A semantic approach to the inclusion of complex nominals in English terminographic resources

Melania Cabezas-García[1][0000-0002-8622-1036] and Pamela Faber[2][0000-0003-0581-0005]

[1] University of Granada, Spain
melaniacabezas@ugr.es
[2] University of Granada, Spain
pfaber@ugr.es

Abstract. Complex nominals (CNs) are characterized by the omission of the semantic relation between their constituents due to noun packing. Despite their frequency in specialized texts written in English [1] their representation and inclusion in knowledge resources has received little research attention. This paper presents a proposal for the inclusion of CNs in an English terminographic resource on renewable energy. For that purpose, we used knowledge patterns and paraphrases to access the meaning of CNs in a wind power corpus. We then filled the definitional templates proposed by Frame-based Terminology [2]. Our main goal was to conceptually organize a term entry to facilitate knowledge of the domain while keeping the entry length to a minimum. Furthermore, this proposal is a valuable starting point toward the development of bilingual and multilingual resources since translation should be based on meaning. Our results also afforded insights into compound term formation in English, as reflected in the addition of specific values to the semantic relations encoded by the hypernym. Term instability and multidimensionality were also prevalent.

Keywords: Complex Nominal, Semantics, Terminography.

1 Introduction

Renewable energies have led to the creation of new terms that should be included in knowledge resources. Complex nominals (CNs) are of particular importance since these phraseological units are very frequent in scientific texts [1][3][4][5][6][7]. Noun packing, the omission of constituents, and the non-specification of the semantic relation between the units forming the CN, often result in a lack of compositionality and transparency in these terms [8][9]. In other words, the meaning of a CN cannot always be predicted from its head and modifiers [10]. The only clear information is that "it denotes something (conveyed by the head) that is somehow related to something else (conveyed by the modifier)" [8: 100].

Cabezas-García M., Faber P. (2017) A Semantic Approach to the Inclusion of Complex Nominals in English Terminographic Resources. In: Mitkov R. (eds) Computational and Corpus-Based Phraseology. EUROPHRAS 2017. Lecture Notes in Computer Science, vol 10596. Springer, Cham. https://doi.org/10.1007/978-3-319-69805-2_11

This paper describes how CNs can be included in an English terminographic resource on renewable energies. For that purpose, a corpus of specialized texts on wind power was used to extract paraphrases and knowledge patterns [11][12] (see section 3), which facilitated access to the semantics of these phraseological units. Furthermore, a specialized corpus on the environment (available in Open Corpora in Sketch Engine) provided a larger amount of data. We then filled out the definitional templates proposed by Frame-based Terminology [2], which include the semantic relations encoded by the CNs and permit the clustering of related terms.

Our objectives were to access the semantics of CNs, verify whether their meaning could be understood in terms of similar CNs [10], and conceptually organize the term entry on the basis of this shared meaning. To the best of our knowledge, the semantic organization of specialized CNs, and the inclusion and description of CNs formed by more than two terms have not received sufficient research attention. Moreover, this proposal facilitates knowledge representation of the domain while keeping the entry length to a minimum. It is also a valuable starting point toward the development of bilingual and multilingual resources [3][13].

2 Complex nominals in dictionaries

Complex nominals (CNs), e.g. *power plant*, are very frequent in English specialized texts [1][3][4][5][6][7]. They are expressions with a head noun preceded by one or more modifiers (i.e. nouns or adjectives) [14]. These multi-word terms are characterized by their syntactic-semantic complexity since two or more concepts are juxtaposed without any explicit indication of the relation between them [10]. This usually entails the formation of long CNs that may be difficult to understand [15], which highlights the need to describe them in specialized resources. CNs can be endocentric (the focus of our study), when one term is the head and the other constituents modify it [1], e.g. *wind power*. Alternatively, they can be exocentric, when the CN is not a hyponym of one of its elements, and thus appears to lack a head [16], e.g. *saber tooth*.

One of the essential characteristics of CNs is the existence of underlying propositions that can be inferred in the term formation processes, as highlighted in Levi [14]: predicate deletion (e.g. *power plant* instead of *a plant produces power*) and predicate nominalization (e.g. *power generation* instead of *power is generated*). Along these lines, Mel'čuk et al. [17] argue that argument structure is fundamental when describing predicates. Given that CNs have concealed or nominalized verbs, the study of micro-contexts (i.e. the relation between a predicate and its argument structure [18]) is crucial in terminographic descriptions.

In the last twenty years, research on CNs has addressed the formation and use of these multi-word units, their semantics, and different methods for interpreting them [14][19][20][21][22][23][24]. More recently, CNs have been investigated for translation purposes [5], and special attention has been paid to their formation and interpretation [7][25], namely by means of paraphrasing verbs and prepositions [1][26]. However, the focus has been on two-term CNs. Furthermore, CNs have not been systematically treated in dictionaries [27] though most authors agree on their inclusion as sub-

lemmas of a main entry [11][28] because other locations could prevent readers from finding the right information [11]. Whereas some authors point out that multi-word units should appear under the first content word [29], others defend that CNs should be a sublemma of the head noun [28][29]. Nonetheless, there is general consensus that the inclusion and treatment of phraseological units depends on user needs [11][27].

3 Meaning access in complex nominals

The non-specification of the semantic relation between CN constituents often makes it difficult to understand the meaning of these phraseological units. Traditionally, inventories of semantic relations have been the preferred way of accessing this conceptual link ([19][20][30][31] *inter alia*). However, these classifications can pose problems such as the choice of the best set of relations, their abstract nature, and the existence of more than one possible relation in the same CN [1].

For these reasons, Downing [21] argues that current inventories of semantic relations cannot capture the conceptual relation between the constituents of a CN. In this respect, authors such as Nakov and Hearst [32] suggest that the best way of ascertaining the meaning of a CN is by means of multiple verb paraphrases. For instance, *malaria mosquito* can be paraphrased using different verbs such as *carry*, *spread*, *cause* or *transmit*, which specifically convey the action carried out by the mosquito (for more examples, see Hendrickx et al. [4], Butnariu et al. [26], and Nulty and Costello [33] *inter alia*). In our opinion, inventories of semantic relations and paraphrases are complementary approaches since the informativity of conceptual relations can be enhanced by means of fine-grained verb paraphrases [18].

Knowledge patterns (KPs) are also very useful for the extraction of semantic relations [34][35][36]. They are lexico-syntactic patterns that encode semantic relations in real texts [11][12]. Table 1 shows some of the most frequent KPs in the environmental domain (as well as in general language) [37: 8].

Table 1. Knowledge patterns in León-Araúz and Reimerink [37: 8].

Semantic relation	Knowledge pattern
IS_A	such as, rang* from, includ*
PART_OF	includ*, consist* of, formed by/of
MADE_OF	consist* of, built of/from, constructed of, formed by/of/from
LOCATED_AT	form* in/at/on, found in/at/on, tak* place in/at, located in/at
RESULT_OF	caused by, leading to, derived from, formed when/by/from
HAS_FUNCTION	designed for/to, built to/for, purpose is to, used to/for
EFFECTED_BY	carried out with, by using

Another kind of KP are ‘grammatical knowledge patterns’ (e.g. noun + verb), that coincide to a great extent with verb paraphrases and are very useful when identifying functional relations [11].

Nevertheless, KPs also have difficulties such as noise and silence, pattern variation, anaphora, linguistic and domain dependence, etc., which must be taken into account [36]. Section 4 describes the use of paraphrases and KPs in this research.

4 Materials and methods

For the purpose of the study, a corpus[1] on wind power of approximately 1 million words was manually compiled. It was composed of highly specialized texts, namely scientific articles and PhD dissertations, originally written in English and published in high-impact academic journals. The corpus was uploaded to Sketch Engine (https://www.sketchengine.co.uk/) [38], a corpus analysis tool that can generate concordance lines, word sketches (frequent word combinations), wordlists, etc.

The 'Keywords/Terms' function of Sketch Engine was then used to extract a list of the single-word (keywords) and multi-word lexical units (terms) most typical of the wind power corpus, which was automatically contrasted with a reference corpus. Given the limited scope of this study, the maximum number of keywords and terms was set at 100. We observed that *turbine* was ranked first in the keywords list and *wind turbine* appeared in third position on the terms list. The high prevalence of these terms was confirmed in a term extractor, TermoStat (http://termostat.ling.umontreal.ca/) [39], where *wind turbine* was the second most frequent term of a list of 8,533[2] CNs of the wind power corpus. After consulting specialized resources [40][41][42], it was found that *wind turbine* was not uniformly treated, and more often than not, its hyponyms were not described, despite the fact that they are recurrent terms in the domain.

As a solution, we analysed the word sketches generated in Sketch Engine. These sketches showed the terms that usually modify *wind turbine* and thus allowed us to select term candidates that were hyponyms of *wind turbine*. We selected those CNs with a higher frequency[3], which did not belong to an extended, irrelevant CN, and whose constituents were linked by semantic relations, with a view to conceptually organizing them and to facilitating knowledge acquisition. CNs formed by general words were rejected (e.g. *prospective wind turbine*, *offshore wind turbine*, *three-bladed wind turbine*) since the *differentiae* with other cohyponyms could be easily inferred. However, other CNs such as *variable speed wind turbine*, apparently easy to understand, convey information that is relevant to their meaning and which cannot be directly elicited from their surface form. They were thus included in our proposal. Our list of term candidates was then composed of 12 CNs, which were hyponyms of *wind turbine*, such as *lift force wind turbine*, *upwind turbine,* and *shroud wind turbine*.

According to Frame-based Terminology [2], each conceptual category has a prototypical template composed of the semantic relations activated by this category. Definitional templates, which were used in Frame-based Terminology since the ONCOTERM research project [2], are thus the basis for homogeneous category-specific definitions that make the semantic relations explicit, as well as the logical organiza-

---

1 This corpus is planned to be annotated with CNs occurrences and made available in Open Corpora (Sketch Engine).

2 Even though TermoStat is an excellent term extractor, it often offers some noise due to the inclusion of wrong CNs (e.g. *page u*) or irrelevant parts of longer CNs (e.g. *mw wind*).

3 Since we focused on CNs that were hyponyms of *wind turbine*, the search was limited to CNs mostly formed by three or more constituents. This explains the relatively low frequency of term candidates (35 occurrences on average). However, since they are key concepts of the domain, they should be described in a resource specialized in wind power.

tion of the microstructure of a term entry. First, the template of the hypernym *wind turbine* was filled with the information extracted by means of KPs and paraphrases, as detailed below, and later applied to its hyponyms. Property inheritance was evidently present, and subtypes added specific values that distinguished them from their cohyponyms. Table 2 shows the template of *wind turbine*, whose properties are inherited by its hyponyms, which add specific values (Table 3), such as the attributes of the parts (*axis of rotation parallel to the ground*). Although, this has some similarities with traditional ontologies, as previously mentioned, it advances the idea of a category-specific template that acts as a blueprint for the definitions of category members.

Table 2. Definitional template of *wind turbine*.

wind turbine	
IS_A	device
HAS_PART	blade, rotor, shaft, generator, nacelle, gearbox, bearings, yaw control, tower
USES_RESOURCE	wind
HAS_FUNCTION	convert wind energy to electrical or mechanical power

Table 3. Definitional template of *horizontal axis wind turbine*.

horizontal axis wind turbine	
IS_A	wind turbine
HAS_PART	axis of rotation parallel to the ground

With a view to accessing the meaning of CNs and filling in these templates, KPs were applied to the wind power corpus in order to ascertain the semantic relations encoded by *wind turbine*. For that purpose, the KP-based grammars developed by León-Araúz et al. [36] were implemented in the wind power corpus. This facilitated the grouping of related terms in word sketches that specify the semantic relation between the terms (e.g. PART_OF, HAS_FUNCTION, LOCATED_AT, etc.). Subsequent CQL queries of the KPs collected in León-Araúz et al. [36] were performed in order to find further knowledge-rich contexts (KRCs), i.e. "a context indicating at least one item of domain knowledge that could be useful for conceptual analysis" [11]. Nevertheless, it was impossible to extract sufficient data because of the reduced size of the corpus (which will be expanded in future work) and the limited number of linguistic forms that a semantic relation can have in specialized texts [35].

Therefore, we decided to use the EcoLexicon English Corpus, a corpus of specialized environmental texts, consisting of more than 23 million words pertaining to different environmental subdomains, which is now available in Open Corpora (Sketch Engine). By using the KP grammars [36] and CQL queries of KPs, the semantic relations activated by *turbine* were ascertained, such as its parts and its function. Even though future work will further refine these grammars and enhance them with more KPs and restrictions, these word sketches permitted us to access the conceptualization of *wind turbine* (essential in the formation of specific hyponymic CNs) and thus elaborate the definitional template of this CN. This template was complemented with and

confirmed by the data extracted by means of paraphrases, as subsequently explained, and the information in specialized resources [40][41][42].

Thus, paraphrases were also used to query the corpus. As argued in Auger and Barrière [35], when elucidating the semantic relation implicit in CNs, KPs must be complemented by an analysis of the syntactic relations that show semantic relations. This can be accomplished by means of paraphrases, which specify the relation between the constituents of the CN [1]. We thus performed CQL queries in Sketch Engine, which allows more sophisticated queries for the optimal extraction of paraphrases with specific lexical or grammatical patterns. Table 4 shows a query to extract words between *turbine* and *lift* and vice versa, in a span of 10 tokens. As can be observed, the paraphrases reveal the semantic content that is concealed in CNs because of noun packing, and they also give access to explanatory segments. Furthermore, they permit the identification of related terms (e.g. *lift force wind turbine* and *drag force wind turbine* are terminological antonyms, as reflected in the use of *instead* and *either*).

Table 4. CQL query of paraphrases of *lift force wind turbine*.

(meet [lemma= "turbine"] [lemma="lift"] -10 10) within <s/>
Modern wind **turbines** are predominantly based on aerodynamic **lift**. Lift force use aerofoils (blades) that interact with the incoming wind.
Wind **turbines** using aerodynamic **lift** can be further divided according to the orientation of the spin axis into horizontal-axis and vertical-axis type turbines.
Turbines can be divided into "**lift**" machines and "drag" machines according to which force is generated by the wind and exploited as "motive force".
In the "**lift**" **turbines**, with respect to the "drag" type, the wind flows on both blade surfaces, which have different profiles, thus creating at the upper surface a depression area with respect to the pressure on the lower surface.
The design of these modern **turbines** uses **lift** instead of drag to spin the blades.
Depending on the design of the **turbine**, either drag or **lift** moves the blades.

As previously mentioned, our objective was to verify whether the semantics of specialized CNs could be at least partly derived from the meaning of similar CNs [10][15][25][43]. To this end, hyponyms were organized in different groups based on the semantic relation between their constituents. The groups were alphabetically arranged, whereas the CNs in each group were listed according to their frequency. For example, *horizontal axis wind turbine* had 33 occurrences and thus appeared before *vertical axis wind turbine*, which had 27 occurrences (see Figure 1, where indentation shows hyponymic relations, semantic relations appear in small caps in square brackets, and attributes are in brackets).

wind turbine
  [HAS_PART (DIRECTION)]
  horizontal axis wind turbine; vertical axis wind turbine
    [HAS_PART (LOCATION)]
    upwind turbine; downwind turbine
  [MOTIVE_FORCE]

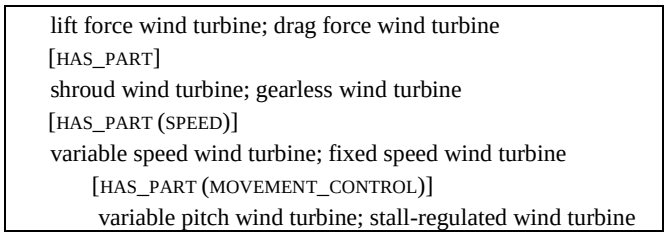

Fig. 1. Organization of hyponyms based on the semantic relations between their constituents.

Finally, we performed CQL queries to find KPs that revealed synonyms of the CNs, such as *is a synonym of*, *also called*, *referred to as*, etc. (see Figure 2 below). Other synonyms were found by the identification of synonymic KPs when reading parallel documents (i.e. online texts and websites on wind power) for documentation. For instance, *diffuser augmented wind turbine* was found to be a synonym of *shroud wind turbine*, as revealed in concordances such as *A shroud wind turbine, often referred to as a diffuser augmented wind turbine* (…).

5 Semantic organization of a term entry

At first glance, a user of a specialized resource might think that all the CNs based on *wind turbine* are subtypes of this concept without any internal differences. However, after an in-depth conceptual analysis of CNs by means of KPs and paraphrases, hyponyms of *wind turbine* were found to belong to different hierarchical levels. In other words, they established different semantic relations, and some of them were found to be hyponyms of other terms. An effective specialized resource should reflect these differences to facilitate understanding and the eventual translation of the terms.

We thus propose the inclusion of CNs as sublemmas of a main entry. Since they are subtypes of a superordinate concept, a logical structure would presumably reflect this conceptual hierarchy. Furthermore, CNs usually designate very specific concepts, which explains the relatively low number of occurrences in the corpus that would validate their inclusion as main entry terms. On the other hand, the head of these CNs is a noun, which is the part of speech most often consulted by users [29], thus avoiding difficulties in finding the CN in question.

Our proposal focuses on the conceptual organization[4] of the microstructure of a term entry. Hyponyms are usually CNs, which, despite their formal similarity, may have quite different meanings, as reflected in the concealed semantic relation between their components. Many authors defend a semantic approach to lexicographic and terminographic resources [2][17][44] since this reveals domain structure, facilitates understanding of the concepts, and provides the basis for translation. This is extremely important because English texts are often translated for knowledge dissemination purposes. Furthermore, in our opinion, hyponymic CNs must be defined since they

[4] The overlap in the CNs of this study is not a general rule since many hyponyms do not show the same linguistic form as their hypernyms, e.g. *abrasion* as a hyponym of *erosion*.

are often formed by more than three constituents with no specification of the relation between them. This evidently makes their comprehension more difficult.

With a view to conceptually organizing a term entry, we studied whether in specialized CNs, similar modifiers complemented the head in the same way [10][15][25][43]. For instance, given the semantic relation CAUSES in *wind erosion*, *water erosion* is expected to establish the same semantic relation [25] since the slots opened by similar heads tend to be filled by similar modifiers [43], and *vice versa*. This would indicate that the semantics of CNs could be partly derived from the meaning of similar CNs [10]. In other words, our assumption was that CNs modified by similar terms (e.g. *variable speed wind turbine* and *fixed speed wind turbine*) establish the same semantic relation between their constituents. Thus, if one of the CNs were defined, it would not be necessary to define the other. Nevertheless, after analysing the meaning of CNs and their implicit semantic relations by means of KPs and paraphrases, it was found that this hypothesis was not satisfactory for all specialized CNs. This occurred because many of the specialized CNs were not compositional, i.e. their meaning could not be directly construed from the meaning of their parts [8][9][10], because there were often concealed constituents that were required for an accurate understanding of the CN. For instance, *stall-regulated wind turbine* is not fully compositional because there is information missing (namely, the high wind speed conditions necessary for the turbine to stall). Without further clarification, the meaning of the CN is opaque since it cannot be understood from the meaning of its parts [10].

Therefore, our assumption was only applied to compositional (i.e. transparent) CNs, where the only difficulty was the specification of the semantic relation between their parts. This was especially true for CNs that are in opposition to each other (e.g. *horizontal axis wind turbine* and *vertical axis wind turbine*). In these cases, it was possible to infer the meaning of one of the related CNs from the definition of other CNs in the same group (i.e. modified by similar terms and linked by the same semantic relation [10][15][25][43]). As for non-compositional CNs (which does not mean that they are idiomatic), both co-hyponyms were defined (e.g. *variable pitch wind turbine* and *stall-regulated wind turbine*). Despite the fact that their constituents belonged to the same family, the meaning of the second CN was not easily construed from the definition of the first because additional information was required. Therefore, we found that the omission of the semantic relation was not the only problem in CNs. More specifically, domain specificity and excessive noun packing (a source of bracketing complexity) do not support the statement that the semantics of CNs can be partly derived from the meaning of similar CNs [10]. However, this hypothesis is of great importance since it considers essential features of CNs, though its application depends on the purposes of the resource.

Figure 2 shows our proposal for the entry of *wind turbine* in a specialized knowledge resource on renewable energies. This model can be applied to any lexicographic or terminographic entry, both in electronic and printed[5] format. Electronic resources are a frequent option in today's world, given the fact that, unlike printed dictionaries, they have no space restrictions and are easily updated. Furthermore, this

[5] Usage examples should be explicitly stated in printed resources.

format offers different access points to the information [45] (e.g. search for phraseological pattern, conceptual category, lemma, conceptual representation, etc.).

wind turbine: *turbine* [USES_RESOURCE] *wind*.
Wind-driven device that converts wind energy to electrical or mechanical power (*syn.* wind generator, windmill, aerogenerator). Usage examples.

horizontal axis wind turbine: *wind turbine* [HAS_PART (DIRECTION)] *horizontal axis*.
Wind turbine whose axis of rotation is parallel to the ground (*syn*. HAWT). Usage examples. *Related terms*: vertical axis wind turbine (*syn.* VAWT). Usage examples.

upwind turbine: *horizontal axis wind turbine* [HAS_PART (LOCATION)] *upwind.*
Horizontal axis wind turbine whose rotor faces the wind. Usage examples. *Related terms*: downwind turbine. Usage examples.

lift force wind turbine: *wind turbine* [MOTIVE_FORCE] *lift force*.
Wind turbine that uses lift forces (perpendicular to the direction of the air flow) to spin the blades and turn the rotor. Usage examples. It contrasts with **drag force wind turbine** (*syn*. impulse wind turbine), which uses drag forces (parallel to the direction of the air flow). Usage examples.

shroud wind turbine: *wind turbine* [HAS_PART] *shroud*.
Wind turbine protected by a shroud that accelerates the incoming wind, significantly increasing the mass and power available to the turbine (*syn.* ducted wind turbine, diffuser augmented wind turbine, DAWT). Usage examples. *Related terms*: gearless wind turbine. Usage examples.

variable speed wind turbine: *wind turbine* [HAS_PART (SPEED)] *variable speed*.
Wind turbine in which the rotor speed increases and decreases with changing wind speeds. Usage examples. *Related terms*: fixed speed wind turbine (*syn.* constant speed wind turbine). Usage examples.

variable pitch wind turbine: *variable or fixed speed wind turbine* [HAS_PART (MOVEMENT_CONTROL)] *variable pitch*.
Variable or fixed speed wind turbine that adjusts the angle of the blades out of the wind when experiencing high operational wind speeds in order to control the output power (*syn*. pitch controlled wind turbine). Usage examples. It contrasts with **stall-regulated wind turbine** (*syn*. passive stall-regulated wind turbine, fixed pitch wind turbine), whose blades respond to high wind speeds by stopping turning. Usage examples.

Fig. 2. Term entry proposal for *wind turbine* and its hyponyms.

As can be observed, the semantic relation between the CN constituents was specified as a first step towards a full understanding of meaning. Related CNs were grouped together when they were modified by similar terms, and the same relation was established between their constituents. In each group, the phraseological unit with the highest frequency in our wind power corpus was defined. As for compositional CNs, the other CNs in the same group were only included as 'related terms'. A definition was not needed since their meaning can be easily inferred from the definition of the first CN. The only difficulty in compositional CNs was the non-specification of the semantic relation. Some examples are *horizontal axis wind turbine* and *vertical axis wind turbine*, *upwind turbine* and *downwind turbine*, *shroud wind turbine* and *gear-*

less wind turbine, and *variable speed wind turbine* and *fixed speed wind turbine*. Alternatively, in non-compositional CNs, it was not possible to deduce the meaning of the CN from the definition of other related CNs (mostly because of the omission of constituents relevant to their meaning). Thus, the description of the related CN (e.g. *drag force wind turbine* and *stall-regulated wind turbine*) was preceded by the expression *it contrasts with*. The latter CN does not require a full definition, given that the characteristics differentiating it from its cohyponym are sufficient.

Definitions were based on the templates proposed by Frame-based Terminology [2]. These templates reflect the semantic relations activated by a conceptual category and provide consistency to term entries. The definitions in our proposal are composed of a *genus* and *differentiae*. The *genus* indicates the category to which the term belongs. In this case, all the CNs were hyponyms of *wind turbine*, although there were also more specific subtypes, such as *upwind turbine*, a hyponym of *horizontal axis wind turbine*. Thus, property inheritance is evident in the sense that hyponyms acquire the characteristics in the definition of their hypernym. On the other hand, the *differentiae* are the features that distinguish a hyponym from its hypernym and the other units in its lexical domain [2]. The *differentiae* of the hyponyms of *wind turbine* are usually based on attributes of the parts of a turbine.

Although CNs are characterized by a high degree of instability [46], as shall be explained, our proposal only included abbreviations and those synonyms whose linguistic form was significantly different from the CN in question. Synonyms were presented in brackets, introduced by the reduced expression *syn*. and followed by a full stop. In non-compositional CNs, where all related CNs were defined, their synonyms were placed immediately after the name of the CN to avoid possible misunderstandings. Finally, usage examples can be consulted by clicking on the hyperlink (simulated with underlined characters), which shows concordance lines of each CN in the EcoLexicon English Corpus, available in Open Corpora (Sketch Engine).

As previously highlighted, the semantic organization of term entries facilitates translation, because meaning is the starting point when rendering terms into another language. English is the *lingua franca* of specialized communication, but there is a need for translation for purposes of knowledge dissemination. For these reasons, this proposal can be the basis for creating resources in other languages, especially given the proliferation of renewable energy solutions in many different countries. In particular, this model can be used for the implementation of multi-word terms in the phraseological module of EcoLexicon (www.ecolexicon.ugr.es), a terminological knowledge base on environmental science that is conceptually organized.

6 Term formation

The hyponyms of *wind turbine* were analysed as an example of CN formation in specialized domains. In English, the creation of CNs is the order of the day [5][7][46]. CNs are generally created to name more specific concepts, and thus are usually hyponyms of a superordinate term. The hyponyms of *wind turbine* were created by adding specific values to the semantic relations encoded by *wind turbine*, which appear in its

definitional template. The CNs studied refer to different parts of a turbine, namely to specific features of these parts, but such components are not explicitly mentioned in the CN. This adds extra syntactic-semantic complexity to these phraseological units.

A distinguishing feature of new CNs is their instability, as reflected in their variants [28][46]. This instability is clearly evident in the concordance lines of the CNs, where the frequent omission of some of the constituents of the CN is noteworthy. This elision usually occurs as longer CNs with a high frequency in specialized discourse are formed. For instance, in the case of (long) hyponyms of *wind turbine*, the constituent usually omitted is *wind* or the part of the turbine in question since this is the most evident information that can be disregarded. In contrast, the *differentiae* are always present, because these are the distinguishing features of the term.

This instability of complex terms is linked in many cases to multidimensionality, an essential phenomenon in specialized domains. The features of a concept are usually specified from different perspectives and the set of characteristics that define a concept is normally multidimensional [47: 120]. Therefore, the hyponyms of *wind turbine* emphasize different features, such as the direction of the axis of rotation, and the location or the speed of the rotor. This conceptual dynamism does not mean that several concepts are involved, but rather that different perspectives are taken in order to highlight one or more characteristics of the same concept. For example, a *horizontal axis wind turbine* can also be regarded as a *variable speed wind turbine* or a *lift force wind turbine*, depending on the information emphasized.

Furthermore, CNs are a special type of term since they have the potential to combine different dimensions in one phraseological unit. The union of these dimensions results in the formation of very long CNs since these dimensions are part of the micro-context of the concealed proposition. In other words, they belong to the argument structure, either as arguments, adjuncts, or attributes of these complements. Different examples of this phenomenon were observed in the wind power corpus. For example, *stall-regulated horizontal axis wind turbine* alludes to the direction of the axis of rotation and the mechanism of movement control in high wind speeds, whereas *horizontal axis offshore wind turbine* refers to the direction of the axis of rotation as well as the location of the turbine. The formation of long CNs adds syntactic-semantic complexity to these units since internal groupings must be identified (i.e. bracketing) in order to ascertain where semantic relations are established.

Because of multidimensionality, the same concept can be involved in different situations, which can affect its relations in the conceptual system, and thus should be considered in knowledge representations [37]. Therefore, the multidimensionality in our CNs underscores the semantic complexity of these phraseological units. It also gives the user a situational picture since it elicits the frame or underlying knowledge structure by making the conceptual dimensions explicit, either by forming long CNs or by highlighting certain dimensions. Frames and multidimensionality thus play a key role in term formation, which is represented in our proposal by means of the conceptual organization of term entries.

Furthermore, micro-contexts are the root of compound term formation since these CNs are the result of concealed propositions. English CNs are characterized by noun packing. However, when translating these phraseological units, this mechanism must

be adapted to the term formation rules of the target language. For instance, in Romance languages the underlying semantic relation must be made explicit, namely in long CNs. This usually produces paraphrase structures that make the concealed verb explicit. The conceptual organization of term entries is thus valuable since translation and idiomatic adaptations must be based on meaning.

7 Conclusions

CNs are very frequent in English specialized texts [1]. These phraseological units are characterized by their syntactic-semantic complexity, which highlights the need to include multi-word terms in linguistic resources. However, up until now CNs have not been systematically treated in dictionaries. This paper presents a proposal for the inclusion of CNs in an English terminographic resource on renewable energies. For that purpose, a wind power corpus was used to extract paraphrases and KPs [11][12], which allowed access to the semantics of CNs. Also used was an environmental corpus that provided further data. We then filled in the definitional templates proposed in Frame-based Terminology [2], which included the semantic relations activated by the CNs and allowed the clustering of related terms. Our main goal was to conceptually organize a term entry in order to accurately structure knowledge and facilitate the understanding of concepts.

The results of this study showed that the description of CNs in specialized resources is essential, because they play a key role in conceptual systems [5][7][47][48]. As stated by Sager et al. [48], the constituents of a CN are linked by a semantic relation in the conceptual system. Thus, the terminological system is connected to the conceptual system since the semantic relations in CNs (see Figure 1) allow the reconstruction of the semantic network of a domain [5]. Accordingly, the semantic organization of term entries allows the specification of the different types of hyponym, which are usually CNs. In addition, it favours awareness of the frame or knowledge structure underlying term formation by including related terms, while keeping the entry length to a minimum.

In this research we studied compound term formation based on an analysis of the hyponyms of *wind turbine*. These multi-word terms added specific values to the semantic relations in their hypernyms. A high degree of instability was also observed, since some of the constituents were frequently omitted. Multidimensionality, which is frequent in specialized domains, was found to take part in compound term formation by selecting one dimension in the CN or combining several dimensions, which resulted in long phraseological units.

This proposal for the inclusion of complex terms in specialized resources can be helpful for different users ranging from specialists and semi-experts in the energy domain to language professionals and students. In spite of being a monolingual resource, it provides the basis for the transfer of knowledge to other languages since meaning is the starting point in translation. In particular, when translating CNs into Romance languages, a concept-based approach is particularly useful, because English noun packing is usually rendered in the form of paraphrase structures that make the

concealed semantic relation explicit. Accordingly, plans for future research include the analysis of the role of predicates in compound term formation as well as the translation of CNs into Spanish, with a view to implementing these multi-word terms in the phraseological module of EcoLexicon (www.ecolexicon.ugr.es). Although the procedure is mostly done manually, its application to different types of CN will allow the extraction of conceptual information (sets of semantic relations, attributes, conceptual categories) to be implemented in EcoLexicon. This will speed up the inclusion of new CNs in the phraseological module. Moreover, the use of a distributional semantic model [49] will help to identify related terms. Nevertheless, in contrast to phraseological information that is automatically included (without classification or filtering) in other resources such as the word sketches of Sketch Engine, manual work is an added value in EcoLexicon.

8 Acknowledgements

This research was carried out as part of project FF2014-52740-P, Cognitive and Neurological Bases for Terminology-enhanced Translation (CONTENT), funded by the Spanish Ministry of Economy and Competitiveness. Funding was also provided by an FPU grant given by the Spanish Ministry of Education to the first author. Finally, we would like to thank the anonymous reviewers for their useful comments.

References

1. Nakov, P.: On the interpretation of noun compounds: Syntax, semantics, and entailment. Natural Language Engineering 19(03), 291–330 (2013).
2. Faber, P., López Rodríguez, C. I., Tercedor Sánchez, M.: Utilización de técnicas de corpus en la representación del conocimiento médico. Terminology 7(2), 167–198 (2001).
3. Daille, B., Dufour-Kowalski, S., Morin, E.: French-English multi-word term alignment based on lexical context analysis. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), pp. 919-922 (2004).
4. Hendrickx, I., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Szpakowicz, S., Veale, T.: SemEval-2013 Task 4: Free Paraphrases of Noun Compounds. In: Second Joint Conference on Lexical and Computational Semantics (*SEM): Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) 2, pp. 138–143 (2013).
5. Sanz Vicente, L.: Análisis contrastivo de la terminología de la teledetección. La traducción de compuestos sintagmáticos nominales del inglés al español. PhD Thesis. University of Salamanca, Salamanca (2011).
6. Fernández-Domínguez, J.: A morphosemantic investigation of term formation processes in English and Spanish. Languages in Contrast 16(1), 54–83 (2016).
7. Kageura, K.: The Quantitative Analysis of the Dynamics and Structure of Terminologies. John Benjamins, Amsterdam/Philadelphia (2012).
8. Grant, L., Bauer, L.: Criteria for Re-defining Idioms: Are we Barking up the Wrong Tree? Applied Linguistics 25(1), 38–61 (2004).
9. Smith, V., Barratt, D., Zlatev, J.: Unpacking noun-noun compounds: Interpreting novel and conventional food names in isolation and on food labels. Cognitive Linguistics 25(1), 99–147 (2014).

10. Ó Séaghdha, D., Copestake, A.: Interpreting compound nouns with kernel methods. Natural Language Engineering 19, 1–26 (2013).
11. Meyer, I.: Extracting Knowledge-Rich Contexts for Terminography: A Conceptual and Methodological Framework. In: Bourigault, D., Jacquemin, C., L'Homme, M.-C. (eds.) Recent Advances in Computational Terminology, pp. 279–302. John Benjamins, Amsterdam (2001).
12. Marshman, E.: Lexical Knowledge Patterns for Semi-automatic Extraction of Cause-effect and Association Relations from Medical Texts: A Comparative Study of English and French. PhD Thesis, Université de Montréal, Montréal (2006).
13. Morin, E., Daille, B., Prochasson, E.: Bilingual Terminology Mining from Language for Special Purposes Comparable Corpora. In: Sharoff, S., Rapp, R., Zweigenbaum, P., Fung, P. (eds), Building and Using Comparable Corpora, pp. 265–284. Springer, Berlin/Heidelberg (2013).
14. Levi, J.: The Syntax and Semantics of Complex Nominals. Academic Press, New York (1978).
15. Rallapalli, S., Paul, S.: A Hybrid Approach for the Interpretation of Nominal Compounds using Ontology. In: 26th Pacific Asia Conference on Language, Information and Computation, pp. 554–563 (2012).
16. Bauer, L.: Les composés exocentriques de l'anglais. In: Amiot, D. (ed.) La composition dans une perspective typologique, pp. 35–47. Artois Presses Université, Arras (2008).
17. Mel'čuk, I., Clas, A., Polguère, A.: Introduction à la lexicologie explicative et combinatoire. Duculot, Louvain-la-Neuve (1995).
18. Cabezas-García, M., Faber, P.: Exploring the Semantics of Multi-word Terms by Means of Paraphrases. In: Temas actuales de Terminología y estudios sobre el léxico, pp. 193–217. Comares, Granada (2017).
19. Vanderwende, L.: Algorithm for automatic interpretation of noun sequences. In: Proceedings of the 15th conference on Computational linguistics, 2. COLING '94, pp. 782–788 (1994).
20. Rosario, B., Hearst, M. A., Fillmore, C.: The Descent of Hierarchy, and Selection in Relational Semantics. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL '02, (July), pp. 247–254 (2002).
21. Downing, P.: On the creation and use of English compound nouns. Language 53, 810–842 (1977).
22. Lauer, M.: Corpus statistics meet the noun compound: Some empirical results. In: The Association for Computational Linguistics Conference (ACL), pp. 47–54 (1995).
23. Warren, B.: Semantic patterns of noun-noun compounds. Acta Universitatis Gothoburgensis, Göteborg (1978).
24. Kageura, K.: The Dynamics of Terminology: A descriptive theory of term formation and terminological growth. John Benjamins, Amsterdam/Philadelphia (2002).
25. Kim, S. N., Baldwin, T.: A Lexical Semantic Approach to Interpreting and Bracketing English Noun Compounds. Natural Language Engineering 1(1), 1–23 (2013).
26. Butnariu, C., Kim, S. N., Nakov, P., Ó Séaghdha, D., Szpakovicz, S., Veale, T.: SemEval-2 Task 9: the interpretation of noun compounds using paraphrasing verbs and prepositions. In: Proceedings of the Fifth International Workshop on Semantic Evaluation (SemEval 2010), pp. 39–44 (2010).
27. Parra Escartín, C., Losnegaard, G. S., Samdal, G. I. L., Patiño García, P.: Representing Multiword Expressions in Lexical and Terminological Resources: An Analysis for Natural Language Processing Purposes. In: Proceedings of the eLex 2013 conference, pp. 338–357 (2013).

28. Lew, R.: The Role of Syntactic Class, Frequency, and Word Order in Looking up English Multi-Word Expressions. Lexikos 22, 243–260 (2012).
29. Béjoint, H.: The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills. Applied Linguistics 2(3), 207–222 (1981).
30. Nastase, V., Szpakowicz, S.: Exploring noun-modifier semantic relations. In: Fifth International Workshop on Computational Semantics (IWCS-5), pp. 285–301 (2003).
31. Girju, R., Moldovan, D., Tatu, M., Andantohe, D.: On the semantics of noun compounds. Journal of Computer Speech and Language 19(4), 479–496 (2005).
32. Nakov, P., Hearst, M.: Using Verbs to Characterize Noun-Noun Relations. Artificial Intelligence Methodology Systems and Applications 4183, 233–244 (2006).
33. Nulty, P., Costello, F. J.: General and specific paraphrases of semantic relations between nouns. Natural Language Engineering 19(3), 357–384 (2013).
34. Condamines, A.: Corpus Analysis and Conceptual Relation Patterns. Terminology 8(1), 141–62 (2002).
35. Auger, A., Barrière, C.: Pattern-Based Approaches to Semantic Relation Extraction: A State-of-the-Art. Terminology 14(1), 1–19 (2008).
36. León-Araúz, P., San Martín, A., Faber, P.: Pattern-based Word Sketches for the Extraction of Semantic Relations. In: Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016), pp. 73–82 (2016).
37. León-Araúz, P., Reimerink, R.: Knowledge Extraction and Multidimensionality in the Environmental Domain. In: Proceedings of the Terminology and Knowledge Engineering (TKE) Conference 2010. Dublin City University, Dublin (2010).
38. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Ková, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography 1(1), 7–36 (2014).
39. Drouin, P.: Term extraction using non-technical corpora as a point of leverage. Terminology 9, 99–115 (2003).
40. Park, C., Allaby, M.: Dictionary of Environment & Conservation. 3rd edn. Oxford University Press, Oxford (2013).
41. Jelley, N.: A Dictionary of Energy Science. 1st edn. Oxford University Press, Oxford (2017).
42. Cleveland, C. J., Morris, C.: Dictionary of Energy. 2nd edn. Elsevier, Amsterdam/Oxford/Waltham (2015).
43. Maguire, P., Wisniewski, E. J., Storms, G.: A Corpus Study of Semantic Patterns in Compounding. Corpus Linguistics and Linguistic Theory 6(1), 49–73 (2010).
44. Cohen, B.: Lexique de cooccurrents. Bourse-conjoncture économique. Linguatech, Montreal (1986).
45. Lorente Casafont, M., Martínez Salom, M. A., Santamaría-Pérez, I., Vargas Sierra, C.: Specialized collocations in specialized dictionaries. In: Torner Castells, S., Bernal, E. (eds.) Collocations and other lexical combinations in Spanish: theoretical, lexicographical and applied perspectives, pp. 200–222. Routledge, Abingdon/New York (2017).
46. Cabezas-García, M., Faber, P.: The role of micro-contexts in noun compound formation. Neologica 11, 101–118 (2017).
47. Kageura, K.: A preliminary investigation of the nature of frequency distributions of constituent elements of terms in terminology. Terminology 4(2), 199–223 (1997).
48. Sager, J. C., Dungworth, D., McDonald, P. F.: English Special Languages. Principles and Practice in Science and Technology. Brandstetter Verlag, Wiesbaden (1980).
49. Bernier-Colborne, G., Drouin, P.: Evaluation of distributional semantic models: a holistic approach. In: Proceedings of the 5th International Workshop on Computational Terminology (CompuTerm2016), pp. 52-61 (2016).