasivo sampler	muestreo pasivo de aire
	ozone total column	columna total de ozono	columna de ozono total	ozono columna total	columna total de ozono
	air pollutant concentration	concentración de contaminantes del aire	concentración de contaminantes en el aire	Aire pollutant concentración	concentración de contaminantes atmosféricos
	surface particulate matter	material particulado superficial	partículas de la superficie	Superficie particulate asunto	materia de partículas superficiales
	organic carbon aerosol	aerosol de carbono orgánico	aerosol de carbono orgánico	Aerosol de carbono orgánico	aerosol de carbono orgánico
	soil dust aerosol	aerosol de polvo del suelo	aerosol de polvo del suelo	Aerosol de polvo de la tierra	aerosoles de polvo de suelo
	sea salt aerosol	aerosol de sal marina	aerosol de sal marina	Aerosol de sal del mar	aerosol de sal marina
	particle size distribution	distribución de tamaño de partícula	distribución del tamaño de las partículas	Distribución de medida de la partícula	distribución de tamaño de partícula
	point source emission	emisión de fuente puntual	emisión de la fuente puntual	emisión de fuente del punto	emisión de fuente de punto
4-word CNs	particle number size distribution	distribución del tamaño del número de partículas	distribución del tamaño del número de partículas	distribución de medida del número de partícula	distribución del tamaño del número de partícula

	coarse mode sea-salt aerosol	aerosol de sal marina de modo grueso	Aerosol de sal marina de modo grueso	Mar de modo tosco-aerosol de sal	aerosol de sal marina en modo grueso
	size distribution of PM	distribución de tamaño de PM	distribución de tamaño del PM	Distribución de medida de PM	distribución de tamaño de PM
	non-methane volatile organic compounds	compuestos orgánicos volátiles no metanos	compuestos orgánicos volátiles no metánicos	No-metano compuestos orgánicos volátiles	compuestos orgánicos volátiles no metano
	surface albedo radiative effect	efecto radiativo de albedo superficial	efecto radiativo del albedo de superficie	Albedo de superficie radiative efecto	efecto radiactivo de albedo superficial
	planetary boundary layer height	altura de la capa límite planetaria	altura de la capa límite planetaria	Altura de capa de frontera planetaria	altura de capa de límite planetario
	air quality monitoring station	estación de monitoreo de la calidad del aire	estación de vigilancia de la calidad del aire	Estación de control de calidad de aire	estación de control de la calidad del aire
	atmospheric aerosol optical depth	profundidad óptica de aerosol atmosférico	profundidad óptica del aerosol atmosférico	Aerosol atmosférico profundidad óptica	profundidad óptica atmosférica de aerosol
	air mass back trajectory	trayectoria de la masa de aire para atrás	trayectoria de la masa de aire hacia atrás	Masa de aire atrás trajectory	trayectoria de retorno de la masa aérea
	portable light-scattering aerosol monitor	monitor portátil de aerosol con dispersión de luz	monitor portátil de aerosol de dispersión de luz	Portátil monitor de aerosol que esparci ligero	monitor portátil de aerosol de dispersión luminosa
5-, 6- and 7-word CNs	hybrid single-particle lagrangian integrated trajectory model	modelo de trayectoria integrada lagrangiana híbrida de una sola partícula	modelo híbrido de trayectoria integrada de una sola partícula de lagrange	híbrido solo-la partícula lagrangiana integrado trajectory modelo	modelo híbrido de una sola partícula de trayectoria lagrangiana integrada
5-, 6- and 7-word CNs	anthropogenic aerosol ERF	aerosol antropogénico ERF	Aerosol antropogénico ERF	anthropogenic Aerosol ERF	aerosol ERF antropogénico
5-, 6- and 7-word CNs	photo-chemically aged biomass burning emissions	emisiones de quema de biomasa fotoquímicamente envejecidas	emisiones de quema de biomasa fotoquímicamente envejecidas	Foto-biomasa envejecida químicamente que quema emisiones	emisiones de combustión de biomasa con edades fotoquímicas
5-, 6- and 7-word CNs	diurnal FRP cycle	ciclo diurno de FRP	ciclo diurno de FRP	Diurno FRP ciclo	ciclo diurno FRP

	atmospheric particle number size distribution	distribución del tamaño del número de partículas atmosféricas	distribución del tamaño del número de partículas atmosféricas	Medida de número de partícula atmosférica distribución	distribución del tamaño del número de partículas atmosféricas
	PUF disk passive air sampler	muestra de aire pasivo de disco PUF	Muestreador de aire pasivo de disco PUF	PUF Disco aire pasivo sampler	muestreador de aire pasivo en disco PUF
	industrial NMVOCs emission	emisión industrial de COVNM	emisión de COVNM industriales	Industrial NMVOCs emisión	emisión industrial de nMVOC
	aged toluene SOA particle	partícula SOA de tolueno envejecido	partícula envejecida de tolueno SOA	nvejecido toluene SOA partícula	partícula de toluene SOA envejecida
	matrix-assisted laser desorption/ionization time-of-flight mass spectrometry	espectrometría de masas por desorción / ionización por láser asistida por matriz tiempo de vuelo	desorción e ionización láser asistida por matriz espectrometría de masas en tiempo de vuelo	matricial-ionización de desorción de láser/asistida tiempo-de-espectrometría de masa del vuelo	desorción láser asistida por matriz/ionización tiempo de espectrometría de masa de vuelo
	CCN-active aerosol fraction	fracción de aerosol activa CCN	CCN-fracción de aerosol activo	CCN-Fracción de aerosol activo	fracción de aerosoles con actividad CCN

Translations performed using the online, open-access versions of the following engines (last accessed June 14th 2020): Google Translate (<https://www.translate.google.com/>), DeepL (<https://www.deepl.com/translate/>), Apertium (<https://www.apertium.org/>), and Systran (<https://translate.systran.net/translationTools/text>).

		Google Translate				DeepL				Apertium				Systran			
		A	B	C	Σ	A	B	C	Σ	A	B	C	Σ	A	B	C	Σ
3-word CNs	biomass burning aerosol	1	1	0	2	1	1	0	2	0	0	0	0	1	1	0	2
	passive air sampler	1	0	1	2	1	0	1	2	0	0	0	0	0	1	1	2
	ozone total column	1	0	1	2	1	0	1	2	1	0	0	1	1	0	1	2
	air pollutant concentration	1	1	1	3	1	1	1	3	0	0	0	0	0	1	1	2
	surface particulate matter	1	1	1	3	1	1	1	3	0	0	0	0	1	0	1	2
	organic carbon aerosol	1	1	1	3	1	1	1	3	1	1	1	3	1	1	1	3
	soil dust aerosol	1	1	1	3	1	1	1	3	1	1	0	2	1	1	0	2
	sea salt aerosol	1	1	1	3	1	1	1	3	1	1	1	3	1	1	1	3
	particle size distribution	1	1	1	3	1	1	1	3	0	1	1	2	1	1	0	2
	point source emission	1	1	1	3	1	1	0	2	0	1	1	2	0	1	0	1
					2,7				2,6				1,3				2,1
4-word CNs	particle number size distribution	1	0	1	2	1	0	1	2	0	0	1	1	1	0	1	2
	coarse mode sea-salt aerosol	1	1	0	2	1	1	0	2	0	0	0	0	1	1	0	2
	size distribution of PM	0	1	0	1	0	1	1	2	0	1	1	2	1	1	0	2
	non-methane volatile organic compounds	1	1	0	2	1	1	1	3	1	0	0	1	1	1	0	2
	surface albedo radiative effect	1	1	0	2	1	1	0	2	0	0	0	0	0	1	1	2
	planetary boundary layer height	1	1	1	3	1	1	1	3	0	1	0	1	1	0	0	1

	air quality monitoring station	1	1	1	3	1	1	1	3	1	1	1	3	1	1	1	3
	atmospheric aerosol optical depth	1	1	0	2	1	1	1	3	1	0	0	1	1	0	0	1
	air mass back trajectory	0	1	1	2	0	1	1	2	0	0	0	0	0	1	1	2
	portable light-scattering aerosol monitor	1	1	0	2	1	1	0	2	0	0	0	0	1	1	0	2
					2,1545				2,4				0,9				1,9
5-, 6- and 7-word CNs	hybrid single-particle lagrangian integrated trajectory model	1	1	0	2	1	1	0	2	0	0	0	0	1	0	1	2
	anthropogenic aerosol ERF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	photo-chemically aged biomass burning emissions	1	1	0	2	1	1	0	2	0	0	0	0	0	1	1	2
	diurnal FRP cycle	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0
	atmospheric particle number size distribution	1	0	0	1	1	0	0	1	0	0	0	0	1	0	1	2
	PUF disk passive air sampler	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1
	industrial NMVOCs emission	1	1	1	3	1	0	1	2	0	0	0	0	0	1	0	1
	aged toluene SOA particle	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
	matrix-assisted laser desorption/ionization time-of-flight mass spectrometry	1	0	0	1	1	0	0	1	0	0	0	0	1	0	0	1
	CCN-active aerosol fraction	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

1,105	1	0	1