Semantic-based Retrieval of Complex Nominals in Terminographic Resources

Melania Cabezas-García, Juan Carlos Gil-Berrozpe
University of Granada
E-mail: melaniacabezas@ugr.es, jcgilberrozpe@ugr.es

Abstract

In English, specialized concepts frequently take the form of complex nominals (CNs), e.g. *greenhouse gas emissions*. The syntactic-semantic complexity of these multi-word terms (MWTs) highlights the need for a systematic treatment in specialized resources. This paper explores how semantic patterns in CNs can be applied to retrieve information in terminological knowledge bases, specifically in EcoLexicon (http://ecolexicon.ugr.es), the practical application of Frame-based Terminology (Faber 2012). For that purpose, we extracted the 250 most frequent CNs in an English wind power corpus. Structural disambiguation was performed to identify the internal groups linked by semantic relations. *Ad-hoc* semantic categories were then assigned to the elements of CNs with a view to studying the formation of CNs and allowing semantic-based queries in EcoLexicon. Then, the semantic relations between the CN constituents were analyzed by means of knowledge patterns and paraphrases. Our preliminary results showed recurrent semantic patterns in CN formation. This facilitates the inference of semantic relations, which is one of the main difficulties of MWTs. Furthermore, a semantic-based view of the CN module of EcoLexicon is presented, which allows different types of semantic query.

Keywords: complex nominals, semantic patterns, semantic categories, terminological knowledge bases

1 Introduction

In English, specialized concepts frequently take the form of complex nominals (CNs), e.g. *greenhouse gas emissions*. These multi-word terms (MWTs) are characterized by their syntactic and semantic complexity, which underlines the need for their systematic treatment in lexicographic and terminographic resources (Cabezas-García and Faber 2017a). Before describing CNs, their meaning must be specified, usually with a set of semantic relations (e.g. in *wind erosion*, wind *causes* erosion) (Rosario et al. 2002; Girju et al. 2005; *inter alia*). Nakov and Hearst (2006) propose the use of verb paraphrases (e.g. *olive oil* is an *oil that comes from/is squeezed from olives*). This indicates that semantic patterns in CN formation (Maguire et al. 2010) should be addressed before including CNs in specialized resources, e.g. *beach erosion* represents a LANDFORM (*beach*) that is the *patient_of* a LOSS PROCESS (*erosion*). These semantic patterns are closely related to the underlying conceptualization of the domain and the semantic relations in CNs, whose non-specification usually poses comprehension problems.

This paper explores how semantic patterns in CNs can be applied to retrieve information in terminological knowledge bases. EcoLexicon (http://ecolexicon.ugr.es) (Faber et al. 2016) is a terminological knowledge base on environmental science, which is currently being redefined to describe CNs. The design of its new phraseological module focuses on the inclusion of different kinds of data regarding CNs, such as the combinations derived from a given term, definitions, translations in English and Spanish, syntactic combinations and semantic co-occurrences. For this purpose, we extracted the 250 most frequent CNs in an English wind power corpus. Structural disambiguation was performed to identify the internal groups between which semantic relations had to be established. *Ad-hoc* semantic categories (e.g.

LANDFORM, WATER BODY) were then assigned to the internal groups in CNs with a view to conceptually analyzing the formation of CNs and allowing semantic-based queries in EcoLexicon. The assignment of semantic categories also facilitated the identification of the semantic relations between CN constituents with paraphrases and knowledge patterns (Meyer 2001; Marshman 2006). Thus, this search allowed users to perform queries based on terms, semantic categories, and semantic relations present in CNs.

The rest of this paper is organized as follows. Section 2 provides a review of the most relevant theoretical aspects concerning CNs, as well as a description of how they are usually dealt with in various terminographic resources. Section 3 explains the methodology applied to this study, focusing on the extraction of CNs and their semantic analysis. Section 4 shows the results of this semantic analysis and displays the most relevant semantic patterns in our wind power corpus. In Section 5 we describe the semantic-based view of the CN module of EcoLexicon. Finally, Section 6 gives the conclusions that can be derived from this research.

2 Complex Nominals in Linguistic Resources

2.1 Complex Nominals and their Semantics

Complex nominals (CNs) are expressions with a head noun modified by one or more elements, usually nouns or adjectives (Levi 1978), e.g. *significant wave height*. CNs reflect the preference of Germanic languages for condensed structures in which economy of expression overcomes clarity of expression (Štekauer et al. 2012). In particular, stacking modifiers on the left of the head (i.e. pre-modification) is the most frequent formation pattern in English (e.g. *water vapor*), although prepositional modification of the head (i.e. post-modification) can also be found, often combined with premodifiers (e.g. *general circulation of the atmosphere*). This permits the expression of specialized knowledge in semantically condensed structures (Sager et al. 1980; Sanz-Vicente 2012).

According to Lauer and Dras (1994), CNs can pose problems, namely insofar as their identification in texts, and their syntactic and semantic analysis. First, identifying them in texts can be challenging since they are often formed by general language words that may not be recognized as part of the CN. Furthermore, CNs are composed of more than two constituents, which also entails the need to disambiguate their internal structure (e.g. *renewable* [*energy source*]). Accessing the semantics of CNs is not an easy task, since these MWTs convey a relation between two or more elements that requires further knowledge (Maguire et al. 2010; Smith et al. 2014). Additional difficulties can arise when translating CNs, since term formation has different patterns in different languages (e.g. the pre-modification in English is not possible in Romance languages). Finally, the representation of CNs in linguistic resources has also been the subject of debate since they should be treated more systematically (Cabezas-García and Faber 2017a).

The semantic analysis and representation of CNs are the focus of this study. The semantics of CNs can be analyzed in terms of predicate nominalization (e.g. a system *transmits* AC > *AC transmission system*) and predicate deletion (an industry *produces* wind power > *wind power industry*) (Levi 1978). These concealed predicates largely correspond to semantic relations (e.g. *AC transmission system* > a system *has_function* [transmitting] AC; and *wind power industry* > an industry *has_function* [producing] wind power). Semantic relations are essential for the conceptualization of CNs, because they are part of the micro-context of these MWTs. The semantics of the head of a CN determines the conceptual nature of its modifiers (e.g. *energy* is usually referred to in terms of the resource that is used for its production, as in *wave energy, solar energy, wind energy*, etc.). In this micro-context, the internal semantic relations in CNs are relevant because they show how the elements of the micro-context are linked (e.g. energy *caused_by* waves/sun/wind).

2.2 Representation of Complex Nominals

As for the representation of CNs, a major problem is that they are rarely defined (e.g. *Vocabulaire et cooccurrents de la comptabilité* [Caignon 2001]). Furthermore, they are often listed alphabetically (e.g. *Dictionary of Energy* [Cleveland & Morris 2015]; *A Dictionary of Translation Technology* [Chan 2004]). However, a representation of domain structure should reflect the relations of the CN with other terms (e.g. *Elsevier's Dictionary of Medicine Spanish-English English-Spanish* [Hidalgo 2014]). There are also resources that show the modifiers and their possible heads in different lines, instead of including the entire CN (e.g. *Diccionario técnico inglés-español español-inglés* [Beigbeder 2006]). Other resources display a modifier along with a list of different and not necessarily related CNs that contain the modifier (e.g. *Routledge French Technical Dictionary* [Arden 2013]), such as *complementary angle*, *complementary code* and *complementary color* (Arden 2013: 141).

For rapid knowledge acquisition (Faber 2012), a conceptual approach seems to be best, since it reflects domain structure, facilitates understanding, and provides the basis for translation (Cabezas-García & Faber 2017a). An increasing number of resources take a semantic approach (e.g. FrameNet [Baker et al. 1998]; WordNet [Fellbaum 2005]; VerbNet [Kipper-Schuler et al. 2006]; DiCouèbe [OLST 2013]; DicoInfo [OLST 2018]; DicoEnviro [OLST 2018]; *inter alia*). However, the conceptual information of units is often based on semantic roles (e.g. AGENT, PATIENT), while other relevant information is not recorded. This is the case of semantic categories (e.g. SUBSTANCE, LANDFORM), which permit generalizations about concepts and conceptual organization as well as semantic-based queries. Resources that allow such queries include the *Pattern Dictionary of English Verbs* (Hanks 2014) where a list of semantic categories can be queried to obtain the verbs with which they co-occur, or vice versa (e.g. the COLOR category establishes verbs such as *shade, paint* or *dye*). The *DELAC Dictionary of Serbian Compounds* (Krstev et al. 2006) also allows the retrieval of compounds based on semantic categories (e.g. the search for the +Zool category shows Serbian compounds including an animal). Furthermore, in the new version of the *Diccionario de términos médicos* of the Spanish *Real Academia Nacional de Medicina* (2012) users can search for terms included in definitions (e.g. if *piel* [*skin*] is queried, entries including *skin* in their definitions will be shown). Additionally, one of the views of the *Diccionario Ideológico de la lengua española* (Casares 2013) organizes concepts in semantic categories (e.g. the MOLLUSK category includes *clam*, *oyster*, *squid*, etc.). Therefore, frame-like representations (Fillmore 1985) (e.g. EcoLexicon) are a good option since they combine conceptual and linguistic representations (L'Homme 2014).

3 Complex Nominal Extraction and Semantic Analysis

For our study, we compiled a corpus of English specialized texts on wind power, composed of approximately 1.8 million words. The corpus, which consisted of specialized articles and PhD dissertations, was uploaded to Sketch Engine (http://www.sketchengine.co.uk) (Kilgarriff et al. 2004), a corpus analysis tool that is used for CN extraction and semantic analysis. The different forms of English CNs were obtained with Corpus Query Language (CQL) regular expressions. Two lists of 1,000 CNs were extracted. One was composed of CNs formed by pre-modification[1] of the head by nouns, adjectives and/or adverbs (e.g. *tip speed*), and a second list composed of CNs formed by post-modification[2] of the head by a prepositional phrase (e.g. *angle of attack*). Both lists were combined and the 250 most frequent CNs were selected. All CNs that were part of a longer CN or which belonged to other domains such as Statistics, Mathematics, or Economics, were discarded.

1 The regular expression used for such query was the following one: [tag="N.*|JJ.*|RB.*"]{1,}[tag="N.*"].

2 The regular expression used for such query was the following one: [tag="N.*|JJ.*|RB.*"]{0,}[tag="N.*"]{1,}[tag="IN"]{1} [tag="DT"]?[tag="JJ.*|RB.*"]{0,}[tag="N.*"]{1,}.

The first step in semantic analysis was the internal structure disambiguation of CNs with more than two constituents (e.g. [*wind power*] [*generation system*]). Indicators of the internal groupings in CNs are the conceptual nature of the possible combination (e.g. *wind power* is a concept in linguistic resources), the existence of monolexical variants or equivalents in another language (e.g. *generation system* is also referred to as *generator*), or the co-occurrence of this CN with different modifiers (e.g. *conventional generation system*, *diesel generation system*, or *electric generation system*).

Conceptual or semantic categories were then used to specify the semantics of CNs, taking into account different categorization levels (Murphy & Lassaline 1997) and conceptual similarity (Hahn & Chater 1997). This categorization was based on concept definitions as well as on the contextual information in the wind power corpus. After determining the characteristics shared by concepts, categories were manually established and organized hierarchically from general to specific. In this way, the 250 CNs were classified in 47 conceptual categories, distributed in four categorization levels. The most general level was composed of the three starter ontological categories: ENTITY (i.e. physical and mental objects), PROCESS (i.e. events extending over time and involving different parties), and ATTRIBUTE (i.e. characteristics of entities or processes). The CNs were then classified in one of four more specific levels. For example, CNs with a generic meaning could only be categorized in one level (e.g. *energy source* [ENTITY]), whereas CNs with a specific meaning could be categorized in all four levels (e.g. *carbon dioxide* [ENTITY > MATTER > SUBSTANCE > CHEMICAL SUBSTANCE]). However, for the sake of simplifying such categorization, we only refer to the most specific category (e.g. *carbon dioxide* [CHEMICAL SUBSTANCE]). Furthermore, this categorization was applied to the CNs in three stages: (1) to the whole CN; (2) to its internal groupings; (3) to its individual constituents. For instance, *offshore wind turbine* as a whole was classified as an INSTRUMENT. Semantic categories were then assigned to the internal groupings in the CN (*offshore* [LOCATION] *wind turbine* [INSTRUMENT]), as well as to each of its constituent parts (*offshore* [LOCATION] *wind* [WIND MOVEMENT] *turbine* [INSTRUMENT]).

The next step involved specifying the semantic relation between CN constituents. Knowledge patterns (KPs) were used to extract the internal relations in CNs in the form of KP-based sketch grammars (León-Araúz et al. 2016). Figure 1 shows an extract of the results of a query that targets the sentences annotated as word sketches between *power* and *plant*, where ws means word sketch, "power-n" and "plant-n" are the terms that have been annotated as part of a word sketch in the corpus; and "\"%w\".*" means any relation defined in the KP-based sketch grammars. As can be seen, these KPs show that the function of plants is power generation.

fresh water use. Conventional	***plants***	generate **power** from fossil fuels and nuclear materials, which
the "representative" wind	***plant***	produces zero **power** approximately 2000 h/y, and full power about
the wind source. In fact, as far as the amount of		**power** produced by the wind ***plant*** is small in
. In particular, the higher the fraction of		**power** produced by a power ***plant*** in comparison with the
following formula, where: P is the total active		**power** generated by the wind power ***plant*** ; Q is the total
by the wind power plant; Q is the total reactive		**power** generated by the wind power ***plant*** ; r is the
not be controlled, at the rated frequency, the		**power** produced by a wind power ***plant*** can be
and availability of the ***plant*** to generate		**power** that is expected at a particular period. The
in large, central power	***plants***	produce **power** at high voltage (up to 25,000 V). These

Figure 1: KPs between *power* and *plant* obtained with the following query: [ws("power-n","\"%w\" .*","plant-n")].

The analysis of semantic relations with KPs was complemented by searching for verb paraphrases in order to access the concealed or nominalized predicates in CNs (Levi 1978; Nakov & Hearst 2006, 2013), which add further semantic precision to semantic relations (Nakov & Hearst 2013; Cabezas-García & Faber 2017b). For instance, *power curve* was found to designate a curve that represents, calculates, simulates or provides the output power of a wind turbine, as shown in Figure 2.

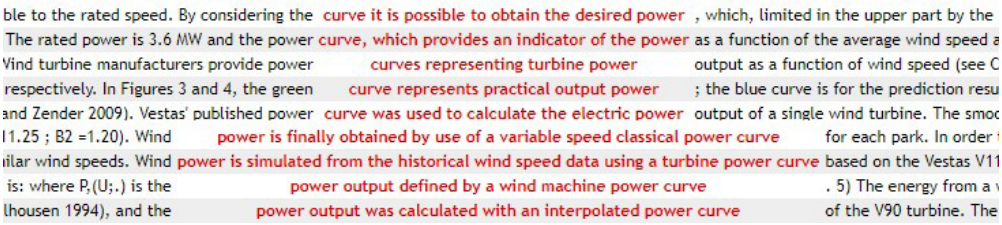

Figure 2: Verb paraphrases for *power curve* obtained with the following queries: [lemma="power"][]{0,10}[tag="V.*"][]{0,10}[lemma="curve"]within <s/> [lemma="curve"][]{0,10}[tag="V.*"][]{0,10}[lemma="power"]within <s/>.

However, verb paraphrases are not always easy to find. One reason for this is that CNs usually have concealed elements, whose omission complicates the extraction of verb paraphrases. Instead, many CNs include adjectives[3] or other units that refer to the hidden noun and thus are not easily linked to the head by means of predicates. For instance, in *renewable generation*, no verb joining *renewable* and *generation* was found. Therefore, we used free paraphrases, i.e. co-occurrences of the constituents of a CN in a sentence, as a way of accessing the meaning of those CNs whose semantics had not been ascertained by means of KPs or verb paraphrases. As shown in Figure 3, there was a missing element in *renewable generation* that complicated the extraction of KPs and verb paraphrases, namely the sources from which power is generated.

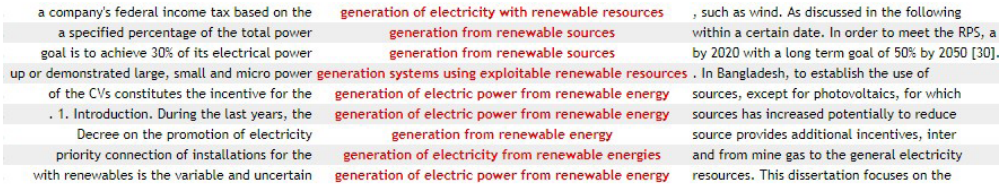

Figure 3: Free paraphrases for *renewable generation* obtained with the following query: [lemma="generation"][]{0,10}[word="renewable"][lemma!="generation"]within <s/>.

In summary, the non-specification of the semantic relation in CNs was addressed by combining different procedures – namely KPs, verb paraphrases and free paraphrases – which offered further semantic insights into these MWTs.

4 Semantic Patterns in Complex Nominal Formation

In CNs, two or more concepts converge. However, these combinations are not random, but are the result of semantic constraints (Štekauer 1998). In CNs, these semantic constraints take the form of micro-contexts, which refer to the opening of slots by the head (Maguire et al. 2010). These slots are filled by semantic categories and roles. The semantics of the head thus determines possible modifiers (Rosario et al. 2002). For instance, *emission* opens two slots. One of them is usually filled by the semantic category of SUBSTANCE, which has the role of PATIENT or the entity being emitted (as in *greenhouse gas emissions* and *CO_2 emission*). The second slot normally represents the category of SPACE, which has the role of DESTINATION or place where substances are emitted (as in *air emission* and *atmospheric emission*). The combination of both slots is also possible and produces longer CNs, such as *atmospheric emission of CO_2*.

3 Some adjectives have an underlying noun. In that case, they can be replaced by this noun, which facilitates the retrieval of verb paraphrases. For instance, in *fluvial sediment*, verbs linking *river* and *sediment* can be queried (e.g. *deposit*, *carry*, *transport*, etc.).

This study focused on semantic categories in CN formation because they capture regularities and allow generalizations (Hoey 2005). The CNs in our study mostly designated the categories of QUANTITY (28 out of 250 CNs, e.g. *net load*), MAGNITUDE (21 CNs, *e.g.* wind speed), SYSTEM (19 CNs, e.g. *distribution system*), ACTIVITY (19 CNs, e.g. *wind project*), ENERGY (18 CNs, e.g. *renewable energy*), and INSTRUMENT (16 CNs, e.g. *offshore wind turbine)*. In addition, the most frequent categories used to form CNs were the following: (i) WIND MOVEMENT (67 times, e.g. *wind resource*); (ii) ENERGY (64 times, e.g. *power production*); (iii) QUANTITY (67 times, e.g. *feed-in tariff*); (iv) INSTRUMENT (30 times, e.g. *rotor disc*); (v) ACTIVITY (29 times, e.g. *system operation*); (vi) SYSTEM (28 times, e.g. *power generation system*); (vii) MAGNITUDE (22 times, e.g. *water depth*); (viii) LOCATION (20 times, e.g. *offshore market*). The combination of these categories reflected recurrent semantic patterns in the formation of CNs, which can be useful for the inference of the semantic relations between the components of CNs (Rosario et al. 2002; Maguire et al. 2010; Smith et al. 2014; Cabezas-García & León-Araúz 2018). Table 1 shows the most productive semantic patterns for term formation in our study, as well as some of the CNs derived from these conceptual combinations and the semantic relation between their constituents.

Table 1: Semantic patterns for term formation in the domain of wind power.

Semantic Pattern	Number of CNs	Examples	Semantic Relation
ENERGY + ACTIVITY	8	*wind-energy project* *wind power development* *wind power forecast*	*has_function* *has_patient*
ORIGIN + ENERGY	8	*electrical power* *kinetic energy* *renewable energy*	*caused_by* *has_attribute_origin*
ENERGY + FORMATION	7	*wind power generation* *energy production* *electricity generation*	*has_result*
ENERGY + SYSTEM	7	*wind power generation system* *electric power system* *energy storage system*	*has_function*
LOCATION + ACTIVITY	5	*offshore wind market* *offshore wind industry* *offshore project*	*has_attribute_location*
QUANTITY + QUANTITY	5	*total installed capacity* *net load* *load demand*	*has_attribute_quantity* *has_patient* *represents*
WIND MOVEMENT + FACILITY	5	*wind power plant* *wind farm* *wind park*	*uses_ resource*

As can be observed, regularities were found between CNs formed by the same conceptual categories and the semantic relations encoded. Namely, the following semantic patterns established the same semantic relation in all of their CNs: ENERGY + FORMATION (*has_result*); ENERGY + SYSTEM (*has_function*); LOCATION + ACTIVITY (*has_attribute_location*); and WIND MOVEMENT + FACILITY (*uses_ resource*). However, other patterns had more than one semantic relation, which were indicative of semantic constraints. For example, ENERGY + ACTIVITY activated two semantic relations: *has_patient*, when the CN was formed by predicate nominalization (e.g. *wind power development*, *wind power forecast*), and *has_function*, when the CN was formed by predicate deletion (e.g. *wind-energy project*, *electricity market*). This highlighted the role of predicates in CN formation (Levi

1978; Cabezas-García & Faber 2017b). As for ORIGIN + ENERGY, the *caused_by* relation was mostly established (as in *electrical energy* and *kinetic energy*) though the *has_attribute_origin* relation was found in CNs with *renewable* (e.g. *renewable energy*, *renewable electricity*), which actually modifies the source from which energy can be produced. Finally, QUANTITY + QUANTITY tended to encode the *has_attribute_quantity* relation, when it included adjectives determining how the quantity must be understood (as in *total installed capacity* and *net load*). Additionally, it can activate the *represents* relation (e.g. *capacity credit*), or the *has_patient* relation, when the CN is formed by predicate nominalization (as in *load demand*). These semantic relations correlated with the most frequent relations in all of the CNs of our study, which were *has_patient* (in 35 CNs), *has_function* (in 32 CNs), and *represents* (in 18 CNs). The broad ontological distinction of conceptual categories between entities, processes, and attributes also shed light on the relevance of certain relations. Namely, the *has_function* relation was prevalent in entities (29 out of 177 CNs), whereas *has_patient* was the most frequent relation in processes (29 out of 60), and *attribute_state* was the most recurrent relation in the attributes (three out of 13 CNs). This was not surprising, since entities are usually described in terms of their use; processes often emphasize the concept that receives the action; and attributes specify properties, such as the state. Therefore, analyzing the semantic aspects of the formation of CNs can help to identify recurrent patterns in the production of these MWTs. The application of this semantic information to knowledge resources facilitates understanding of concepts, as well as domain and frame reconstruction (Sager et al. 1980).

5 Phraseological Module of EcoLexicon: The Semantic Combinations View

5.1 EcoLexicon and its Phraseological Module

EcoLexicon (Faber et al. 2016), based on the theoretical premises of Frame-based Terminology (Faber 2012), is a multidimensional terminological knowledge base on environmental science. It targets user knowledge acquisition through different types of multimodal and contextualized information to respond to cognitive and communicative needs. Its public is any user group interested in broadening their knowledge of the environment for text comprehension and/or generation (e.g. environmental experts, technical writers, translators). This resource is currently available in English and Spanish, though five more languages (German, Modern Greek, Russian, French and Dutch) are being gradually implemented. To date, its database consists of a total of 3,950 concepts and 21,720 terms.

EcoLexicon has a visual interface with different modules for conceptual, linguistic, and graphical information. Because of the importance of phraseological information in terminological resources, a new phraseological module is currently under construction. Although the module describes verbal collocations and CNs, this paper only deals with the CN submodule. This submodule contains four views: (1) the Modifiers + Head view: (2) the EN-ES view; (3) the Syntactic Combinations View; and (4) the Semantic Combinations view. The Modifiers + Head view focuses on CN formation and allows users to search a list of CNs that contain a given term (e.g. if the user looks for *turbine*, the query will show a list of CNs including *horizontal axis wind turbine*, *fixed speed wind turbine*, *wind turbine foundation*, etc.). Moreover, the search tool in the EN-ES view offers bilingual results in English and Spanish (e.g. the user can search *wind turbine* and then find *aerogenerador*, or vice versa). The Syntactic Combinations view offers advanced queries according to the part of speech (e.g. if the user looks for N+N+N, the results will include *sea level rise*, *incident wave height*, *sediment transport rate*, etc.). Regarding the Semantic Combinations view, it allows users to perform queries based on the semantics of CNs in terms of their conceptual categories and internal semantic relations, as will be shown in Section 5.2.

5.2 The Semantic Combinations View

In the Semantic Combinations view of the CN submodule, users can perform a simple or an advanced query. Figure 4 shows the query screen and the results screen of the simple query "wind power". The simple query box can be used to perform a proximity search. As shown in the results screen, the system automatically converts the user's search into a query expression ("wind[TERM] AND power[TERM]") and displays a list of results in EcoLexicon that include the terms (e.g. *offshore wind power*, *wind power generation*). The alphabetically listed results show the CN divided into its components with the bracketing, along with the conceptual category of each constituent (below, in red) and the internal semantic relation that links these components (on the right, in blue). For instance, the first result in Figure 4 is *offshore wind power*, which is described as *offshore* [LOCATION] *wind power* [ENERGY] and contains the semantic relation *has_attribute_location*. To the right edge of each result, there is a "+" symbol that displays a box with additional information about the entry: definition, semantic information, corpus concordances, verbal collocations, access to the full entry in EcoLexicon, and notes. The definition describes the concept. Additionally, the semantic information section provides conceptual information, such as the semantic category of the CN, each of its constituents, and the internal relations in CNs formed by more than two terms (e.g. in *wind power forecast*, forecast *has_patient* wind power, and power *is_caused_by* wind). The corpus concordances allow users to access the EcoLexicon corpus in Sketch Engine. Furthermore, the verbal collocations option links the two sections of the phraseological module, namely CNs and verbal collocations, because many collocations have a nominal equivalent in the form of a CN (e.g. *a volcano erupts* > *volcano eruption*). Finally, the term entry in EcoLexicon provides synonyms of the CN, equivalents in different languages, images, conceptual networks, etc.

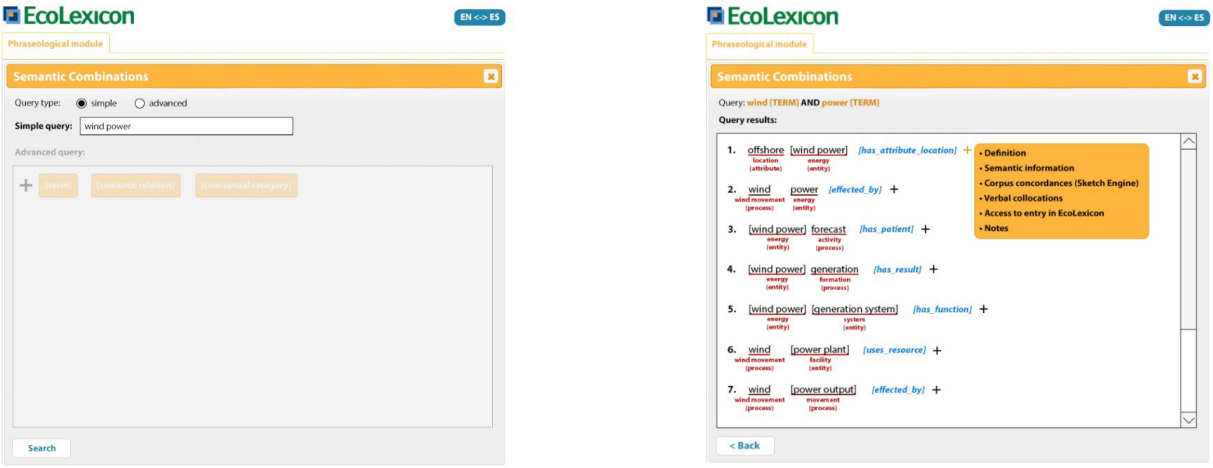

Figure 4: Simple query (left side) and results (right side) in the Semantic Combinations view using the following expression: wind[TERM] AND power[TERM].

The advanced query presents a series of particularities that allow users to perform more complicated searches. As shown in Figure 5, the advanced query is based on three elements: (i) terms; (ii) semantic relations; (iii) conceptual categories. By clicking on the orange bubbles next to the "+" symbol, the user can add as many elements to the query as they want and in any order, since this query allows for free element combination (e.g. CATEGORY + CATEGORY, term + CATEGORY, CATEGORY + relation + CATEGORY, etc.). In the same way, any element can also be deleted. The term bubble has a free text box to type anything, whilst the semantic relation and the conceptual category bubbles display a picklist showing all the relations or categories contained in EcoLexicon. However, it is also possible

to choose the options “ANY RELATION” or “ANY CATEGORY”. In fact, displaying all the possibilities with a picklist is the simplest way for users to easily find and choose the most suitable option for their query. In addition, each bubble contains an “ADD” and an “OR” button, which are useful if the user wants to look for more than one term, relation and/or category found in the same position.

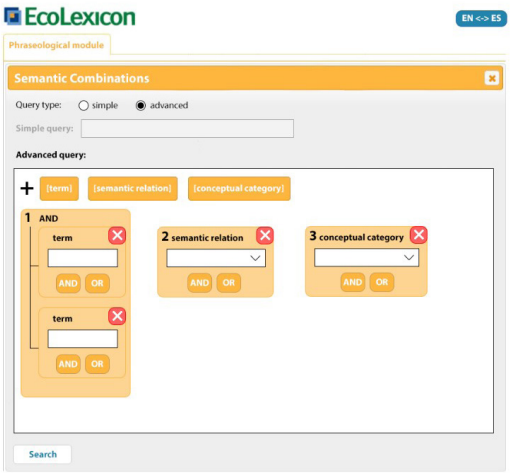

Figure 5: Advanced query in the Semantic Combinations view, showing the free element combination.

Figure 6 shows the query screen and the results screen of the advanced query “offshore[TERM] + [ANY CATEGORY]”. In order to perform this search, the user must select the option “advanced” next to “Query type”, and this will activate the advanced query box, where the user will then create a term bubble in order to type “offshore”, and a conceptual category bubble in order to select “ANY CATEGORY”. As a consequence, this expression displays a series of results that include *offshore capacity*, *offshore market* and *offshore project*, to name a few examples. As in the simple query, the results are also listed alphabetically, showing the conceptual categories of the groupings in the CN and the internal semantic relation.

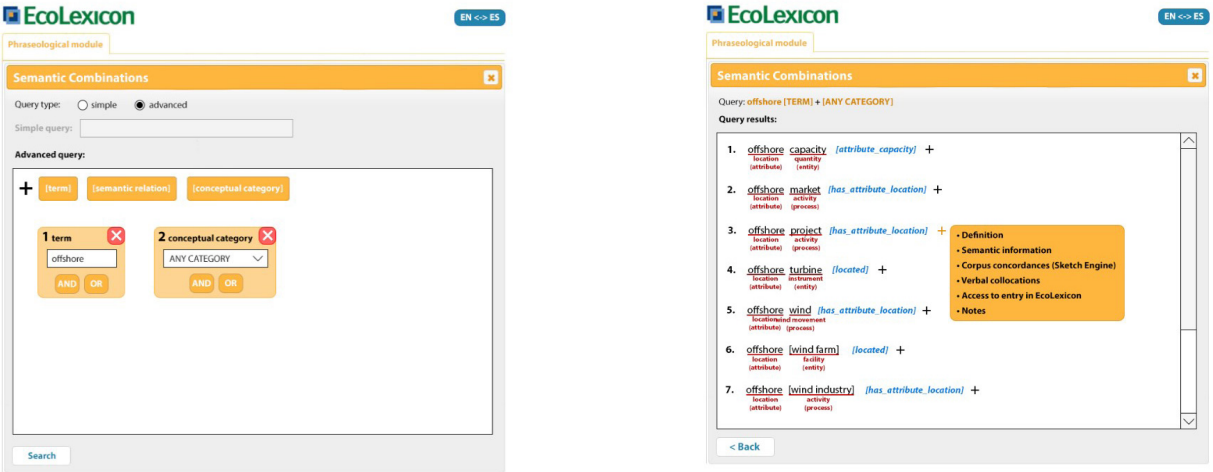

Figure 6: Advanced query (left side) and results (right side) in the Semantic Combinations view with the following expression: offshore[TERM] + [ANY CATEGORY].

Finally, another feature of the results screen is the possibility of offering similar results based on the search criteria. In Figure 7 is shown an advanced query of CNs made of two conceptual categories [“energy (entity)[CATEGORY] + formation (process)[CATEGORY]”], which has six results (e.g.

electricity generation, *electricity production*, *energy generation*). In this way, the Semantic Combinations view allows the user to see more results by clicking on "+ Show similar results", which would display those CN that meet the search criteria entered by the user, but in reverse order (e.g. *generated power*, *generated energy*, *produced power*). Accordingly, what matters in CN formation is meaning and content, not shape (Štekauer 1998).

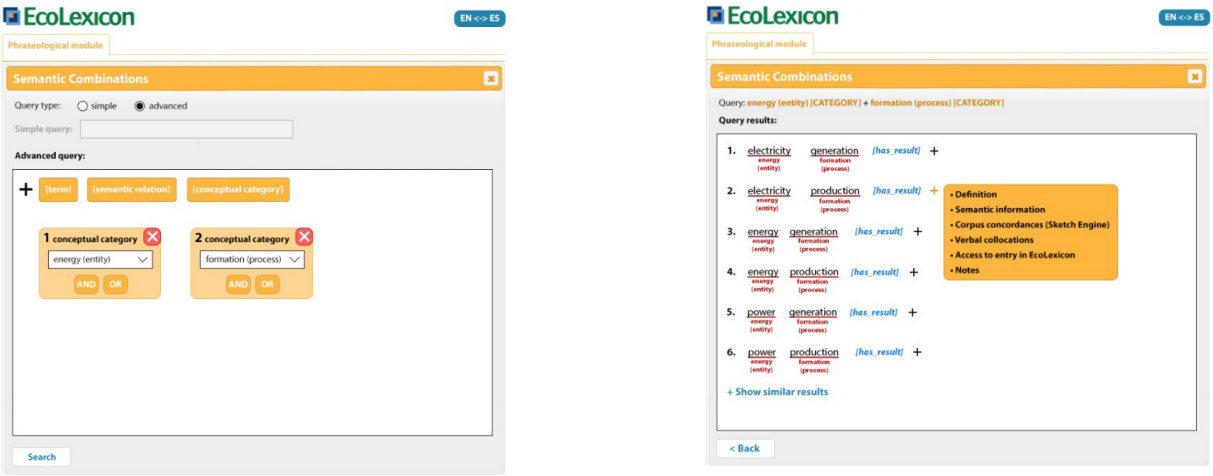

Figure 7: Advanced query (left side) and results (right side) in the Semantic Combinations view using the following expression: energy (entity)[CATEGORY] + formation (process)[CATEGORY].

This Semantic Combinations view enhances the CN submodule by adding a conceptual approach to the queries that users can perform inside the new phraseological module. Since specialized knowledge is not conceptualized in isolation, but rather as part of a context or frame (Faber 2012), the semantic and conceptual features contained in this tool help to describe concept structure and interrelationships within the environmental domain.

6 Conclusion

Complex nominals (CNs) pose different problems, such as their identification in texts, their syntactic and semantic description, their translation into different languages, their representation in lexicographic and terminographic works, *inter alia*. Accordingly, we carried out a semantic analysis of a set of CNs extracted from a wind energy corpus. This analysis was performed by assigning conceptual categories both to the CNs as a whole and to their constituent parts. The semantic relations within the CNs were also described. As part of this process, the most recurring knowledge patterns in CN formation were detected with a view to applying this semantic information to the CN module in EcoLexicon. In this way, users could perform queries based on the meaning of these linguistic units.

Our results showed that the constituent concepts of CNs did not combine randomly, but rather as a result of semantic constraints. In particular, within the wind energy domain the most frequent patterns of CN formation were the following combinations: ENERGY + ACTIVITY, ORIGIN + ENERGY, ENERGY + FORMATION, ENERGY + SYSTEM, LOCATION + ACTIVITY, QUANTITY + QUANTITY, and WIND MOVEMENT + FACILITY. Since certain semantic relations tended to be established between these category pairs, this helped to infer the internal semantic relations in the CNs composed of these categories. Therefore, the existence of a certain relation in a category pair indicated that most of the CNs with these same categories would also have the same internal semantic relation (Rosario et al. 2002; Maguire et al. 2010; Smith et al. 2014; Cabezas-García & León-Araúz 2018).

Our ultimate objective was to apply the semantic information extracted from CNs to one of the new views of the CN module in EcoLexicon: the Semantic Combinations view, which is complemented by the Modifiers + Head view (showing CNs from a given term); the EN-ES view (allowing users to perform bilingual queries); and the Syntactic Combinations view (offering the possibility to combine syntactic categories in order to obtain CNs with specific syntactic patterns). The Semantic Combinations view permits users to enter different search elements, namely terms, semantic relations and conceptual categories, and combine them in any order or number. Examples of possible queries include "energy (entity)[CATEGORY] + formation (process)[CATEGORY]" (e.g. *electricity generation*, *energy production*), "offshore[TERM] + [ANY CATEGORY]" (e.g. *offshore turbine*, *offshore wind farm*), and "[ANY CATEGORY] + has_function [RELATION] + [ANY CATEGORY] (e.g. *control system*, *power plant*).

In conclusion, this new tool based on semantic information is adapted to the new tendencies in linguistic resources (Krstev et al. 2006; RANM 2012; Casares 2013; Hanks 2014), where semantics plays a key role. EcoLexicon, a resource characterized by its emphasis on conceptualization and knowledge structuration, will be enhanced with the creation of a specific section for CNs, allowing users to perform a wide range of queries. Although the set of semantic categories may vary in different domains, the semantic-based queries presented in this paper can be implemented in any electronic resource with a view to offering an enhanced user experience.

In future research, we plan to annotate every CN in EcoLexicon based on its domain or subdomain within the environment, as well as their annotation based on their semantic roles (Cabezas-García & San Martín 2017). Furthermore, semi-automatic annotation of CNs will be explored and user feedback in the CN module of EcoLexicon will be assessed. Additionally, since hyponymy is the semantic relation that is at the core of CN formation, further research will also include the analysis of the hyponymic nuances within CNs and their relation to the decomposition of hyponymy into subtypes (Gil-Berrozpe et al. 2017).

References

Arden, Y. (2013). *Routledge French Technical Dictionary / Dictionnaire technique anglais*. London/New York: Routledge.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING '98)*, 10–14 August 1998, pp. 86–90. Université de Montréal, Canada.

Beigbeder, F. (2006). *Diccionario técnico inglés-español español-inglés*. Madrid: Díaz de Santos.

Cabezas-García, M., Faber, P. (2017a). A Semantic Approach to the Inclusion of Complex Nominals in English Terminographic Resources. In R. Mitkov (ed.), *Computational and Corpus-Based Phraseology*, pp. 145–159. Cham: Springer.

Cabezas-García, M., Faber, P. (2017b). Exploring the Semantics of Multi-word Terms by Means of Paraphrases. In M.A. Candel-Mora, C. Vargas-Sierra (eds.), *Temas actuales de terminología y estudios sobre el léxico*, pp. 193–217. Granada: Comares.

Cabezas-García, M., León-Araúz, P. (2018). Towards the Inference of Semantic Relations in Complex Nominals: A Pilot Study. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, 7-12 May 2018, pp. 2511-2518. Miyazaki, Japan: ELRA.

Cabezas-García, M., San Martín, A. (2017). Semantic annotation to characterize contextual variation in terminological noun compounds: a pilot study. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 4 April 2017, pp. 108–113. Valencia: Association for Computational Linguistics.

Caignon, P. (2001). *Vocabulaire et cooccurrents de la comptabilité*. Montréal: Linguatech.

Casares, J. (2013). *Diccionario ideológico de la lengua española (de la idea a la palabra y de la palabra a la idea)*. Madrid: Gredos.

Chan, S.W. (2004). *A Dictionary of Translation Technology*. Hong Kong: Chinese University Press.

Cleveland, C., Morris, C. (2015). *Dictionary of Energy*. Amsterdam: Elsevier.

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

Faber, P., León-Araúz, P., and Reimerink, A. (2016). EcoLexicon: new features and challenges. In I. Kernerman., I. Kosem Trojina, S. Krek, and L. Trap-Jensen (eds.), *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, pp. 73–80. Portorož, Slovenia.

Fellbaum, C. (2005). WordNet and wordnets. In K. Brown et al. (eds.), *Encyclopedia of Language and Linguistics*, pp. 665–670. Oxford: Elsevier.

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

Gil-Berrozpe, J.C., León-Araúz, P., and Faber, P. (2017). Specifying Hyponymy Subtypes and Knowledge Patterns: A Corpus-based Study. In *Proceedings of the Fifth International Conference on Electronic Lexicography in the 21st Century (eLex 2017)*, 19–21 September 2017, pp. 63–92. Leiden, The Netherlands.

Girju, R., Moldovan, D., Tatu, M., and Andantohe, D. (2005). On the semantics of noun compounds. In *Computer Speech and Language*, 19(4), pp. 479–496.

Hahn, U., Chater, N. (1997). Concepts and Similarity. In K. Lamberts and D. Shanks (eds.), *Knowledge, Concepts, and Categories*, pp. 93–131. Cambridge (MA)/London: MIT Press.

Hanks, P. (2014). *Pattern Dictionary of English Verbs*. Accessed at: http://pdev.org.uk/ [27/03/2018].

Hidalgo, A. (2014). *Elsevier's Dictionary of Medicine Spanish-English English-Spanish*. Amsterdam: Elsevier.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Abingdon: Routledge.

Kilgarriff, A., Rychlý, P., Smrž, P. and Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress*, 6–10 July 2004, pp. 105–116. Lorient, France.

Kipper-Schuler, K., Korhonen, A., Ryant, N., and Palmer, M. (2006). Extending VerbNet with Novel Verb Classes. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, 22–28 May 2006, pp. 1027–1032. Genoa, Italy: ELRA.

Krstev, C., Vitas, D., and Savary, A. (2006). Prerequisites for a Comprehensive Dictionary of Serbian Compounds. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala (eds.), *Advances in Natural Language Processing: Lecture Notes in Computer Science*, 4139, pp. 552–563. Berlin/Heidelberg: Springer.

L'Homme, M.C. (2014). Terminologies and taxonomies. In J.R. Taylor (ed.), *Handbook of the Word*. Oxford: Oxford University Press.

Lauer, M., Dras, M. (1994). A probabilistic model of compound nouns. In *Proceedings of the 7th Australian Joint Conference on Artificial Intelligence*, pp. 474–481. Singapore, Singapore.

León-Araúz, P., San Martín, A., and Faber, P. (2016). Pattern-based Word Sketches for the Extraction of Semantic Relations. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm 2016)*, 12 December 2016, pp. 73–82. Osaka, Japan.

Levi, J. (1978). *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

Maguire, P., Wisniewski, E.J., and Storms, G. (2010). A corpus study of semantic patterns in compounding. In *Corpus Linguistics and Linguistic Theory*, 6(1), pp. 49–73.

Marshman, E. (2006). 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, Canada.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework. In D. Bourigault, C. Jacquemin, and M.C. L'Homme (eds.), *Recent Advances in Computational Terminology*, pp. 279–302. Amsterdam/Philadelphia: John Benjamins.

Murphy, G.L., Lassaline, M.E. (1997). Hierarchical Structure in Concepts and Basic Level of Categorization. In K. Lamberts and D. Shanks (Eds.), *Knowledge, Concepts, and Categories*, pp. 93–131. Cambridge (MA)/London: MIT Press.

Nakov, P., Hearst, M. (2006). Using Verbs to Characterize Noun-Noun Relations. In *Artificial Intelligence Methodology Systems and Applications*, 4183, pp. 233–244.

Nakov, P., Hearst, M. (2013). Semantic Interpretation of Noun Compounds Using Verbal and Other Paraphrases. In *ACM Transactions on Speech and Language Processing*, 10(3), pp. 1–51.

Observatoire de Linguistique Sens-Texte (OLST) (2013). *DiCouèbe: Dictionnaire en ligne de combinatoire du Français*. Accessed at: http://olst.ling.umontreal.ca/dicouebe/index.php [27/03/2018].

Observatoire de Linguistique Sens-Texte (OLST) (2018). *DiCoInfo: Le dictionnaire fondamental de l'informatique et de l'Internet*. Accessed at: http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi [27/03/2018].

Observatoire de Linguistique Sens-Texte (OLST) (2018). *DiCoEnviro: Dictionnaire fondamental de l'environnement*. Accessed at: http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi [27/03/2018].

Real Academia Nacional de Medicina (RANM) (2012). *Diccionario de términos médicos*. Madrid: Panamericana.

Rosario, B., Hearst, M., and Fillmore, C. (2002). The Descent of Hierarchy, and Selection in Relational Semantics. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, 7–12 July 2002, pp. 247–254. Philadelphia, Pennsylvania, United States.

Sager, J.C., Dungworth, D., and McDonald, P.F. (1980). *English Special Languages: Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter Verlag.

Sanz-Vicente, L. (2012). Approaching secondary term formation through the analysis of multiword units: An English–Spanish contrastive study. In *Terminology*, 18(1), pp. 105–127.

Smith, V., Barratt, D., and Zlatev, J. (2014). Unpacking noun-noun compounds: interpreting novel and conventional food names in isolation and on food labels. In *Cognitive Linguistics*, 25(1), pp. 99–147.

Štekauer, P. (1998). *An Onomasiological Theory of English Word-Formation*. Amsterdam/Philadelphia: John Benjamins.

Štekauer, P., Valera, S., and Körtvélyessy, L. (2012). *Word-formation in the world's languages*. Cambridge/New York: Cambridge University Press.

Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. Funding was also provided by the FPU grants given by the Spanish Ministry of Education to both authors. Finally, we would like to thank the anonymous reviewers for their useful comments.