Multiword-term bracketing and representation in terminological knowledge bases

Pilar León-Araúz[1], Melania Cabezas-García[1], Pamela Faber[1]

[1] University of Granada, Granada, Spain
E-mail: pleon@ugr.es, melaniacabezas@ugr.es, pfaber@ugr.es

Abstract

Multiword terms (MWTs) are frequently consulted in terminological resources due to their structural, cognitive, and conceptual complexity. However, in most terminological resources they are not always well described, since they are often included as independent term entries with no information on how their constituents are related. An accurate management of MWTs of three or more constituents requires, as a first step, their structural disambiguation, also called bracketing. This paper examines MWT bracketing in order to enhance MWT representation by describing their structural dependencies. Based on NLP advances in bracketing, a protocol has been designed through corpus queries and evaluated according to the reliability of corpora and rules as well as the causes underlying failure. Automatising bracketing can help enhance the representation of MWTs in terminological knowledge bases, assisting both the terminologist and the final user, since making their relational structure explicit can favour knowledge acquisition.

Keywords: multiword term; bracketing; terminological knowledge base; terminology

1. Introduction

Multiword terms (MWTs) are frequently consulted in terminological resources due to their structural, cognitive, and conceptual complexity. However, in most terminological resources they are not always well described (Cabezas-García & Faber, 2017) or even well related to their heads and/or modifiers, since they are often included as independent term entries or unanalysed text strings, with no other information about their underlying relational structure. An accurate management and description of these terms requires an initial step that traditionally has not been among the main interests in terminology or specialised lexicography. This is bracketing, or structural disambiguation (Nakov & Hearst, 2005; Barrière & Ménard, 2014), which is necessary for the right interpretation of MWTs having three or more constituents, as in [*reactive power*] *consumption*. Knowledge of these dependencies facilitates MWT comprehension (i.e. reactive power is consumed instead of power consumption is reactive) and, consequently, translation. In Spanish, *consumo de potencia reactiva* would be the right choice instead of **consumo energético reactivo* or **consumo reactivo energético*, which would be the result of a misunderstood bracketing. The inclusion of MWTs in knowledge-based resources can benefit from their prior structural disambiguation, whose automatisation can assist both

terminologists and final users. For instance, their representation can be enhanced by placing them in relation to other concepts' entries based on their dependencies, such as their hypernyms (consumption), thus facilitating knowledge acquisition.

Cabezas-García and León-Araúz (2019) proposed a series of manual steps for the bracketing of MWTs based on their linguistic properties and advances from NLP. At a later stage, León-Araúz and Cabezas-García (in press) added new steps in the form of a bracketing protocol and designed queries in Sketch Engine (Kilgarriff et al., 2004) with a view to automatising bracketing and analysing the reliability of every rule in two different English corpora: (i) a wind power corpus (since the set of MWTs belonged to this domain); and (ii) the Open Access Journal (DOAJ) corpus. The Sketch Engine's API was used to automatically query the corpora. Based on the results of the queries, rules were collectively applied to provide the bracketing of a 103 three-term MWT set. Although the automatic protocol worked in 83% of the cases, the bracketing failed in both corpora for 13 MWTs, thus suggesting a more qualitative study of the results, by analysing those MWTs and looking for possible causes.

This paper examines MWT bracketing in order to enhance MWT representation by describing their structural dependencies. The bracketing errors in León-Araúz and Cabezas-García (in press) were analysed and our results showed that an in-depth analysis of bracketing errors can be used to enhance the protocol. In turn, using an automatised bracketing protocol can result in a more accurate representation of MWTs in terminological resources. In particular, a specific module for MWT representation (Cabezas-García, 2019; 2020) has been designed in the terminological knowledge base EcoLexicon (https://ecolexicon.ugr.es/), which will include bracketing-related information. The remainder of this paper is structured as follows: Section 2 describes the procedure followed in order to automatise bracketing and evaluate its output; Section 3 proposes a new module for the description of MWTs in a terminological knowledge base; and Section 4 draws some conclusions and future lines of research.

2. Multiword-term bracketing

NLP has particularly focused on the structural disambiguation of MWTs, given their difficulties for NLP systems (Lauer, 1995; Girju et al., 2005; Nakov, 2007; Barrière & Ménard, 2014). Likewise, their difficulties in translation (i.e. one of the ultimate purposes of term bases) have been widely acknowledged. However, to the best of our knowledge, none of these findings have been applied in the design of MWT entries in terminological knowledge bases. In Section 2.1 the main bracketing models found in the literature are briefly described. In Section 2.2 the protocol applied in this research, based on the latter, is explained and evaluated.

2.1 Bracketing models

NLP has proposed two main models for the bracketing of three-term MWTs: the adjacency and dependency models. The adjacency model (Marcus, 1980; Pustejovsky et al., 1993) takes an MWT p1 p2 p3 and compares if p2 is more related to p1 or p3. For that purpose, the number of occurrences of p1 p2 and p2 p3 are compared. For instance, in *renewable energy technology* there are more occurrences of *renewable energy* than of *energy technology*. Thus, a left-bracketing structure is adopted ([*renewable energy*] *technology*). The dependency model (Lauer, 1995) compares whether p1 is more strongly associated with p2 or p3. Therefore, the analysis does not start from the central term, as in the adjacency model, but rather from the first one to the left. When p1 is more strongly associated with p2 than to p3, there is a left bracketing ([*tip speed*] *ratio*). In contrast, when p1 is dependent on p3, there is a right bracketing (*mean* [*wind speed*]).

Along the same lines, Grefenstette (1994) states that dependency structures govern how MWTs can be shortened: "*civil rights activist* can be bracketed as [*civil rights*] *activist*, which can be shortened to *rights activist* but not to *civil activist*. On the other hand, *Yale medical library* is properly bracketed as *Yale* [*medical library*] which can then be reduced to *Yale library* or *medical library*, but not to *Yale medical*" (Grefenstette, 1994, p. 65). Based on Grefenstette's approach, for a right bracketing, both p2 p3 (*medical library*) and p1 p3 (*Yale library*) should be more frequent than p1 p2 (*Yale medical*), whereas for a left bracketing p1 p2 (*civil rights*) should be more frequent than p1 p3 (*civil activist*), the latter actually being the same rule as the one proposed by the dependency model.

Apart from these models, Nakov and Hearst (2005) propose a series of surface patterns (i.e. hyphens and slashes, possessive genitive, internal capitalisation, brackets, concatenation, internal inflection, etc.) as signs indicating an internal grouping. For example, *brain's stem cell* would suggest a right bracketing (*brain* [*stem cell*]) because of the possessive genitive, whereas *tyrosine kinases activation* would indicate a left bracketing ([*tyrosine kinase*] *activation*) because of the internal inflection. They also suggest that paraphrases are useful for identifying internal dependencies in MWTs. For instance, *health care reform* is left-bracketed because paraphrases separating those groups can be found, as in "reform *in* health care". Paraphrases can be either verbal or prepositional.

2.2 Bracketing automatisation

Based on the models and patterns above, a set of queries was designed and sent to Sketch Engine's API in order to retrieve and compare the frequencies of all the possible groupings contained in a list of 103 MWTs selected from the wind energy specialised domain (Section 2.2.1). As mentioned above, two corpora were used to

compare whether corpus size and/or domain specificity had an influence on the output: (i) a wind power corpus (WPC) specifically compiled for this research; and (ii) the Open Access Journal (DOAJ) corpus. The first consisted of wind energy specialised texts (i.e. scientific articles and PhD dissertations originally written in English) and had approximately three million words, whereas the latter covered all areas of science, technology, medicine, social science, and humanities and had approximately two billion words.

After that, the results were compared with the baseline (manually disambiguated by three annotators) and the protocol was evaluated in terms of rule and corpus reliability (Section 2.2.2). Since the protocol failed in both corpora for 13 MWTs, a more in-depth analysis was performed in order to discover the causes of protocol failure (Section 2.2.3) and improve it accordingly.

2.2.1 Preparing the dataset: queries and rules

The list of 103 MWTs, manually bracketed as a baseline, is included in Table 1.

offshore [wind farm]	installed [wind power]	[permanent magnet] generator
[tip speed] ratio	[wind turbine] design	[wind farm] project
[wind power] plant	[wind penetration] level	[wind speed] distribution
[wind power] generation	[wind speed] datum	[wind energy] production
[wind power] capacity	novel [wind turbine]	extreme [wind speed]
mean [wind speed]	domestic [hot water]	[wind tunnel] test
[wind power] production	[power generation] system	[wind energy] penetration
average [wind speed]	offshore [wind market]	offshore [wind park]
offshore [wind turbine]	[renewable energy] technology	[renewable energy] system
[renewable energy] source	[wind power] penetration	[wind speed] measurement
offshore [wind power]	[wind power] forecast	shrouded [wind turbine]
offshore [wind energy]	[wind power] development	[wind turbine] control
[wind energy] system	total [installed capacity]	micro [hydropower plant]
small [wind turbine]	conventional [power plant]	hybrid [wind farm]
high [wind turbine]	[power system] reliability	[blade element] theory
rated [wind speed]	offshore [wind project]	[reactive power] consumption
large [wind farm]	[wind turbine] model	[wind energy] potential
onshore [wind farm]	power [electronic converter]	installed [wind generation]
[wind turbine] blade	[wind turbine] generator	offshore [wind resource]
[wind power] output	[sound pressure] level	[wind turbine] application
low [wind speed]	[wind turbine] manufacturer	power [spectral density]
[wind turbine] rotor	[wind energy] project	[wind speed] forecasting
large [wind turbine]	[wind power] fluctuation	[wind power] integration
[control system] design	[heat transfer] medium	[transmission system] operator
average [capacity factor]	[wind power] project	[thermal power] plant
[wind energy] sector	[hydroelectric power] station	[time domain] simulation
unity [power factor]	urban [wind turbine]	[reactive power] control
[full load] hour	[hydro power] plant	[grid connection] cost

[wind turbine] component	[wind energy] capacity	[wind energy] application
[power system] operation	[hydroelectric power] plant	[voltage source] converter
net [capacity factor]	[wind resource] assessment	[sound power] level
[mass flow] rate	[wind farm] development	net [present value]
[wind energy] converter	[wind energy] density	conventional [wind turbine]
[wind turbine] system	[reactive power] compensation	[renewable energy] resource
[wind turbine] technology		

Table 1: List of MWTs manually bracketed

Based on bracketing models (2.1), the terms in Table 1 were decomposed in all possible groupings and/or searched for within different structures, as pointed out by the following 12 indicators:

1. MWTs decomposed in all possible groupings according to adjacency, dependency and shortening models (p1 p2, p2 p3, p1 p3) (for *offshore wind farm*, *offshore wind, wind farm, offshore farm*);

2. Insertions within the MWTs (p1 * p2 p3 and p1 p2 * p3) (*offshore* [*wind farm*] because *offshore* ***shrouded*** *wind farm*);

3. Longer MWTs where adjacent groupings act as modifiers (p1 p2 *, p2 p3 *), head (* p1 p2, * p2 p3) or middle modifiers (* p1 p2 *, * p2 p3 *) (*offshore* [*wind farm*] because ***onshore*** *wind farm*);

4. MWTs with a hyphen between adjacent groupings (p1-p2 p3, p1 p2-p3) ([*cell cycle*] *analysis* because *cell-cycle analysis*);

5. MWTs with the possessive genitive between adjacent groupings (p1's p2 p3, p1 p2's p3) (*brain* [*stem cell*] because *brain****'s*** *stem cell*);

6. MWTs showing brackets around a single element (p1 p2 (p3), p1 (p2) p3, (p1) p2 p3) or a grouping ((p1 p2) p3, p1 (p2 p3)) ([*cell cycle*] *analysis* because **(***cell cycle***)** *analysis*);

7. MWTs where one of the adjacent groupings forms a monolexical compound (p1p2 p3, p1 p2p3) ([*gear box*] *manufacturer* because ***gearbox*** *manufacturer*);

8. MWTs where one of the first two elements is inflected for number (p1 p2s p3, p1s p2 p3) ([*tyrosine kinase*] *activation* because *tyrosine kinase****s*** *activation*);

9. MWTs showing a different word order of the first two elements (p2 p1 p3) (*mean* [*total consumption*] because *total mean consumption*);

10. MWTs decomposed in all possible groupings having a prepositional paraphrase in between (p3 PREP p1 p2, p2 p3 PREP p1, p1 p3 PREP p2) ([*permanent magnet*] *generator* because *generator* ***with*** *permanent magnets*; [*mean wind*]

speed because *mean speed **of** wind*);

11. MWTs decomposed in all possible groupings having a verbal paraphrase in between (p3 V p1 p2, p1 p2 V p3, p2 p3 V p1, p1 V p2 p3) ([*permanent magnet*] *generator* because *generator **has** permanent magnets*);

12. MTWs where one of the adjacent groupings is followed by two capital letters (expecting an acronym) in brackets (p1 p2 (AA) p3, p1 p2 p3 (AA)) ([*direct current*] *generator* because *direct current **(DC)** generator*).

Consequently, 34 specific CQL (Corpus Query Language) queries were designed for the extraction of occurrences of each of the above structures (Table 2).

Bracketing indicators	**Structure retrieved**	**CQL queries**
Decomposed MWTs	p1 p2	[tag!="JJ.*|N.*"][lemma="**p1**"][lemma="**p2**"][tag!="N.*|JJ.*"]
	p2 p3	[tag!="JJ.*|N.*"][lemma="**p2**"][lemma="**p3**"][tag!="N.*|JJ.*"]
	p1 p3	[tag!="JJ.*|N.*"][lemma="**p1**"][lemma="**p3**"][tag!="N.*|JJ.*"]
Insertions	p1 * p2 p3	[lemma="**p1**"][tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]+ [lemma="**p2**"][lemma="**p3**"]
	p1 p2 * p3	[lemma="**p1**"][lemma="**p2**"][tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]+ [lemma="**p3**"]
Longer MWTs	p1 p2 *	[tag!="N.*|JJ.*"][lemma="**p1**"][lemma="**p2**"] [tag="JJ.*|N.*|RB.*|VVG.*|VVN.*" & lemma!= "**p3**"]* [tag="N.*" & lemma!= "**p3**"]
	* p1 p2	[tag="N.*|JJ.*"]+[lemma="**p1**"][lemma="**p2**"] [tag!="N.*|JJ.*"]
	p2 p3 *	[tag!="N.*|JJ.*"] [lemma="**p2**"][lemma="**p3**"] [tag="JJ.*|N.*|RB.*|VVG.*|VVN.*"]* [tag="N.*"]
	* p2 p3	[tag="N.*|JJ.*" & lemma!= "**p1**"]+ [lemma="**p2**"] [lemma="**p3**"] [tag!="N.*|JJ.*"]
	* p1 p2 *	[tag="N.*|JJ.*"]+ [lemma="**p1**"][lemma="**p2**"] [tag="JJ.*|N.*|RB.*|VVG.*|VVN.*" & lemma!="**p3**"]*[tag="N.*" & lemma!= "**p3**"]
	* p2 p3 *	[tag="N.*|JJ.*" & lemma!="**p1**"]+[lemma="**p2**"][lemma="**p3**"]

		[tag="JJ.*|N.*|RB.*|VVG.*|VVN.*" & lemma!="**p3**"]*[tag="N.*"]
Hyphen	p1-p2 p3	[lemma="**p1-p2**"][lemma="**p3**"]
	p1 p2-p3	[lemma="**p1**"][lemma="**p2-p3**"]
Possessive genitive	p1 p2's p3	[lemma="**p1**"][word="**p2**'s"][lemma="**p3**"]
	p1's p2 p3	[word="**p1**'s"][lemma="**p2**"][lemma="**p3**"]
Brackets	p1 p2 (p3)	[lemma="**p1**"][lemma="**p2**"][word="\("][lemma="**p3**"][word="\)"]
	(p1) p2 p3	[word="\("][lemma="**p1**"][word="\)"][lemma="**p2**"][lemma="**p3**"]
	p1 (p2) p3	[lemma="**p1**"][word="\("][lemma="**p2**"][word="\)"][lemma="**p3**"]
	(p1 p2) p3	[word="\("][lemma="**p1**"][lemma="**p2**"] [word="\)"] [lemma="**p3**"]
	p1 (p2 p3)	[lemma="**p1**"][word="\("][lemma="**p2**"] [lemma="**p3**"] [word="\)"]
Monolexical compound	p1p2 p3	[lemma="**p1p2**"][lemma="**p3**"]
	p1 p2p3	[lemma="**p1**"][lemma="**p2p3**"]
Inflection	p1 p2s p3	[lemma="**p1**"][lemma="**p2**" & tag="NNS"][lemma="**p3**"]
	p1s p2 p3	[lemma="**p1**" & tag="NNS"][lemma="**p2**"][lemma="**p3**"]
Word order	p2 p1 p3	[lemma="**p2**"][lemma="**p1**"][lemma="**p3**"]
Prepositional paraphrases	p3 PREP p1 p2	[lemma="**p3**"][]{0,2}[tag="IN" & lemma!="like"] []{0,2}[lemma="**p1**"][lemma="**p2**"][lemma!="**p3**"]
	p2 p3 PREP p1	[lemma!="**p1**"][lemma="**p2**"][lemma="**p3**"][]{0,2}[tag="IN" & lemma!="like"][]{0,2}[lemma="**p1**"]
	p1 p3 PREP p2	[tag!="JJ.*|N.*"][lemma="**p1**"][lemma="**p3**"][]{0,2}[tag="IN" & lemma!="like"][]{0,2}[lemma="**p2**"][tag!="JJ.*|N.*"]
Verbal paraphrases	p3 V p1 p2	[lemma="**p3**"][]{0,2}[tag="VV.*"][]{0,2}[lemma="**p1**"] [lemma="**p2**"][lemma!="**p3**"]

	p1 p2 V p3	[lemma="**p1**"][lemma="**p2**"][lemma!="**p3**"]{0,2}[tag="VV.*"] []{0,2}[lemma="**p3**"]
	p2 p3 V p1	[lemma!="**p1**"][lemma="**p2**"][lemma="**p3**"][]{0,2}[tag="VV.*"] []{0,2}[lemma="**p1**"]
	p1 V p2 p3	[lemma="**p1**"][lemma!="**p2**"]{0,2}[tag="VV.*"][lemma!="**p1**"]{0,2} [lemma="**p2**"][lemma="**p3**"]
Acronyms	p1 p2 (AA) p3	[lemma="**p1**"][lemma="**p2**"][word="\("][word="[A-Z]{2}(s)?"][word="\)"][lemma="**p3**"]
	p1 p2 p3 (AA)	[lemma="**p1**"][lemma="**p2**"][lemma="**p3**"][word="\("][word="[A-Z]{2}(s)?"][word="\)"]

Table 2: CQL queries

To retrieve all the data, each constituent of the 103 MWTs was automatically filled in the placeholders of p1, p2 and p3 and queries were sent to both corpora through Sketch Engine's API, which means that a total of 7,004 queries were performed. In order to avoid noise, all queries were applied to a single sentence (within <s/>) and sub-hits (lazy results causing a multiplying effect) were filtered out. For the same reason, some of the queries need to exclude certain elements. For example, when looking for the MWTs decomposed in three independent terms (p1 p2, p2 p3, p1 p3), the queries exclude any adjective or noun before and after them ([tag!="N.*|JJ.*"]) to avoid structures where the groupings are only part of longer MWTs.

Based on the figures retrieved through the Sketch Engine's API, the following 16 rules were developed in order to automatically compute the bracketing of each MWT (Table 3).

Adjacency	1. If p1 p2 > p2 p3 then (p1 p2) p3; If p1 p2< p2 p3 then p1 (p2 p3); Else, N/A
Dependency	2. If p2 p3 > p1 p3, then (p1 p2) p3; If p2 p3 < p1 p3, then p1 (p2 p3); Else, N/A
Shortening	3. If p1 p2 > p1 p3, then (p1 p2) p3 If p1 p3 & p2 p3 > p1 p2, then p1 (p2p3) Else, N/A

Insertions	4. If p1 p2 * p3 > p1 * p2 p3, then (p1 p2) p3; If p1 p2 * p3 < p1 * p2 p3, then p1 (p2 p3); Else, N/A
Longer MWTs	5. If p1 p2 * + * p1 p2 + * p1 p2 * > p2 p3 * + * p2 p3 + * p2 p3 *, then (p1 p2) p3; If p1 p2 * + * p1 p2 + * p1 p2 * < p2 p3 * + * p2 p3 + * p2 p3 *, then p1 (p2 p3); Else, N/A
Hyphen	6. If p1-p2 p3 > p1 p2-p3, then (p1 p2) p3; If p1-p2 p3 < p1 p2-p3, then p1 (p2 p3); Else p1 N/A
Possessive genitive	7. If p1 p2's p3 > p1's p2 p3, then (p1 p2) p3; If p1 p2's p3 < p1's p2 p3, then p1 (p2 p3); Else N/A
Brackets	8. If p1 p2 (p3) > (p1) p2 p3 + p1 (p2) p3, then (p1 p2) p3; If p1 p2 (p3) < (p1) p2 p3 + p1 (p2) p3, then p1 (p2 p3); Else N/A
	9. If (p1 p2) p3 > p1 (p2 p3), then (p1 p2) p3; If (p1 p2) p3 < p1 (p2 p3), then p1 (p2 p3); Else N/A
Monolexical compound	10. If p1 p2 p3 > p1 p2p3, then (p1 p2) p3; If p1 p2 p3 < p1 p2p3, then p1 (p2 p3); Else N/A
Internal inflection	11. If p1 p2s p3 > p1s p2 p3, then (p1 p2) p3; If p1 p2s p3 < p1s p2 p3, then p1 (p2 p3); Else N/A
Word order	12. If p2 p1 p3 > 0, then p1 (p2 p3); Else N/A
Prepositional paraphrases	13. If p3 PREP p1 p2 > p2 p3 PREP p1, then (p1 p2) p3; If p3 PREP p1 p2 < p2 p3 PREP p1, then p1 (p2 p3); Else N/A

	14. If p1 p3 PREP p2 > 0, then p1 (p2 p3) Else, N/A
Verbal paraphrases	15. If p3 V p1 p2 + p1 p2 V p3 > p2 p3 V p1 + p1 V p2 p3, then (p1 p2) p3; If p3 V p1 p2 + p1 p2 V p3 < p2 p3 V p1 + p1 V p2 p3, then p1 (p2 p3); Else N/A
Acronyms	16. If p1 p2 (AA) > p1 p2 p3 (AA), then (p1 p2) p3; If p1 p2 (AA) < p1 p2 p3 (AA), then p1 (p2 p3); Else N/A

Table 3: Bracketing rules

Most of the rules lead to either left or right bracketing (or N/A if no results or equal results are obtained), but two of them are only indicative of one. If rules 12 and 14 apply, they will indicate a left or right bracketing, respectively, but if they do not, that does not mean that the opposite bracketing applies. For instance, when applying rule 12 to *micro hydropower plant*, the word order *hydropower micro plant* is not found. However, this does not mean that it has a left bracketing. Furthermore, most of the rules compare the figures of two queries, but some others include the addition of several from different queries (5, 8 and 15). For instance, when rule 5 is applied to *wind power fluctuation*, longer MWTs formed by each of the possible groupings are compared and added (e.g. for *wind power*, longer MWTs, such as *wind power system*, *onshore wind power*, and *offshore wind power consumption*, are added and compared to the figures associated with *power fluctuation*). Finally, except for rules 12 and 14, all the rules but one (3) are composed of two opposing conditions. Rule 3 is a mixture of the left-bracketing condition of the dependency model and two nested conditions (p1 p3 > p1 p2 & p2 p3 > p1 p2).

In sum, the protocol is composed of 12 indicators formulated in 34 queries, whose results are compared in 16 bracketing rules. Once the rules were applied and the bracketing candidates obtained (based on the agreement of most rules, which all have the same weight), the results were compared to the baseline.

2.2.2 Evaluating the protocol: rules and corpus reliability

Our results showed that the protocol allows for the correct bracketing of MWTs in more than 83% of the cases as the average in both corpora, but some of the rules are more productive and/or reliable than others, certain differences between the corpora can also be found, and the confidence level of all rules (i.e. the probability to match with the baseline based on the number of rules agreeing on the same result) shows differences among the MWTs in the dataset.

The performance of the rules for disambiguating purposes is based on their likelihood to retrieve results from corpora and their ability to actually solve MWT bracketing as compared to the baseline. The balance between frequency and reliability is what constitutes the basis for a weighted protocol. This means that there are rules that do not retrieve any result very often, but they are highly reliable when they do. For instance, the possessive rule had a 100% matching rate but could only be used with seven MWTs. In contrast, there are rules that are always likely to retrieve results but do not always deliver an output matching the baseline.

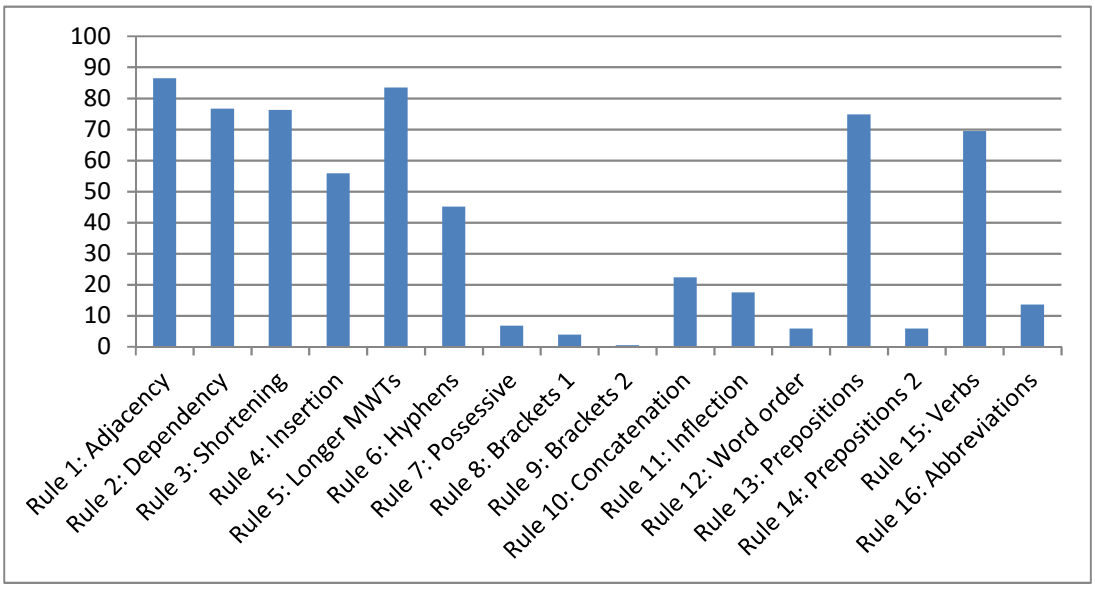

Figure 1: Performance of bracketing rules

Figure 1 shows the performance of each of the rules considering both factors. Adjacency (86.4%), longer MWTs (83.5%), dependency (76.7%) and shortening (76.2%) are, collectively, the most useful rules.

As for the corpora, the agreement with the baseline based on the queries on the WPC outperformed that of the DOAJ. Another difference is the varying performance of the protocol on left or right bracketing. Generally speaking, left bracketing is better identified in both corpora, but the difference is even more noticeable in the WPC.

Corpus size and type were thus found to have an influence on the results. The WPC, although smaller in size, provided better bracketing results for the MWT dataset (86.4% vs. 79.6%), as it belongs to the wind power domain. Domain-specificity is thus a key factor for the performance of the protocol over size.

When looking at the rules individually, differences can also be found when comparing corpora (Figure 2) from both quantitative and qualitative points of view.

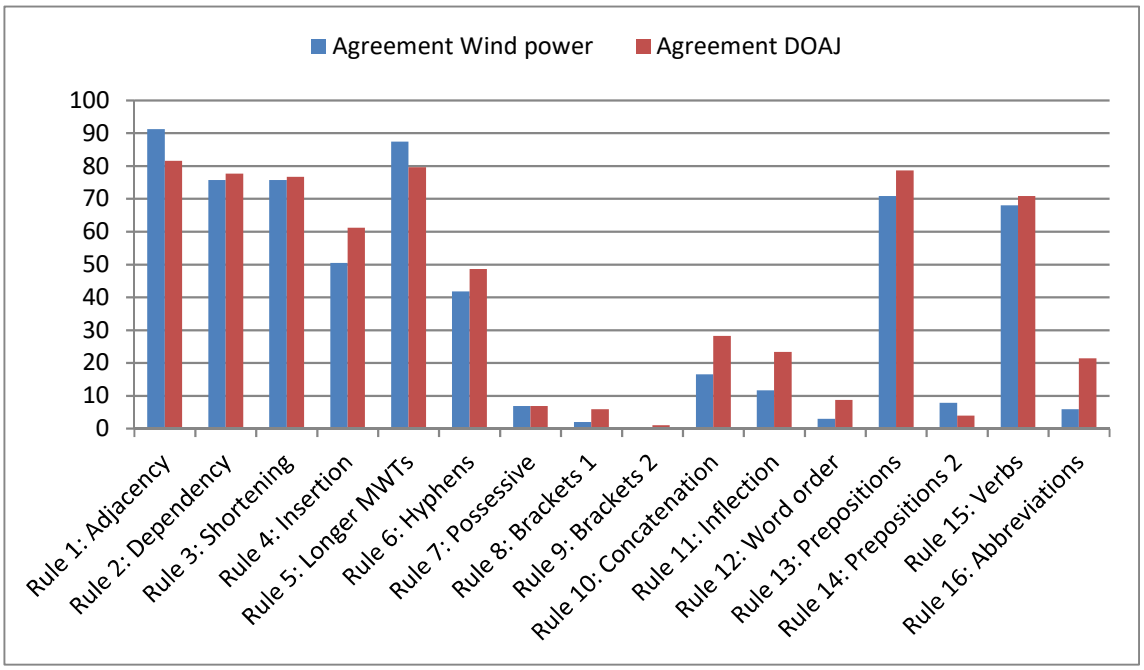

Figure 2: Quantitative and qualitative performance of bracketing rules in both corpora

As previously mentioned, the most reliable rules in both corpora were those related to adjacency, dependency, or the capacity to form new longer MWTs, followed by prepositional or verbal paraphrases and insertions. However, the DOAJ provided better results for certain indicators related to the "surface patterns" reported by Nakov (2007) (e.g. hyphens, concatenation, inflection, abbreviations, etc.), since such patterns will be more likely found in larger corpora. Among the most reliable rules, adjacency and longer MWTs performed better in the WPC, whereas dependency, shortening and insertion performed better in the DOAJ, which might indicate that the former are domain-dependent and the latter size-dependent. This can be verified when looking at the figures (Figure 3) from a purely qualitative way (i.e. not taking into account when no results are retrieved from the corpora and bracketing cannot be computed). In that case most of the rules except for brackets, prepositional paraphrases (and only that of p1 p3 PREP p2) and abbreviations were more reliable in the WPC.

The fact that right bracketing has a lower matching rate with the baseline, especially in the DOAJ, opens a new line of inquiry regarding the nature of these MWTs and their syntactic structure, since the choice of the dataset, based on frequency, was not balanced in terms of left/right bracketing or syntactic structures. The main differences between the corpora are the following: adjacency is equally reliable for left and right bracketing in the WPC as opposed to the DOAJ, where right bracketing reliability scores higher; the insertion and longer MWTs rules work in opposing directions; the inflection rule in the DOAJ only shows reliability for left bracketing.

In the WPC, 100% reliability is shown for hyphens and possessives in the case of left bracketing and for word order for right bracketing. In the DOAJ, 100% reliability is

found for bracketed groupings in the case of right bracketing. In both of them, 100% reliability is found for bracketed single words, word order, type 2 prepositional paraphrases and abbreviations in the case of right bracketing.

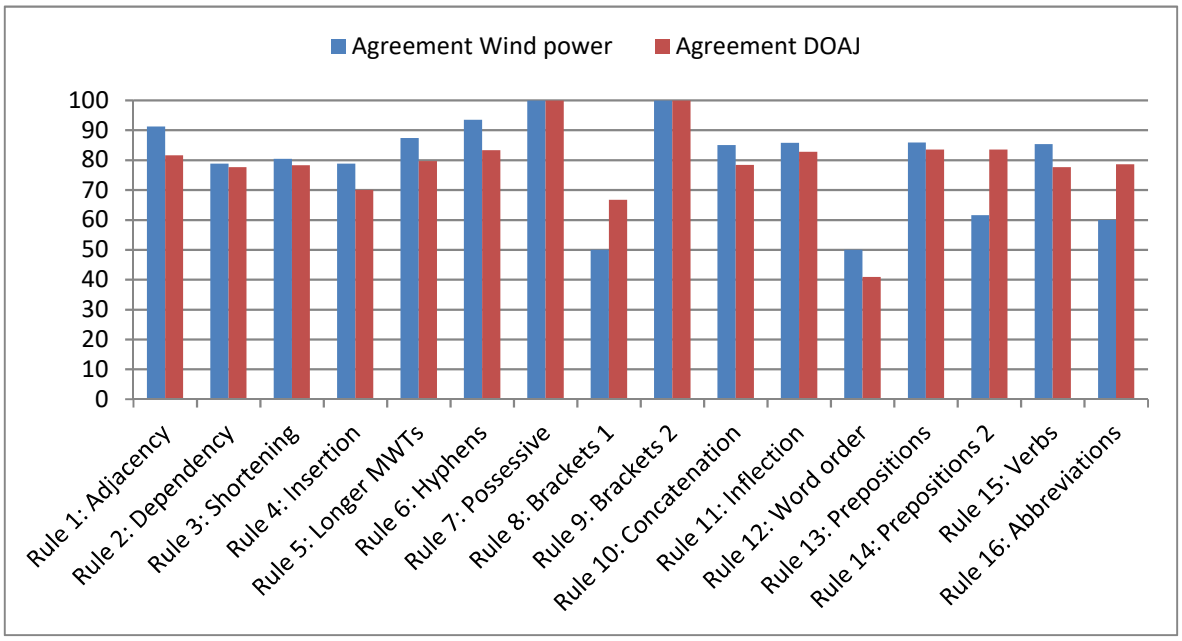

Figure 3. Qualitative performance of bracketing rules in both corpora

Regarding the overall evaluation of the protocol, the output was analysed based on the following: (1) whether the resulting bracketing agreed with the baseline; (2) whether the candidate bracketing was the same in both corpora: and (3) the confidence of each bracketing based on the number of rules pointing in the same direction without considering N/A results (no results from the queries). For instance, for [*wind turbine*] *blade*, even if only 68.75% of the rules could be applied, 100% of them pointed to a left bracketing.

In half the cases, the rules showed a 100% confidence, 51.45% for the WPC and 41.74% for the DOAJ, from which 96.22% and 95.34%, respectively, agreed with the baseline. The only failed bracketings with a 100% confidence were *offshore wind project* (in both corpora), *sound power level* in the WPC, and *offshore wind park* in the DOAJ. From the bracketings showing 80 to 99% confidence (20.38% in both corpora), 90.47% and 85.71% agreed with the baseline. From 50 to 79% confidence (28.15% and 37.86%), 65.51% and 58.97% agreed with the baseline.

In the WPC alone, erroneous bracketing only occurred for *hydroelectric power station* (and only because the application of all rules gave a N/A output), whereas in the DOAJ failures included *wind power plant*, *wind power generation*, *wind power output*, *power electronic converter*, *sound pressure level*, *wind energy density*, *wind energy production*, and *reactive power consumption*. The fact that more erroneous bracketings were found through the DOAJ might indicate again that domain-specificity is what matters the most, since in this corpus many different domains converge and the constituents of these MWTs might accept very different combinations outside the wind

power domain.

In both corpora the bracketing failed for the following 13 MWTs: *offshore wind power*, *offshore wind energy*, *wind penetration level*, *offshore wind project*, *hydroelectric power plant*, *hydro power plant*, *micro hydropower plant*, *installed wind generation*, *offshore wind resource*, *thermal power plant*, *sound power level*, *mass flow rate* and *offshore wind park*. We have thus selected this list to perform a more in-depth analysis of possible causes.

2.2.3 Understanding the causes of protocol failure

The 13 MWTs where the protocol failed in both corpora are shown in Table 4 with both outputs and confidence levels.

Baseline	**WPC output**	**Confidence**	**DOAJ output**	**Confidence**
offshore [wind power]	N/A	50%	[offshore wind] power	62.5%
offshore [wind energy]	[offshore wind] energy	62.5%	[offshore wind] energy	55.5%
[wind penetration] level	N/A	50%	wind [penetration level]	66.6%
offshore [wind project]	[offshore wind] project	100%	[offshore wind] project	100%
[hydroelectric power] plant	N/A	50%	hydroelectric [power plant]	70%
[hydro power] plant	hydro [power plant]	55.5%	hydro [power plant]	55.5%
micro [hydropower plant]	N/A	50%	[micro hydropower] plant	66.6%
installed [wind generation]	[installed wind] generation	71.4%	[installed wind] generation	66.6%
offshore [wind resource]	[offshore wind] resource	85.7%	[offshore wind] resource	85.7%
[thermal power] plant	N/A	50%	thermal [power plant]	75%
[sound power] level	sound [power level]	100%	sound [power level]	83.3%
[mass flow] rate	mass [flow rate]	66.6%	mass [flow rate]	83.3%
offshore [wind park]	[offshore wind] park	80%	[offshore wind] park	100%

Table 4: 13 MWTs where the bracketing protocol failed

In most cases, the system delivered the baseline's opposite bracketing, but in five cases the results retrieved by the WPC were N/A, since the results pointed to a 50% confidence, which indicates again that a domain-specific corpus outperforms a large one. The results of these MWTs were analysed based on the following possible causes: (i) the nature of the MWTs (e.g. the left/right bracketing, omission of constituents, their syntactic structure, exceptions to the rule); (ii) the formulation of the corpus queries; and (iii) the rules' confidence level; and (iv) the fact that some rules might be noisier than helpful, thus biasing the results.

Based on their syntactic structures, most of the MWTs (9) follow the structure A+N+N; only one MWT shows the Participle+N+N structure, which could be subsumed under the latter; and three N+N+N structures are found. Considering that A+N+N structures only amount to 30% of the initial 103 MWT dataset, this could point to a degree of bracketing difficulty for such structures, although this should be confirmed by replicating the study with a more balanced dataset in terms of syntactic structure.

In terms of left or right bracketing, the set of failed MWTs is really balanced (six and seven respectively) as compared to their proportion in the original 103 MWT set (34 right-bracketed and 69 left-bracketed MWTs), which suggest that this factor does not necessarily influence the success of the protocol.

There seems to be a trend in failure for MWTs having *plant* or *level* as their head and *offshore wind* as modifiers. In some of these cases, a variable bracketing could occur even in human scenarios, which is often the result of multidimensionality and could explain why the rules did not solve the bracketing of most MWTs in this 13-element set, since 11 of them contain the above mentioned heads or modifiers. For instance, the constituents *offshore wind* could be bracketed together indicating the wind type (e.g. [*offshore wind*] *power* would refer to the energy produced from this type of wind).

Alternatively, the opposite grouping would instead highlight the location relation between *offshore* and the head. For example, *offshore* [*wind power*] would allude to a type of energy produced in that specific location. Furthermore, constituents such as *turbine* or *farm* could have been elicited between *wind* and *power* (the true term being *offshore wind turbine/farm power*), in which case *offshore* would refer to the place where those devices are located. The same concept can thus be seen from different angles, so both human and automatic procedures could be likely to provide contradictory bracketed structures. In this sense, the cases of *offshore wind project* and *offshore wind resource* might be a case of human bracketing failure (despite inter-annotator agreement), since confidence figures are particularly striking. These are the only two MWTs, together with *sound power level* and *offshore wind park*, where confidence level scored so high in the wrong direction as compared to the baseline. In contrast, most failed bracketings showed a confidence level of 50-60%, which points to the possibility of setting a threshold above 60%.

Something similar could happen with *hydro power plant*, where [*hydro power*] *plant* would be a plant that uses water power (in a more general sense of energy) and *hydro* [*power plant*] would imply a plant generating power (in the sense of electricity) that uses water. *Hydroelectric power plant* would fall under the same hypothesis, however, with its synonym *hydroelectric power station* the protocol did not fail. The same happened with sound power level, which got a failed bracketing while a very similar term (*sound pressure level*) got it right. This reinforces the hypothesis that *power*, due to polysemy, is especially prone to multidimensionality.

Regarding the formulation of corpus queries, no errors possibly influencing the results were found. The last step was to wonder whether there were certain rules that might be more misleading than helpful in these MWTs, opening the possibility of constraining the protocol for the rest of the MWTs in the set. Table 5 shows the performance of each rule for each MWT in the WPC/DOAJ. However, no significant patterns were found, which means that if each rule were to have a different weight, weights cannot be inferred by analysing erroneous bracketings.

	offshore [wind power]	offshore [wind energy]	[wind penetration] level	offshore [wind project]	[hydroelectric power] plant	[hydro power] plant	micro [hydropower plant]
Rule 1: Adjacency	Agree / Agree	Agree / Agree	Agree / Fail	Fail / Fail	Fail / Fail	Fail / Fail	Agree / Agree
Rule 2: Dependency	Fail / Fail	Fail / Fail	Agree / Agree	Fail / Fail	Agree / Agree	Agree / Agree	Fail / Fail
Rule 3: Shortening	Fail / Fail	Fail / Fail	Agree/Agree	Fail / Fail	Agree / Agree	Agree / Agree	Fail / Fail
Rule 4: Insertion	N/A/Fail	N/A / Fail	Fail / Fail	Fail / Fail	N/A / Fail	Fail/ Agree	N/A/ N/A
Rule 5: Longer MWTs	Agree / Agree	Fail / Agree	Fail / Fail	Fail / Fail	Fail / Fail	Fail / Fail	Agree / Agree
Rule 6: Hyphens	Agree / N/A	Agree / N/A	Agree / N/A	N/A / N/A	N/A / Agree	Agree / Fail	N/A / Fail
Rule 7: Possessive	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 8: Brackets 1	N/A /	N/A /	N/A / N/A	N/A /	N/A / N/A	N/A / N/A	N/A / N/A

	Agree N/A		N/A			
Rule 9: Brackets 2	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 10: Concatenation	Agree / N/A	N/A / Agree	N/A / N/A	Fail / N/A	N/A / Agree	Agree / Fail	N/A / Fail
Rule 11: Inflection	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 12: Word order	N/A / N/A	N/A / Agree	N/A / N/A	N/A / N/A	N/A / N/A	Fail / N/A	N/A / N/A
Rule 13: Prepositions	Fail / Fail	Fail/ Fail	Fail/ Fail	Fail/ Fail	N/A/ Fail	Fail/ Fail	N/A / N/A
Rule 14: Prepositions 2	N/A / N/A	Agree / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / Fail	N/A / N/A
Rule 15: Verbs	Fail / Fail	Fail / Fail	Fail / N/A	Fail / Fail	N/A / Fail	N/A / Fail	N/A / N/A
Rule 16: Abbreviations	N/A / N/A	N/A/ N/A	N/A/ N/A	N/A/ N/A	N/A/ N/A	N/A/ N/A	N/A/ N/A

Table 5: Rules' performance on 13 failed bracketings in the WPC

	installed [wind generation]	offshore [wind resource]	[thermal power] plant	[sound power] level	[mass flow] rate	offshore [wind park]
Rule 1: Adjacency	Agree / Agree	Fail / Agree	Fail / Fail	Fail / Fail	Fail / Fail	Fail / Fail
Rule 2: Dependency	Fail / Fail	Fail / Fail	Agree / Agree	Fail / Fail	Agree / Agree	Fail / Fail
Rule 3: Shortening	Fail / Fail	Fail / Fail	Agree / Agree	Fail / Fail	Agree/Agree	Fail / Fail
Rule 4: Insertion	Fail / Fail	Fail / Fail	N/A / Fail	N/A / N/A	N/A / Fail	N/A / Fail

Rule 5: Longer MWTs	Fail / Agree	Fail / Fail	Fail / Fail	Fail / Fail	Fail / Fail	Fail / Fail
Rule 6: Hyphens	N/A / N/A	N/A / N/A	N/A / Agree	N/A / N/A	N/A / Fail	N/A / Agree
Rule 7: Possessive	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 8: Brackets 1	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 9: Brackets 2	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 10: Concatenation	N/A / N/A	N/A / N/A	N/A / Fail	N/A / N/A	Fail / Fail	Agree / N/A
Rule 11: Inflection	N/A / N/A	N/A / N/A	N/A / Fail	N/A / N/A	N/A / N/A	N/A / N/A
Rule 12: Word order	N/A / N/A	N/A / N/A	N/A / Fail	N/A / N/A	Fail / Fail	N/A / N/A
Rule 13: Prepositions	Agree / N/A	Fail/ Fail	Fail/ Fail	N/A / Agree	N/A / Fail	N/A / N/A
Rule 14: Prepositions 2	N/A / N/A	N/A / N/A	N/A / N/A	N/A / Fail	N/A / Fail	N/A / N/A
Rule 15: Verbs	Fail / Fail	Agree / Fail	Agree / Fail	N/A / N/A	N/A / Fail	N/A /N/A
Rule 16: Abbreviations	N/A/ N/A	N/A/ N/A	N/A / Fail	N/A/ N/A	N/A / Fail	N/A/ N/A

Table 5: Rules' performance on 13 failed bracketings in the WPC II

In any case, the protocol delivered promising results that could be applied in any terminology management scenario needing a thorough description of MWTs.

3. Multiword-term representation in terminological knowledge bases

An accurate representation of MWTs in terminological knowledge bases involves providing users with access to the implicit information codified in such specialised

units, namely their structural dependencies and the semantic relations encoded among the constituents. Since the second depends on the first, automatising bracketing facilitates the inclusion of such information. Furthermore, establishing equivalence and performing cross-lingual comparisons are only possible through the semantics implied.

In EcoLexicon, a new module for the description of MWTs has been designed. When users query a monolexical term, they can access all of the MWTs where the search term appears as a constituent, whether it is the head or a modifier. Figure 4 shows the summary view of four different tabs where different types of information are provided, in this case regarding the search term *turbine*.

The results of this view are a summary of what is obtained in the specific views that will be described below, namely (i) MWT formation, (ii) Equivalents, (iii) Morphosyntactic combinations, and (iv) Semantic combinations. As can be observed in Figure 4, the CN formation bubble shows some of the MWTs that include the term turbine. These examples are also shown in the Equivalents bubble along with their main Spanish equivalents. The Morphosyntactic combinations bubble focuses on bracketing and part-of-speech tagging. Finally, the Semantic combinations bubble also shows bracketing, as well as annotation with semantic categories (blue), semantic roles (red), and the internal semantic relation (grey, on the right).

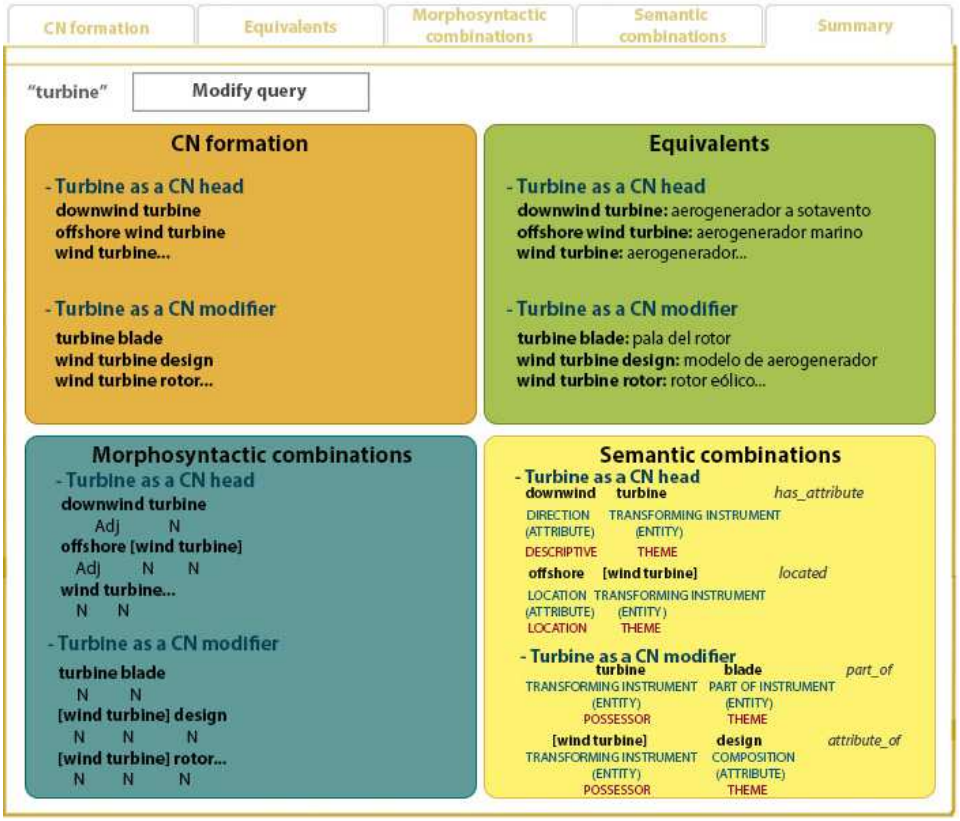

Figure 4: Summary view of the MWT module in EcoLexicon

Figure 5 shows an extract from the MWT formation tab, where the term *generator* is shown as the head of three terms hierarchically organised, linked to their definitions and highlighted conceptual dimensions (i.e. rotor or grid connection) as well as related term variants (i.e. *SCIG*, *DFIG*). MWTs whose modifier is generator (e.g. *generator torque control*) can also be obtained.

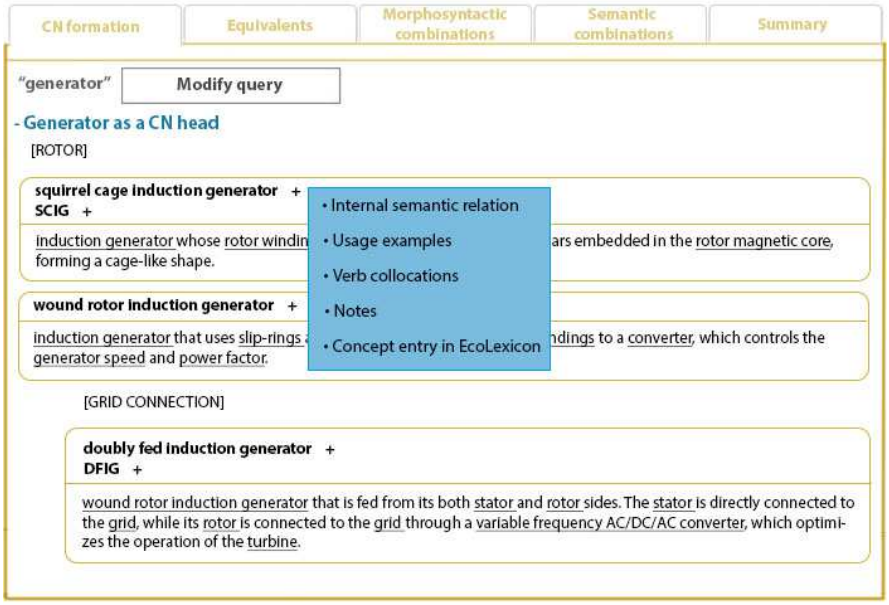

Figure 5: Extract from the MWT formation tab for *generator*

By clicking on the plus sign next to each term, users can access additional information: (i) internal semantic relations between the constituents of the MWT, (ii) usage examples, (iii) verb collocations; (iv) notes, (v) and the main term entry in the knowledge base. The *internal semantic relation* option shows the MWT head and modifier, as well as the semantic relation that links them. In MWTs formed by more than two constituents, bracketing facilitates this distinction between head and modifier, and is thus included in this view (e.g. *wound rotor induction generator* > [wound rotor] *part_of* [induction generator]).

Figure 6 shows an extract from the MWT equivalents tab, where the MWTs with *generator* as their head are now related to their corresponding terms in Spanish. Additional languages, such as French, are planned to be included in the near future. The same secondary options are offered as in the previous view, except for the definition, which is included here as a secondary option.

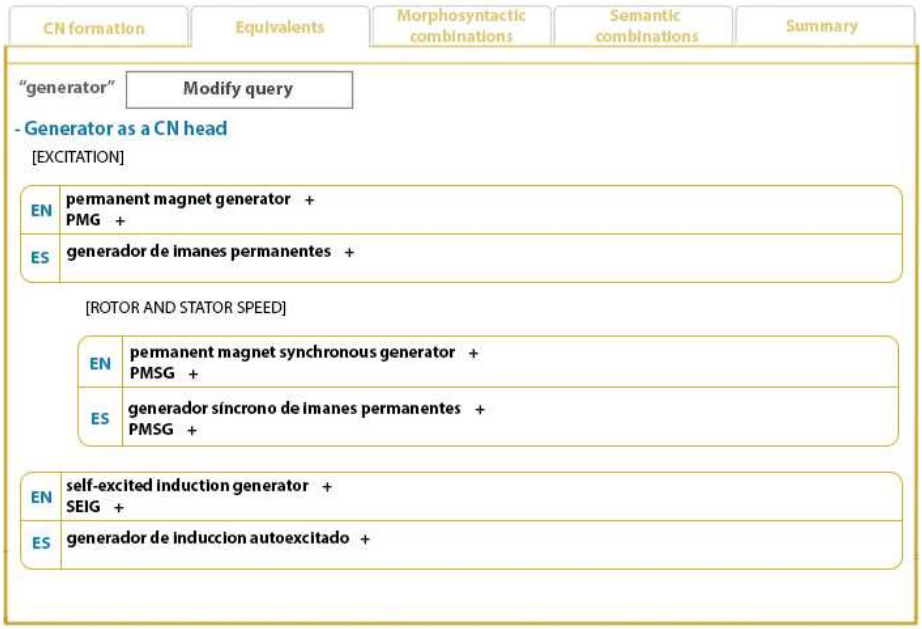

Figure 6: Extract from the MWT equivalents tab for *generator*

Figure 7 shows an extract from the morphosyntactic combinations tab, where the MWTs with *turbine* as their head are presented according to their morphosyntactic structure and bracketing.

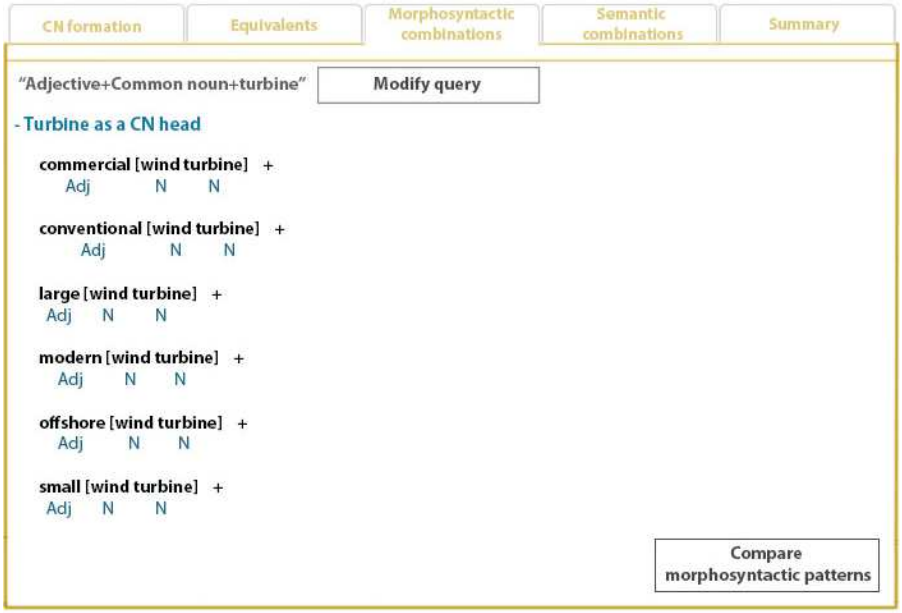

Figure 7: Extract from the Morphosyntactic combinations tab for *turbine*

By clicking on the plus sign next to each term, users can access additional information. In this view, the semantic relation is not provided since such semantic information is not relevant in this section. However, bracketing plays a central role, as it facilitates morphosyntactic analysis and MWT management.

Furthermore, when clicking in Compare morphosyntactic patterns, a bilingual view will be displayed (Figure 8). The results that meet the search criteria will be shown, together with their main variants in the target language. These are annotated with the part-of-speech of each constituent, so that the morphosyntactic patterns of term formation in both languages can be compared. Users can also observe that bracketing does not always correspond in the two languages (e.g. when the equivalent has fewer constituents, as in *power output curve* and *curva de potencia*).

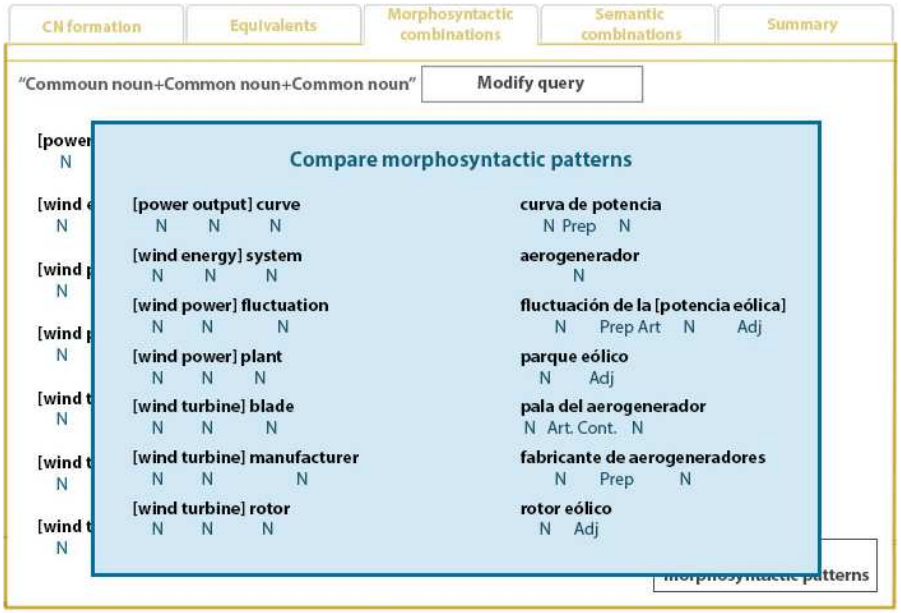

Figure 8: Extract from the Compare morphosyntactic combinations tab for *turbine*

Figure 9 shows an extract from the semantic combinations tab, where the semantic categories MAGNITUDE(ATTRIBUTE) and CHANGE(PROCESS) are queried to obtain the MWTs that include them. The MWTs retrieved are tagged with their bracketing structure (if they have three or more constituents), and their semantic categories and roles. For instance, in *voltage control*, *voltage* belongs to the category of MAGNITUDE(ATTRIBUTE) and *control* belongs to the category of CHANGE(PROCESS). In this MWT, *control* is the agent since it affects *voltage*, the patient. Next to each MWT, its internal semantic relation is also shown.

By clicking on the plus sign next to each term, users can access additional information. Unlike the previous views, an *additional semantic information* option is provided, which displays more specific data for users interested in further conceptual characterisation.

The *Compare semantic patterns* option is also provided (Figure 10). This section can be used to compare the semantic pattern of our results with that of their translation equivalents. A cross-linguistic approach to common phenomena such as variation or multidimensionality can thus be obtained, and the semantic annotation of MWTs in both languages can be contrasted (e.g. *small wind turbine*, based on size, vs its

equivalent *aerogenerador de baja potencia*, based on power). Not surprisingly, bracketing is the key to ascertaining the basic parts of MWTs and facilitate their understanding.

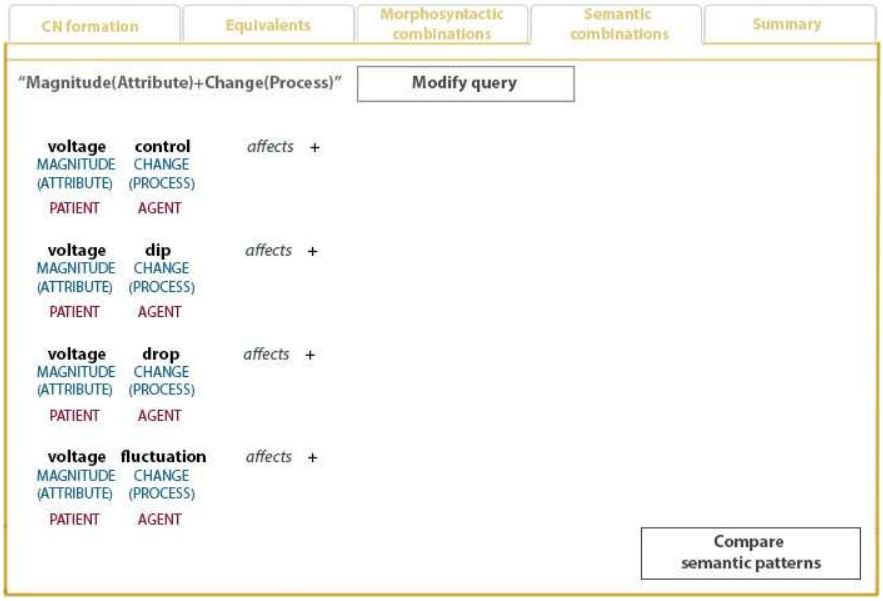

Figure 9: Extract from the Semantic combinations tab

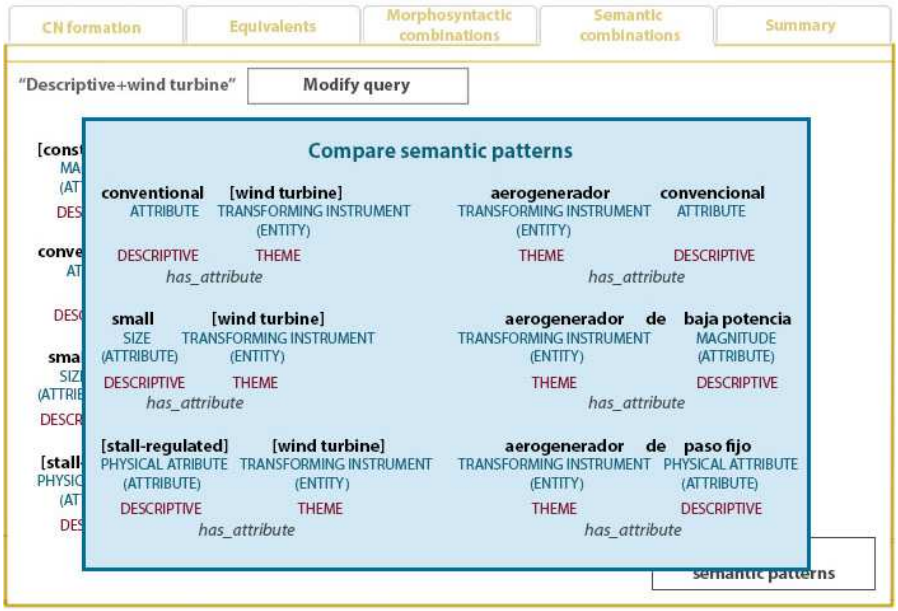

Figure 10: Extract from the Compare semantic patterns tab

4. Conclusions

In this paper, a bracketing protocol has been presented together with its practical application in the design and compilation of a MWT module in a terminological knowledge base. Regarding the protocol, we concluded that the most productive rules

are adjacency, longer MWTs, dependency, shortening, and paraphrases.

It is also advisable to perform the queries in domain-specific corpora, and not necessarily large ones. When large corpora are available, other surface patterns might prove more useful in terms of precision.

As for the MWT module described in this paper, it is intended to be useful for a wide variety of users, ranging from translators and interpreters, terminologists and technical writers, to students and environmental specialists. This resource includes different types of information that assists in both comprehension and production tasks. A systematic approach was adopted with a view to enhancing the heterogeneous description of MWTs in language resources, as well as specific problems such as the lack of consideration of internal dependencies or bracketing.

5. Acknowledgements

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

6. References

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

Cabezas-García, M. & Faber, P. (2017). A Semantic Approach to the Inclusion of Complex Nominals in English Terminographic Resources. In R. Mitkov (ed.) *Computational and Corpus-Based Phraseology*, Lecture Notes in Computer Science, 10596. Cham: Springer, pp. 145-159.

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

Cabezas-García, M. & León-Araúz, P. (2019). On the Structural Disambiguation of Multi-word Terms. In G. Corpas Pastor & R. Mitkov (eds.) *Computational and Corpus-Based Phraseology*, Lecture Notes in Computer Science, 11755. Cham: Springer, pp. 46-60.

Cabezas-García, M. (2020). *Los términos compuestos desde la Terminología y la Traducción*. Berlin: Peter Lang.

EcoLexicon. Accessed at: https://ecolexicon.ugr.es/. (1 February 2021)

Girju, R., Moldovan, D., Tatu, M. & Antohe, D. (2005). On the semantics of noun compounds. *Computer Speech & Language*, 19(4), pp. 479-496.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G.

Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress*. EURALEX, pp. 105-116.

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

León-Araúz, P. & Cabezas-García, M. (in press). Evaluating a bracketing protocol for multiword terms. In *Recent Advances in Multiword Units in Machine Translation and Translation Technology*. John Benjamins.

Marcus, M. (1980). *A Theory of Syntactic Recognition for Natural Language*. MIT Press.

Nakov, P. (2007). *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. PhD dissertation. Berkeley, University of California at Berkeley.

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

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