On the Structural Disambiguation of Multi-word Terms

Melania Cabezas-García(✉) and Pilar León-Araúz(✉)

University of Granada, Granada, Spain
{melaniacabezas,pleon}@ugr.es

Abstract. Multi-word terms pose many challenges in Natural Language Processing (NLP) because of their structure ambiguity. Although the structural disambiguation of multi-word expressions, also known as bracketing, has been widely studied, no definitive solution has as yet been found. Although linguists, terminologists, and translators must deal with bracketing problems, they generally must resolve problems without using advanced NLP systems. This paper describes a series of manual steps for the bracketing of multi-word terms (MWTs) based on their linguistic properties and recent advances in NLP. After analyzing 100 three- and four-term combinations, a set of criteria for MWT bracketing was devised and arranged in a step-by-step protocol based on frequency and reliability. Also presented is a case study that illustrates the procedure.

Keywords: Multi-word term · Structure ambiguity · Bracketing

1 Introduction

The creation of multi-word expressions is a frequent word-formation mechanism. When they belong to a specialized domain, they are known as 'multi-word terms' (MWTs) and usually specify a broader concept (e.g. *timing belt*, which is a type of belt that controls the opening and closing times of an engine's valves). MWTs are often composed by more than two elements when even more characteristics of the concept are conveyed (e.g. *permanent magnet synchronous generator*).

Surprisingly, studies that specifically address these longer combinations, characterized by their structural ambiguity, are fewer in number than those focusing on two-term MWTs [1–5]. In these MWTs a dependency analysis must be performed, which has cognitive implications since it is the basis for an accurate semantic analysis of the MWT and its subsequent applications (i.e. translation). This structural disambiguation, often known as 'bracketing' [6], involves the grouping of the dependent elements so that the MWT is reduced to its basic form of modifier+head, as in [*permanent magnet*] [*synchronous generator*].

Natural Language Processing (NLP) has particularly focused on multi-word expression bracketing because of its inherent difficulties for NLP systems [1, 2, 5, 7, 8]. However, problems that have still not been solved include MWTs formed by more than three elements. In fact, the more constituents an MWT has, the more difficult its bracketing. In addition, other areas such as Terminology or Translation Studies have

G. Corpas Pastor and R. Mitkov (Eds.): Europhras 2019, LNAI 11755, pp. 46–60, 2019.
https://doi.org/10.1007/978-3-030-30135-4_4

not addressed this phenomenon despite its relevance for MWT description in knowledge bases or the establishment of interlingual equivalences of MWTs. Since terminologists or translators do not usually resort to NLP systems to analyze MWT structures by means of algorithms, they require manual solutions.

This paper describes a series of steps for the bracketing of MWTs, based on the linguistic properties of these terms and recent advances in NLP. For this purpose, a corpus of English specialized texts on wind power was used to extract candidate MWTs and perform CQL queries based on a dataset to clarify dependences. Our goals included the following: (i) to devise a set of criteria for the manual bracketing of MWTs; and (ii) to propose generalizations to make the inference of bracketing structures easier. This proposal facilitates MWT disambiguation by terminologists, translators or any linguist, who do not have access to advanced NLP systems. In addition, it can also be used to enrich NLP algorithms for this task.

2 Structural Ambiguity in Multi-word Terms

MWTs are sequences of two or more elements that designate a specialized concept. For instance, the term that designates SOLAR PANEL is *solar panel* in English, *panel solar* in Spanish, *panneau solaire* in French, and *Solarmodul* in German. Since these terms usually have a nominal head, they are known as complex nominals, noun compounds or nominal compounds. These compounds can be endocentric or exocentric. Endocentric complex nominals have a head and a modifier, and specify a broader concept (*solar generator* is a type of *generator*), whereas exocentric complex nominals lack a head and, thus, are not subtypes of any of their constituents (*saber tooth* is not a hyponym of either *saber* or *tooth* because it designates a *saber-toothed tiger* by means of metonymy) [3]. The most frequent MWTs in specialized texts are endocentric complex nominals, which are the specification of a hypernym.

MWTs do not usually allow the insertion of elements [9], namely those modifying their conceptual content (*wind generator* can become *wind turbine generator*, but not **wind small generator*). They are a well-known term formation procedure, because they allow meaning condensation by means of the deletion of some elements, i.e. the semantic relation held by the constituents and, often, even some MWT constituents. MWTs are formed by a slot-filling mechanism. For example, WIND TURBINE can be specified depending on the orientation of its rotation axis (*horizontal/vertical axis wind turbine*), its location (*offshore/onshore wind turbine*), its number of blades (*two-/three-bladed wind turbine*), etc. Accordingly, these slots allow the formation of sets of MWTs [10].

As previously stated, two-term combinations have been the main focus in MWT research, whereas longer sequences have received less specific attention. However, when dependencies are not analyzed, the interpretation of longer MWTs can be erroneous. For instance, in *offshore wind turbine, offshore* modifies *wind turbine.* The bracketing would thus be *offshore* [*wind turbine*], and not [*offshore wind*] *turbine.* In fact, incorrect assumptions regarding this MWT can lead to translation errors. For example, translating *offshore wind turbine* into Spanish as *turbina de viento marino* would imply that *offshore* modifies *wind.* The correct translation would be

aerogenerador marino, which conveys that it is the *turbine* that is located *offshore*. Bracketing is thus a complex process that requires linguistic knowledge, world knowledge, as well as manual or computational techniques. An example of the complexity of this analysis are the MWTs *offshore* [*wind power*] and *wind* [*power output*], where the same combination (*wind power*) participates in two different bracketing structures.

NLP has proposed two models for the bracketing of three-term MWTs: the adjacency model and dependency model. The adjacency model [11, 12] takes an MWT p1p2p3 and compares if p2 is more related to p1 or p3. For that purpose, the number of occurrences of p1p2 and p2p3 are compared. For instance, in *renewable energy technology* there are more occurrences of *renewable energy* than of *energy technology*. Thus, a left-bracketing structure is adopted, and the system interprets the MWT as [*renewable energy*] *technology*. The dependency model [1] also takes an MWT p1p2p3 and compares whether p1 is more strongly associated with p2 or p3. Therefore, the analysis does not start from the central term, as in the adjacency model, but rather from the first one to the left. When p1 is more strongly associated with p2 than to p3, there is a left bracketing ([*tip speed*] *ratio*). In contrast, when p1 is dependent on p3, there is a right bracketing (*mean* [*wind speed*]).

In the same line, [13] states that these types of term are usually characterized by an internal structure that joins words in a binary branching dependency tree. The author adds that these structures govern how the terms can be abbreviated: "*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*" [13, p. 65].

However, following both models the two possible combinations may have a similar frequency. Especially in four-term combinations, this criterion may not be as straightforward. Moreover, in the case of having a small or unbalanced corpus, frequency cannot be the single disambiguating rule. For this reason, the structural disambiguation proposed by these models should rely on additional factors.

Apart from frequency, [7, pp. 19–21] point out other signs that can clarify the dependencies in English MWTs. These include the identification of term variants on the web. If they have the following characteristics (see Table 1), they point to an internal group. [7] also suggest that paraphrases are useful for identifying internal dependencies in MWTs. For instance, *health care reform* is left-bracketed because paraphrases separating those groups can be found, as in "reform *in* health care". The bracketing indicators in [7] are very useful for the disambiguation of English MWTs. However, they may not apply to other languages, such as those not having the possessive genitive or internal inflection.

Additional clues to the structure of MWTs are offered in [5], one of the few studies addressing the bracketing of multi-word expressions of more than three constituents. They argue that internal dependencies are based on relational, coordinating or lexical links. Their proposal is characterized by the use of Wikipedia, as a term and named entity list, and as a corpus merging the information in all its pages [5, p. 72].

To initially determine that certain constituents are linked by a semantic relation, [5] rely on the use of prepositions. For instance, they search for n1 *for* n2 in the corpus. If

Table 1. Bracketing signs in [7, pp. 19–21].

CN to be disambiguated	Term variant	Bracketing indicator	Bracketing structure
cell cycle analysis	*cell-cycle analysis*	hyphen	[*cell cycle*] *analysis*
brain stem cell	*brain's stem cell*	possessive genitive	*brain* [*stem cell*]
plasmodium vivax malaria	*Plasmodium vivax Malaria*	internal capitalization	[*plasmodium vivax*] *malaria*
leukemia lymphoma cell	*leukemia/lymphoma cell*	slash	*leukemia* [*lymphoma cell*][a]
growth factor beta	*growth factor (beta)*	brackets	[*growth factor*] *beta*
tumor necrosis factor	*tumor necrosis factor (NF)*	abbreviation	*tumor* [*necrosis factor*]
health care reform	*healthcare reform*	concatenation	[*health care*] *reform*
adult male rat	*male adult rat*	change of order	*adult* [*male rat*][b]
tyrosine kinase activation	*tyrosine kinases activation*	internal inflection	[*tyrosine kinase*] *activation*

[a]The authors argue that there is right bracketing because the terms separated by a slash are alternatives [7, p. 20].

[b]The authors defend that there is right bracketing because the two first terms separately modify the head.

occurrences are found, n1 and n2 are said to encode a semantic relation and are thus bracketed. Nevertheless, this criterion cannot be applied to specialized discourse, where all MWT constituents usually belong to a concept system, and thus encode different semantic relations.

Therefore, in MWTs such as *offshore wind industry*, there are semantic relations between all of its constituents: industry *located* offshore, industry *uses_resource* wind, and wind *located* offshore. In this sense, further studies could be undertaken to address this issue by establishing a prioritized order in which semantic relations apply in the dependency link. Additionally, [5] argue that the existence of coordinating elements prevents their dependency. Thus, they look for the conjunctions *or, and*, and *nor* to ascertain whether two constituents are coordinated. For example, *cotton* and *polyester* are coordinated, because both terms are usually linked by those conjunctions. Consequently, in *cotton polyester shirt* those elements are not bracketed. This example is similar to the change of order mentioned in [7] (e.g. *adult male rat*, which is right-bracketed, see Table 1).

Although both studies highlight that the modifiers of these examples individually complement the head, which is undoubtedly true, in our opinion, these modifiers should be bracketed since none of the modifiers is more closely linked to the head than the other. That is why none of them should primarily be grouped with the head. Furthermore, bracketing is an internal grouping mechanism that facilitates analysis. For this reason, we argue that modifiers should be bracketed in this case, representing thus a coordinate MWT (*cotton polyester*) inside a broader MWT ([*cotton polyester*] *shirt*).

Finally, [5] search for determinants and plural uses of the elements to ascertain their lexical links. In *cotton polyester shirt*, for example, a large number of occurrences of *the cotton shirts* was found. The fact that considerably fewer occurrences of *the cotton polyesters* were retrieved suggests that *cotton* and *polyester* separately modify the head *shirt*. Accordingly, they relied on the presence of the possible combination of these elements in a lexical resource, i.e. Wikipedia. However, this criterion does not appear to be as useful for all MWTs since lexical resources can lack many terms. Moreover, in the same way as frequency, we can also find entries for the different possible combinations.

In short, more than twenty years after the development of bracketing models, structural disambiguation still remains problematic. In particular, this applies to combinations of more than three elements, which apart from [5], have not been studied in any depth. A list of bracketing indicators is also necessary for MWTs in specialized texts, formed by three or more constituents. In specialized discourse, these complex terms have specific characteristics that complicate their dependence analysis, such as the formation of long MWTs and the deletion of some constituents. Still another factor is their internal semantic relation, which can be domain specific (*uses_resource*). Bracketing trends should also be explored. In addition, bracketing is also useful for terminologists, translators and other linguists that rarely have access to NLP systems. Thus, manual techniques are a viable solution for these professionals.

3 Materials and Methods

For the purposes of this research, a corpus on wind power of approximately 3 million words was manually compiled. It consisted of specialized texts, such as scientific articles and PhD dissertations, originally written in English. The corpus was analyzed in Sketch Engine (https://www.sketchengine.eu/) [14], a corpus analysis tool that can generate concordance lines, wordlists, and word sketches (frequent word combinations), among many other utilities. We performed the following CQL (Corpus Query Language) queries to extract three- and four-term MWTs, respectively:

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

On the rightmost part of the queries, a nominal head is specified as [tag="N.*"], which can be preceded by nouns (N.*), adjectives (JJ.*), adverbs (RB.*), past participles (VVN.*) or present participles (VVG.*) on the order of two for three-term MWTs ({2}) (first CQL query) and three in the case of four-term MWTs ({3}) (second CQL query). This query was based on the different elements that have been found to premodify[1] nouns in English MWTs, which have been analyzed in previous work. We selected three-term MWTs because bracketing is necessary for this number of

[1] Although MWTs can also be postmodified (*angle of attack*), premodification is the preferred MWT formation pattern [3, 15, 16].

constituents. In addition, they are the most frequent MWTs of those that require bracketing. Four-term MWTs were also addressed because of their greater difficulty, since more patterns are possible. However, our results showed that the more constituents an MWT has, the fewer occurrences are found.

After verifying the terms in concordance lines, those MWTs that had been erroneously extracted were rejected. Examples of discarded MWT candidates were *axis wind turbine*, which always appeared as *horizontal/vertical axis wind turbine* (thus, they were included in the list of four-term MWTs), or *different wind penetration level*, which was a three-word term modified by an adjective. In contrast, there are other apparently similar MWTs that were nonetheless selected because the adjective modifying the head was not only an attribute but also conveyed specialized knowledge. For example, *large wind farm* was considered a concept because *large* actually points to power capacity, as in "In North America, typically, wind farms are larger than 50MW, with some projects of up to 200MW".

Other MWTs, in particular those referring to named entities, were also rejected, since these concepts will be addressed in future work. In the end, we focused on a set of 100 MWTs composed of the 50 most frequent three- and four-term MWTs (Table 2). As can be observed, four-term MWTs suffered a drastic reduction of frequency, which complicated the extraction of results.

Table 2. Sample of most frequent three and four-term MWTs in the corpus

Three-term MWTs	Freq.	Four-term MWTs	Freq.
Offshore wind farm	1024	Horizontal axis wind turbine	129
Tip speed ratio	445	Wind power generation system	105
Wind power plant	419	Installed wind power capacity	101
Wind power generation	374	Doubly feed induction generator	84
Wind power capacity	333	Vertical axis wind turbine	68
Mean wind speed	311	Offshore wind power plant	58
Wind power production	298	Annual mean wind speed	56
Average wind speed	284	Design tip speed ratio	55
Offshore wind turbine	281	Variable speed wind turbine	50
Renewable energy source	265	Wind power forecast error	48
Offshore wind power	264	Large offshore wind farm	46
Offshore wind energy	213	Annual average wind speed	46
Wind energy system	211	Squirrel cage induction generator	43
Small wind turbine	210	Blade root bending moment	39
High wind speed	199	Wind power output fluctuation	38
Variable-speed wind turbine	184	Micro hydro power plant	36
Rated wind speed	183	Permanent magnet synchronous generator	33
Large wind farm	179	Wind energy conversion system	30
Onshore wind farm	177	PMSG wind turbine system	29
Wind turbine blade	170	Optimum tip speed ratio	25
Wind power output	165	Hub height wind speed	25

(*continued*)

Table 2. (*continued*)

Three-term MWTs	Freq.	Four-term MWTs	Freq.
Low wind speed	157	Wind power penetration level	24
Wind turbine rotor	156	Small signal stability analysis	24
Large wind turbine	149	Gross final energy consumption	24
Wind energy converter	148	Wound rotor induction generator	22
Wind turbine system	147	Axial flow induction factor	22
Installed wind power	145	Wind turbine drive train	21
Wind turbine design	144	Wind turbine control system	21
Wind penetration level	144	Wind speed time series	21
Wind speed data	138	Power factor correction capacitor	21
Novel wind turbine	134	Short-term wind speed forecasting	20
Domestic hot water	127	Offshore wind supply chain	19
Power generation system	126	Average wind power density	19
Offshore wind market	125	Offshore wind energy resource	17
Renewable energy technology	121	Average annual wind speed	17
Wind power penetration	120	Optimal tip speed ratio	16
Wind power forecast	120	Wind energy penetration level	15
Wind power development	117	Wind turbine power curve	14
Total installed capacity	115	Wind farm power output	14
Conventional power plant	115	Free stream wind speed	14
Power system reliability	113	Fixed speed wind turbine	14
Offshore wind project	113	Wind turbine power production	13
Wind turbine model	111	Total wind power capacity	13
Power electronic converter	111	Offshore wind power development	13
Wind turbine generator	108	Insulated gate bipolar transistor	13
Sound pressure level	108	Hourly mean wind speed	13
Wind turbine manufacturer	106	Constant tip speed ratio	13
Wind energy project	105	Wind power grid integration	12
Wind power fluctuation	76	Short-term mean wind speed	12
Heat transfer medium	74	Network impedance phase angle	12

After extracting the list of terms to be analyzed, new queries were performed in order to disambiguate all possible groupings. These queries were aimed at researching different occurrences and frequencies based on the bracketing models described in the literature and our own observations:

1. Based on [9], we checked whether all adjacent groupings were found in the corpus intersected by external elements: $p1 * p2 + p3$; $p1 + p2 * p3$ (in the case of three-term MWTs) and $p1 * p2 + p3 + p4$; $p1 + p2 * p3 + p4$; $p1 + p2 + p3 * p4$; $p1 * p2 * p3 + p4$; $p1 * p2 + p3 * p4$; $p1p2 * p3 + p4$ (added in the case of four-term MWTs).
2. In contrast, we checked whether all adjacent groupings, while sticking together, were combined with other modifiers and/or heads: $*p1 + p2$; $p1 + p2*$; $*p2 + p3$;

p2 + p3∗; ∗p1 + p2∗; ∗p2 + p3∗ (in the case of three-term MWTs) and ∗p1 + p2; p1 + p2∗; ∗p2 + p3; p2 + p3∗; ∗p3 + p4; p3 + p4∗; ∗p1 + p2∗; ∗p2 + p3∗; ∗p3 + p4∗; ∗p1 + p2 + p3; p1 + p2 + p3∗; ∗p2 + p3 + p4; p2 + p3 + p4∗; ∗p1 + p2 + p3∗; ∗p2 + p3 + p4∗ (added in the case of four-term MTWs).

3. According to the adjacency model [11, 12], for three-term MWTs, we checked and compared the occurrence and frequency of the following combinations in the corpus: p1 + p2; p2 + p3. Evidently, in the case of four-term MWT, the searches were: p1 + p2; p2 + p3; p3 + p4; p1 + p2 + p3; p2 + p3 + p4.
4. According to the dependency model [1], for three-term MWTs, we checked and compared the occurrence and frequency of p1 + p3 and p1 + p2; whereas for four-term MTWs, we compared p1 + p2; p2 + p3; p1 + p4; p1 + p3; p2 + p4; p1 + p2 + p3; p1 + p2 + p4;p1 + p3 + p4.
5. According to the shortening model [13], we checked and compared the occurrence and frequency of p2 + p3 and p1 + p3 (for three-term MWTs) and p3 + p4; p1 + p4; p2 + p4; p2 + p3 + p4; and p1 + p3 + p4 (for four-term MWTs).
6. Based on the bracketing signs of [7], together with synonymic and antonymic patterns, we checked whether all possible groupings showed any variants or antonyms in the corpus (e.g. p1+p2 *also known as* x).

These queries are illustrated and discussed in more detail in Sect. 4.

4 Experiments and Discussion

4.1 Bracketing indicators

The analysis of the terms in Table 2 and the queries described in Section 3 produced the following list of indicators to perform MWT structural disambiguation tasks in a prioritized order. Unfortunately, to the best of our knowledge, there is no baseline available on specialized terms with which to compare the precision of our method. Therefore, the verification of our results relies on the specialized knowledge acquired during the development of EcoLexicon, a terminological knowledge base on the environment [17, 18].

This first indicator is most often sufficient to infer the bracketing, especially in three-term combinations. However, since specialized corpora are not always available or are sufficiently representative, the rest of the indicators can also be used to confirm the results of the first one. Furthermore, in MWTs consisting of more than three terms more disambiguation steps may be required, since when there are more possible combinations, the results are also more contradictory.

Therefore, for a combination of two or more elements to be grouped together (e.g. *large* [*wind farm*]), the candidate MWT should comply with at least two of the following requirements in the following order:

Step 1: Adjacent Groupings Within the MWT Appear as Independent Terms in the Corpus. MWTs formed by more than two components are usually the combination of other (shorter) MWTs (*wind power* + *power output* = *wind power output*). Thus, these MWTs integrate different concepts combined in a single more complex concept

[15]. Adjacency models are based on this MWT property, which constitutes the main criterion for structural disambiguation (e.g. lexical links in [5]). However, very often, because of the effects of compositionality in specialized language, all possible adjacent combinations appear in the corpus. However, a significant frequency difference among the results can resolve the ambiguity.

In this step, the identification of all possible groupings in the corpus is performed through a query that searches for each adjacent grouping with no other elements susceptible of forming larger compounds. In this way, the starting MWT or other MWT consisting of more than two components are excluded. This is exemplified in the following queries for the MWT *offshore wind turbine*, which search for *offshore wind* and *wind turbine* isolated from other elements, on the left ([tag!="JJ.*|N.*|RB.*|VVG.*|VVN.*"]) and on the right ([tag!="N.*|JJ.*"]).

```
[tag!="JJ.*|N.*|RB.*|VVG.*|VVN.*"][lemma="offshore"][lemma="wind"][tag!="N.*|JJ.*"]
[tag!="JJ.*|N.*|RB.*|VVG.*|VVN.*"][lemma="wind"][lemma="turbine"][tag!="N.*|JJ.*"]
```

In the case of four-term MWTs the same types of queries are performed. This involves adding or reducing lemmas for all possible combinations, which can consist of two or three elements. For example, in *power factor correction capacitor*, all possible combinations are *power factor*, *factor correction*, *correction capacitor*, *power factor correction* and *factor correction capacitor*. Thus, five different queries need to be made.

If only one of the combinations is found in the corpus, the bracketing is straightforward. For instance, for *rated wind speed* two queries searched for *rated wind* (with no results) and *wind speed* (with 1,974 occurrences). Therefore, the bracketing is undoubtedly *rated* [*wind speed*]. In addition, if one of the combinations is clearly more frequent than the other, the most frequent combination drives the bracketing. For example, for *wind turbine system*, *wind turbine* shows 4,567 results and *turbine system* 10. The resulting bracketing is thus [*wind turbine*] *system*.

Nevertheless, if the corpus is not sufficiently large or representative, and especially in the case of four-term MWTs, new indicators should be sought. In the same way, when frequency differences are not significant, further queries should be made. For instance, the queries for *offshore wind power plant* do not show conclusive results (Table 3), since different groupings show similar results. Thus, it is still unknown whether the bracketing is [*offshore wind*] [*power plant*], *offshore* [*wind power plant*] or [*offshore wind power*] *plant*.

Table 3. Frequencies of possible bracket groupings in *offshore wind power plant*

Offshore wind power plant	Freq.
Offshore wind	426
Wind power	2568
Power plant	226
Offshore wind power	104
Wind power plant	262

Something similar can also occur with a particular type of MWT, such as *power generation system*. In this type of MWTs the middle word is the nominalization of a predicate that explicitly codifies the semantic relation between the other two components. In this case, the system *generates* power. Therefore, both structures (*power generation* and *generation system*) can be expected to be equally frequent and other criteria should be applied for disambiguation.

Step 2: The Most Frequent Adjacent Grouping is Still More Frequent Than Other Dependencies. This is in consonance with the dependency model and uses the same types of query in the corpus. For example, in order to corroborate that *wind turbine system* is bracketed as [*wind turbine*] *system*, *wind turbine* should again show more occurrences (4,567) than *wind system* (146), which is the case. A variant of the original dependency model is the shortening technique as mentioned above in the words of [13], which in the case of *wind turbine system* would involve comparing the results of *wind system* (146) and *turbine system* (10). As in the case of *civil rights activist*, if the bracketing is [*wind turbine*] *system,* the term could be shortened to *turbine system* and not *wind system*, since *turbine* is the head of the bracketed nominal group. However, the results point in the wrong direction, contradicting the two previous indicators (Table 4).

Table 4. Adjacency, dependency and shortening indicators for *wind turbine system* and *wind power output*

Wind turbine system			
Adjacency	wind turbine (4,567)	turbine system (10)	[wind turbine] system
Dependency	wind turbine (4,567)	wind system (146)	[wind turbine] system
Shortening	wind system (146)	turbine system (10)	wind [turbine system]
Wind power output			
Adjacency	wind power (2,568)	power output (421)	[wind power] output
Dependency	wind power (2,568)	wind output (10)	[wind power] output
Shortening	wind output (10)	power output (421)	[wind power] output

In our study, this last model was not worth including in the protocol, since most of the time, the results led to more confusion than clarification. There were certain MWTs that comply with all of these three criteria, such as *wind power output* (Table 4), but most of them were not helpful. Therefore, this indicator should be included, if at all, in the protocol, as a last resort, and only in combination with others. Further studies will explore whether the MWTs, for which the shortening technique does not work, are different those to which the three criteria apply (*wind turbine system* vs. *wind power output*).

Step 3: Bracketing Groupings Do Not Allow the Insertion of External Elements Modifying Their Meaning. This determines its lexical unit [9] and its "concepthood", since there are a few cases where external elements can be inserted but no meaning change occurs in the combination. For example, in *wind power plant*, occurrences of

wind power generation plant were not regarded as insertions since meaning is not affected. In fact, *power plant* could still be a bracket grouping because *generation* is a frequently omitted component of term that makes the relationship between power and plant explicit. In contrast, MWTs such as *low wind speed* are found to be "broken" in examples such as *low average wind speed* or *low cut-in wind speed*, where *low* and *wind* can already be ruled out from the bracketing.

In order to find possible insertions among the elements, we apply the following queries, exemplified with the MWT *offshore wind farm*:

```
[lemma="offshore"][]{1,3}[lemma="wind"][lemma="farm"] within <s/>
[lemma="offshore"] [lemma="wind"] []{1,3}[lemma="farm"] within <s/>
```

[]{1,3} identifies possible insertions of one to three words among the elements of the MWT; within <s/> is included to make sure that all occurrences are extracted within the same sentence. If only one of the possible groupings complies with this rule, the bracketing is solved. For instance, in *large wind farm*, the distinction between [*large wind*] *farm* and *large* [*wind farm*] is clear when no elements are found between *wind* and *farm* and several elements are found between *large* and *wind*: *large offshore wind farm*, *large onshore wind farm*, *large commercial wind farm*, etc.

If more than one possible grouping complies with the rule, or none of them complies with it, other criteria should also be applied. This indicator is very reliable, but it is not the first one because very often no external elements can be found among the components of the MWT, which indicates the fixed degree of lexicalization of many idiomatic specialized terms.

Step 4: Bracketing Groupings are Found Combined With Other Elements. This indicator is the opposite of the previous one, as it shows how a bracketing grouping represents a concept and can thus be combined with other modifiers and heads. This means that the relation with the rest of the MWT is not necessary, marking the boundaries of the bracketing. This also represents the previously mentioned slotopening mechanism.

The following queries were performed, exemplified with the MWT *high wind speed*:

```
[lemma="high"][lemma="wind"][tag="N.*" & lemma!="speed"]
[tag="JJ.*|N.*|RB.*|VVG.*|VVN.*"][lemma="high"][lemma="wind"][lemma!="speed"]
[lemma!="high"][lemma="wind"][lemma="speed"][tag="N.*"]
[tag="JJ.*|N.*|RB.*|VVG.*|VVN.*" & lemma!="high"][lemma="wind"][lemma="speed"]
```

In the first query, *high wind* is searched for followed by any other noun than speed ([tag="N.*" & lemma!="speed"]) in order to extract other MWT where *high wind* is combined with different heads. In the second query, *high wind* is searched for in combination with new modifiers and, again, a different head. The other two queries do the same with *wind speed*. For four-term MWT, the same queries are used, expanding

or reducing the number of lemmas according to all possible combinations. For example, in *optimum tip speed ratio*, one of the possible groupings is *tip speed ratio*. Therefore, the first query is the following:

[lemma!="optimum"][lemma="tip"][lemma="speed"][lemma="ratio"][tag="N.*"]

If only one of the possible groupings appears to be combined with other elements, or if frequency figures differ in a meaningful way, the bracketing is solved. For instance, in *wind turbine system*, *wind turbine* appears integrated within a total of 6,509 other MWTs: *wind turbine access*, *wind turbine aerodynamics*, *wind turbine airfoils*, *wind turbine blade*, *wind turbine certification*, etc. and *synchronous wind turbine generator*, *DIG-based wind turbine*, *fixed-speed with turbine*, etc. In contrast, *turbine system* only appears within other MWTs in 10 occurrences (e.g. *shrouded turbine system's performance*, *turbine system model* and *horizontal axis turbine system*, *air-driven turbine system*).

If all possible groupings, or none of them, can be integrated with other MWTs, the next indicators should be applied. For example, all the possible groupings of *offshore wind energy resource*, can be combined within other MWTs. Moreover, contradictory groupings show very similar frequencies (e.g. *offshore wind* and *wind energy*) (Table 5).

Table 5. Possible groupings in *offshore wind energy resource* and other combinations

Offshore wind energy resource		
Possible groupings	Integration within other MWTs	Frequency
Offshore wind	-offshore wind plant/industry/technology -installed offshore wind capacity/global offshore wind market	3345
Wind energy	-wind energy capacity/development -available/land-based wind energy	2292
Energy resource	-energy resource availability/development -renewable/clean energy resource	133
Offshore wind energy	-offshore wind energy potential/project -future/global offshore wind energy market/plant	96
Wind energy resource	-wind energy resource assessment/study -onshore/European/future wind energy resource	50

Step 5: Bracketing Groupings Have Synonyms or Antonyms. This indicator also highlights the "conceptthood" of the groupings, since many concepts show denominative variations (e.g. *wind energy*, *wind power*) or antonyms (*horizontal axis wind turbine* vs. *vertical axis wind turbine*) that reinforce the conceptual nature of the possible combinations. It is related to term-formation mechanisms (different concepts emerge by opposition: *small* [*wind turbine*], *large* [*wind turbine*]) and to slot filling (*power* in this case).

According to [7], some of these variants could include the possessive genitive, slashes, hyphens or acronyms, and monolexical variants would be the most reliable proof of the conceptthood of one of the possible groupings. For example, in *insulated-gate bipolar transistor* and *power factor correction (PFC) capacitor*, the hyphen and the acronym disclose the bracketing mechanism. The following sentence extracted from the corpus reveals *wind turbine* to be the bracket grouping in most of the MWTs where it takes part, since it has a monolexical variant (*aerogenerator*): "Wind turbines or aerogenerators transform the kinetic energy of the wind into electrical energy".

Therefore, apart from searching for possible punctuation-related issues that could unfold [7]'s variants, synonyms in the corpus were also searched through queries based on the knowledge patterns that usually convey synonymy in real texts:

[tag="RB.*"]?[word="known|called|referred"][tag="RB.*"]?[word="to"]?[word="as"]? [tag="N.*"]
[be]?[]{0,3}[lemma="synonym"] [word="of|for|to"] [tag="N.*"]
[tag="N.*"] []{1,3} [word="spelling"] [word="of"] []{0,3}

If only one of the possible groupings shows synonyms or antonyms, bracketing is solved. For instance, from all possible groupings in *offshore wind power plant*, *wind power plant* is related to a synonym in the corpus (*wind power station*), as codified in the pattern *also known as*: "Wind turbines can be installed as individual units or grouped in wind power stations, also known as wind power plants". Therefore, the bracketing *offshore* [*wind power plant*] is chosen. Regarding antonyms, in *reactive power compensation*, one of the possible groupings (*reactive power*) has an antonym (*active power*), which means that the bracketing is [*reactive power*] *compensation*.

This indicator is quite reliable but is suggested as the last step because explicit synonymy or antonymy are not as frequent as other phenomena in specialized corpora. The queries could thus be complemented with the consultation of external terminological resources.

4.2 Case study: *Wind Farm Power Output*

In order to illustrate the application of the previous indicators in a protocol-like manner (on the analysis of the same MWT), we present the example of *wind farm power output*. The possible combinations were the following:

[*wind farm*] [*power output*]
wind[*farm power*]*output*
[*wind farm power*]*output*
wind[*farm power output*]

After applying the first and second indicators, i.e. adjacency and dependency rules, the four possibilities were reduced to two: [*wind farm*] [*power output*] (whose groups had respectively 2,498 and 421 occurrences) and [*wind farm power*] *output* (*wind farm power* having 3 occurrences). This criterion initially seemed to be conclusive. However, we carried out the remaining steps of the procedure for further confirmation.

The third indicator, i.e. the impossibility for external elements to be inserted in the possible combination, also pointed to this organization. External elements were found between *wind farm* and *power output*, which were of the order of 13, as in *wind farm electrical power output* and *wind farm average power output,* or propositional examples such as *wind farm ramps its power output up* and *wind farm will produce similar power output*. On the contrary, external elements were not found between *wind farm power* and *output* in the other possible structure: [*wind farm power*] *output*. This suggested that *power output* should not be separated.

The fourth indicator, i.e. the formation of other MWTs, was also conclusive since the two groups in [*wind farm*][*power output*] formed many more MWTs than the grouping in [*wind farm power*] *output*. In particular, *wind farm* formed 3,786 MWTs (e.g. *wind farm design, commercial wind farm*) and *power output* was present in 489 MWTs (e.g. *power output fluctuation, real power output*), while *wind farm power* formed 13 MWTs (e.g. *wind farm power prediction, real wind farm power curve*). The preferred structure was again [*wind farm*][*power output*].

Finally, according to the fifth indicator, we searched for synonyms or antonyms of the possible combinations, especially monolexical variants. Although no synonyms or antonyms were found for *power output* (supporting the structure [*wind farm*][*power output*]) nor for *wind farm power* (in the structure [*wind farm power*] *output*), synonyms of *wind farm* were retrieved (e.g. *wind power plant, wind power station*). These highlighted the conceptual link of these elements and confirmed the validity of the structure [*wind farm*][*power output*].

5 Conclusions

In this paper, we have addressed the structural disambiguation of MWTs based on the information extracted from a specialized corpus on wind power. We have extracted and analyzed a set of 100 MWTs. This has led us to study the most frequent bracketing structures as well as the most reliable methods to perform disambiguation. Based on these methods, we devised a set of indicators and steps in order to disambiguate the structural dependencies of all possible combinations.

As a future line of research, we plan to enhance these structural indicators with more semantic-oriented criteria, since the relation held between the components of an MWT could also enrich the protocol. Moreover, the combination of structural and semantic criteria would lead to a more accurate interpretation of MWTs since semantic elicitation depends on bracketing, and bracketing can also benefit from semantic-based indicators. Other lines of research will also include the testing of the protocol in other languages and types of multiword expressions.

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 an FPU grant given by the Spanish Ministry of Education to the first author.

References

1. Lauer, M.: Designing statistical language learners: experiments on noun compounds. Ph.D. Macquarie University, Australia (1995)
2. Girju, R., Moldovan, D., Tatu, M., Antohe, D.: On the semantics of noun compounds. Comput. Speech Lang. **19**(4), 479–496 (2005)
3. Nakov, P.: On the interpretation of noun compounds: syntax, semantics, and entailment. Nat. Lang. Eng. **19**(03), 291–330 (2013)
4. Kim, S.N., Baldwin, T.: A lexical semantic approach to interpreting and bracketing English noun compounds. Nat. Lang. Eng. **19**(3), 385–407 (2013)
5. Barrière, C., Ménard, P.A.: Multiword noun compound bracketing using Wikipedia. In: Proceedings of the First Workshop on Computational Approaches to Compound Analysis, Dublin, Ireland, pp. 72–80 (2014)
6. Marsh, E.: A computational analysis of complex noun phrases in navy messages. In: Proceedings of the 10th International Conference on Computational Linguistics, Standford, CA, pp. 505–508 (1984)
7. Nakov, P., Hearst, M.: 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, Ann Arbor, MI, pp. 17–24 (2005)
8. Utsumi, A.: A semantic space approach to the computational semantics of noun compounds. Nat. Lang. Eng. **20**, 185–234 (2014)
9. Johnston, M., Busa, F.: Qualia structure and the compositional interpretation of compounds. In: Viegas, E. (ed.) Breadth and Depth of Semantic Lexicons, pp. 167–187. Springer, Dordrecht (1999). https://doi.org/10.1007/978-94-017-0952-1_9
10. Cabezas-García, M., Faber, P.: A semantic approach to the inclusion of complex nominals in english terminographic resources. In: Mitkov, R. (ed.) EUROPHRAS 2017. LNCS (LNAI), vol. 10596, pp. 145–159. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69805-2_11
11. Marcus, M.: A Theory of Syntactic Recognition for Natural Language. MIT Press, Cambridge (1980)
12. Pustejovsky, J., Anick, P., Bergler, S.: Lexical semantic techniques for corpus analysis. Comput. Linguist. **19**(2), 331–358 (1993)
13. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic Press, Boston (1994)
14. Kilgarriff, A., et al.: The sketch engine: ten years on. Lexicography **1**(1), 7–36 (2014)
15. Levi, J.: The Syntax and Semantics of Complex Nominals. Academic Press, New York (1978)
16. Sager, J.C., Dungworth, D., McDonald, P.F.: English Special Languages. Principles and Practice in Science and Technology. Brandstetter Verlag, Wiesbaden (1980)
17. Faber, P.: A Cognitive Linguistics View of Terminology and Specialized Language. De Gruyter Mouton, Berlin/Boston (2012)
18. San Martín, A., Cabezas-García, M., Buendía, M., Sánchez-Cárdenas, B., León-Araúz, P., Faber, P.: Recent advances in EcoLexicon. Dictionaries: J. Dictionary Soc. North Am. **38**(1), 96–115 (2017)