Semantic annotation to characterize contextual variation in terminological noun compounds: a pilot study

Melania Cabezas-García
LexiCon Research Group
University of Granada
Granada, Spain
melaniacabezas@ugr.es

Antonio San Martín
LexiCon Research Group
Maynooth University
Maynooth, Ireland
antonio.sanmartin@nuim.ie

Abstract

Noun compounds (NCs) are semantically complex and not fully compositional, as is often assumed. This paper presents a pilot study regarding the semantic annotation of environmental NCs with a view to accessing their semantics and exploring their domain-based contextual variation. Our results showed that the semantic annotation of NCs afforded important insights into how context impacts their conceptualization.

1 Introduction

In English, noun compounds (NCs) are the lexical units that are most often used to convey expert knowledge (Daille et al., 2004; Nakov, 2013; Hendrickx et al., 2013). Terminological NCs can be considered a type of multi-word term (MWT) because they are non-idiomatic multi-word units that belong to a specialized domain and lie in the intersection between terms and multi-word expressions (MWEs) (SanJuan et al., 2005; Frantzi et al., 2000; Ramisch, 2015). They are characterized by their semantic complexity since two or more concepts are juxtaposed without any explicit indication of the relation linking them (Ó Séaghdha and Copestake, 2013). This relation is determined largely by the context and the frame (i.e. system of concepts related in such a way that one concept evokes the entire system (Fillmore, 1982)) to which the NC belongs. In other words, they are not fully compositional and their conceptualization can differ depending on the context and the semantic frame in which it is embedded.

This paper describes the use of semantic annotation to explore how domain-based context modulates the meaning of NCs. To this end, the annotated concordance lines were used to identify and analyze the argument structure of the propositions underlying this kind of MWT. The micro-contexts (i.e. the relation of a predicate with its arguments and adjuncts) are directly related to the semantic load of the compound term, because they specify the hidden relation between its components (Cabezas-García and Faber, 2016).

In the following section, a short account of the particularities of NCs and the phenomenon of contextual variation is provided. Then, section 3 describes the materials and methods used in this pilot study. Section 4 expounds the results of the study and discusses their significance. Finally, section 5 presents the conclusions derived from this research and mentions the issues that will be addressed in future work.

2 Contextual Variation in Noun Compounds

2.1 Noun Compounds

NCs are very frequent in specialized texts written in English (Daille et al., 2004; Nakov, 2013; Hendrickx et al., 2013). They are a sequence of nouns that function as a single noun (Downing, 1977), e.g., *water loss* or *population growth*. In endocentric NCs, one term is the head and the other is its modifier (Nakov, 2013) (e.g., *power generation*). Alternatively, in exocentric NCs, the MWT is not a hyponym of one of its elements, and thus appears to lack a head (Bauer, 2008) (e.g. *saber tooth*). Endocentric NCs (the focus of this study) are characterized by their (i) headedness; (ii) transparency, (iii) syntactic ambiguity; and (iv) language-dependency (Nakov, 2013).

NCs have underlying propositions, which can be inferred by the term formation processes highlighted in Levi (1978), involving predicate deletion (e.g. *power system*, instead of a *system produces power*) and predicate nominalization (e.g.

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 108–113,
Valencia, Spain, April 4. ©2017 Association for Computational Linguistics

heat transfer, instead of *heat is transferred*). These propositions underlying the NCs take the form of a predicate with its arguments, which are necessary for the meaning of the verb, and its adjuncts (optional complements) (Tesnière, 1976). The relation of a predicate to its argument structure is known as micro-context. This is a key factor that provides access to the conceptual load of terms, since the predicate, which is the syntactic-semantic core of the sentence, can only be successfully addressed through its complement structure (Cabezas-García and Faber, 2016).

2.2 Contextual Variation

The notion of context plays a crucial role in various disciplines that employ it in different ways. In this paper, context refers to any factor that affects the interpretation of a sign or an expression (Kecskes, 2014). This sense includes linguistic factors (different types of co-text), discursive factors (channel, communicative purpose, degree of formality, topic, and level of specialization), sociocultural factors (social activity in which communication is embedded, and the relation between participants) as well as spatiotemporal factors (San Martín, 2016).

Lexical units do not carry meaning in themselves, but rather trigger the mental representation of meaning in context (Fauconnier, 1994). Meaning is construed in every usage event. Depending on the context, certain segments of the knowledge conventionally associated with a lexical unit are activated and give rise to meaning. Therefore, meaning does not exist outside of context. Without contextual restrictions, lexical units can be said to have semantic potential, which is all the conceptual content that a lexical unit is capable of invoking (Evans, 2009). The semantic potential of a lexical unit constitutes a considerable amount of information, all of which is never fully activated in a single use event. It includes one or more concepts and their underlying conceptual frames.

Given that context is never identical, the meaning of a lexical unit is variable. This phenomenon by which the semantic potential of a lexical unit produces different meanings depending on the context is called contextual variation. Although in practice, it is sometimes difficult to distinguish between a high degree of contextual variation and polysemy, these two phenomena are theoretically different. Polysemy occurs when the semantic potential of a lexical unit refers to more than one concept. For example, *organism* is a polysemic term because it designates two different concepts: ORGANISM (living being) and ORGANISM (system or organization). In contrast, *ozone* is an example of contextual variation because it designates a single concept (OZONE). When *ozone* appears in the context of Atmospheric Science, it is conceptualized as an important allotropic form of oxygen that is present in the atmosphere. However, in the context of Water Treatment and Supply, it is conceived as a powerful virucidal agent used to disinfect water.

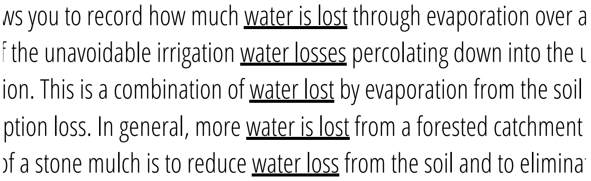
In this paper, we focus on domain-based contextual variation because discourse topic is the contextual factor that best predicts how the semantic potential of a term is restricted in actual usage events (San Martín, 2016). In our analysis, *domain* is synonymous to *knowledge field*.

3 Materials and Methods

A corpus of English texts on environmental science was manually compiled. The corpus consisted of 4,743,025 tokens, and was composed of 16 subcorpora of specialized and semi-specialized texts. Each subcorpus had approximately 300,000 tokens and focused on a specific environmental domain (e.g. Agronomy, Hydrology, etc.).

Each subcorpus was uploaded separately to the term extractor TermoStat (Drouin, 2003) (http://termostat.ling.umontreal.ca/). The search was set to complex terms. The 16 resulting lists of terms were automatically compared. In order to ensure representativeness and significant contextual variation, we only retained the two-term NCs designating processes that had a minimum of 10 occurrences in at least three subdomains (i.e. 10 NCs in total). The MWTs chosen were those designating processes because these units have underlying propositions with a clear argument structure, thus enabling the analysis of micro-contexts (i.e. the relation between a predicate and its arguments and adjuncts), which are key factors in the conceptualization of this kind of MWT (Cabezas-García and Faber, 2016). This pilot study focuses solely on the analysis of *water loss*, with a view to developing an annotation protocol for the rest of MWTs.

We also uploaded the corpus to Sketch Engine (Kilgarriff et al., 2014) (https://www.sketchengine.co.uk/), an online corpus analysis application that allowed us to generate concordance lines, which were subsequently processed with an annotation tool. As previously mentioned, NCs designating processes all have underlying propositions. Nakov and Hearst (2006) confirmed that verb paraphrases are useful for disambiguating these compound terms and eliciting their meaning. Thus, in order to access the concordances that allude to the semantics of the MWT in question, we not only downloaded the concordance lines where the NC appeared but also the concordances where paraphrases had been used (see Figure 1).



ws you to record how much water is lost through evaporation over a
f the unavoidable irrigation water losses percolating down into the u
ion. This is a combination of water lost by evaporation from the soil
ption loss. In general, more water is lost from a forested catchment
of a stone mulch is to reduce water loss from the soil and to elimina

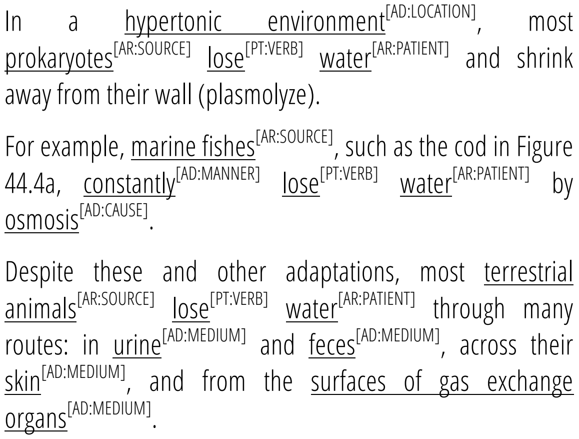
Figure 1: Concordance lines of *water loss* and its verb paraphrases in the domain of Hydrology.

For example, in the case of *water loss*, concordance lines such as "...combination of *water lost* by evaporation from..." were analyzed, as well as those where the NC occurred. This made it possible to access a larger number of examples of the process conveyed by *water loss* (i.e. "a SOURCE ceases to have [LOSE] a PATIENT [WATER]"). The loss process is encoded by verbs conveying a similar meaning though from different perspectives (e.g. *lose*, *evaporate*, *extract*, *release*, etc.).

The next step was the annotation of the concordance lines, following the semantic annotation methodology in L'Homme (2012), which is based on FrameNet (Ruppenhofer et al., 2010). Two human annotators, who established a common tagset and guidelines, annotated the concordance lines with the help of the UAM CorpusTool (O'Donnell, 2008) (http://www.corpustool.com/), an open-source environment for the annotation of text corpora. This tool also allows users to search the corpus, perform statistical studies, analyze file information, etc. The semantic labels [1]used were: (i) PREDICATIVE_TERM, (ii) ARGUMENT, and (iii) ADJUNCT. The predicative term was further specified as VERB or NOUN, and the arguments and adjuncts as AGENT, PATIENT, SOURCE, TIME, LOCATION, RESULT, CAUSE, MANNER, QUANTITY, MEDIUM, DESTINATION, INSTRUMENT, or AIM. The annotation was performed on all the concordance lines given the limited size of the study, but larger annotation tasks would benefit from a selection of contexts, as proposed in L'Homme and Pimentel (2012). Once the texts were annotated, the UAM Corpus Tool software generated summaries of the linguistic designations that filled the arguments and adjuncts slots depending on the contextual domain, and their frequency of occurrence, which were subsequently compared.

4 Results and Discussion

The analysis of the NCs by means of semantic annotation afforded insights into their specific conceptualization for each given contextual domain. Thanks to the annotated concordances, it was possible to compare the conceptualization of the micro-contexts of the NCs in each contextual domain. Particularly, we made use of the automatic generation of lists of the linguistic instantiations that filled each argument and adjunct slots, depending on the contextual domain. This allowed the characterization and analysis of the argument structure of the predicate (see Figure 2).



In a hypertonic environment[AD:LOCATION], most prokaryotes[AR:SOURCE] lose[PT:VERB] water[AR:PATIENT] and shrink away from their wall (plasmolyze).

For example, marine fishes[AR:SOURCE], such as the cod in Figure 44.4a, constantly[AD:MANNER] lose[PT:VERB] water[AR:PATIENT] by osmosis[AD:CAUSE].

Despite these and other adaptations, most terrestrial animals[AR:SOURCE] lose[PT:VERB] water[AR:PATIENT] through many routes: in urine[AD:MEDIUM] and feces[AD:MEDIUM], across their skin[AD:MEDIUM], and from the surfaces of gas exchange organs[AD:MEDIUM].

Figure 2: Annotation of propositions underlying *water loss* in the domain of Biology.

Since the linguistic realizations of the arguments and adjuncts were summarized in the annotation tool, it was possible to compare the conceptualization of the NC, thus allowing the characterization of contextual variation.

Therefore, the semantic annotation of the concordance lines confirmed that contextual variation

[1] It is well-known that the distinction between arguments and adjuncts and the choice of the number and types of semantic labels is problematic. Although this did not cause problems in this work (due to the limited coverage of this pilot study), it is an issue that will be carefully considered in further research.

in NCs is reflected in their argument structure. In other words, the arguments and adjuncts of the predicate underlying a NC, such as *water loss*, were filled by different conceptual categories, depending on the contextual domain.

In regard to *water loss*, the contextual variation was found to manifest itself in the SOURCE of water loss, an argument that is not explicit in the compound. This means that the SOURCE (as reflected in its linguistic designations and those of the adjuncts) varies, depending on the specialized domain. When used in Agronomy, the water loss SOURCE was usually a plant entity (e.g. *plant*, *leaf*, etc.). In contrast, in Hydrology, this SOURCE was generally a waterbody (e.g. *river*, *aquifer*, *lake*, etc.). Finally, in Biology, the preference was for animals (e.g. *animal*, *animal cell*, *blood*, etc.) or some type of living organism. Table 1 shows the linguistic instantiations of the water loss SOURCE in Biology, which highlight the frequency of animal entities in this argument slot.

Category	**Designations**
ANIMAL	*animal* (7), *animal cell* (5), *blood* (3), *filtrate* (3), *egg* (1), *waste* (1), *body* (1), *tissue* (1)
PLANT	*plant* (3), *leaf* (1), *plant cell* (1)
BACTERIA	*prokaryote* (2), *endospore* (1), *Halobacterium cell* (1)
AIR	*air* (2)
SOIL	*soil* (1)

Table 1: Linguistic designations (with frequency of occurrence) filling the SOURCE argument in Biology for *water loss*.

As previously noted, depending on the domain context (Biology, Agronomy or Hydrology), the argument slot (i.e. SOURCE of *water loss*) is designated by a different set of semantically related units. Furthermore, this preference for a specific semantic category in the argument determining the variation (i.e. SOURCE of *water loss*) is reflected in the linguistic realizations of the adjuncts. For example, in Agronomy, the SOURCE argument is filled by plant entities, and the most frequent adjuncts were MEDIUM or CAUSE with linguistic realizations that also belong to the vegetable kingdom: *stoma* and *leaf*, and *transpiration* and *evaporation*, respectively.

Moreover, even though the same NC (*water loss*) sometimes involved the same SOURCE (*waterbody*), its conceptualization was found to have different nuances in each context. For instance, when comparing *water loss from a waterbody* in the domains of Agronomy and Hydrology, it was found that their conceptualizations differed. Whereas in Agronomy texts, *water loss* generally referred to the natural loss of water, in Hydrology texts, *water loss* referred to an artificial process with specific purposes.

This was reflected in the adjuncts and their linguistic realizations. For example, the INSTRUMENT adjunct in Hydrology texts was mainly designated by manmade structures, such as *canal*, *well*, *aqueduct*, *floodgate*, etc. Contextual differences were also evident in the verbs used in the paraphrases. More specifically in Agronomy, the most frequent predicates were *lose*, *evaporate*, *remove*, *transpire*, *absorb*, *draw*, *leave*, and *move*, whereas in Hydrology, there was a preference for predicates with a human/instrument AGENT (e.g. *extract*, *release*, *transmit*, *transfer*, *draw*, *divert*, and *abstract*).

The analysis of micro-contexts and of the linguistic realizations of the arguments and adjuncts was also found to be a useful method for frame-based terminological management (Faber, 2015; L'Homme, 2016). When the argument structure of *water loss* and its linguistic realizations are analyzed, a general picture of the conceptualization of the MWT in each subdomain can be obtained. For instance, this analysis reveals the type of entities that can lose water, the medium in which water is lost, the causes and results of the water loss, etc. For this reason, the identification and annotation of the arguments and adjuncts of the verbs provide insights into the conceptualization of terms and their relations with concepts in larger frames.

5 Conclusions

This research focused on the use of semantic annotation to characterize the micro-contexts that underlie a NC. The results confirmed that contextual variation in NCs designating processes is manifested in their underlying argument structure. Access to the domain-specific conceptualization was accomplished by annotating the NCs as well as the paraphrases that made the hidden verb explicit. This made it possible to identify the conceptual relations between the terms in the compound, which is one of the difficulties of MWTs. Moreover, in regard to the methodology, our results confirmed that the semantic annotation of micro-contexts is an effective technique to study the conceptualization of NCs, namely those representing specialized processes.

In future work, a more in-depth research on the advantages of semantic annotation will be carried out with a view to identifying the role of micro-contexts in NC formation. For the characterization of the different phenomena arising from domain-based contextual variation in MWTs, we also plan to further refine our semantic annotation methodology using WordNet synsets and combine them with the extraction of semantic relations by means of knowledge patterns.

We will also implement the semantic annotation of MWTs for the modeling of this kind of term in the environmental terminological knowledge base EcoLexicon (http://ecolexicon.ugr.es/). Since both endeavors will be multilingual, the results will ultimately be applied to the development of translation rules for MWTs.

Acknowledgements

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

References

Laurie Bauer. 2008. Les composés exocentriques de l'anglais. In D. Amiot, editor, *La composition dans une perspective typologique*, pages 35–47. Artois Presses Université, Arras.

Melania Cabezas-García and Pamela Faber. 2016. Exploring the Semantics of Multi-word Terms by Means of Paraphrases. In *EnTRetextos International Conference on Specialized Translation*, Valencia.

Béatrice Daille, Samuel Dufour-Kowalski, and Emmanuel Morin. 2004. French-English multi-word term alignment based on lexical context analysis. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 919–922.

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

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Vyvyan Evans. 2009. *How Words Mean: Lexical Concepts, Cognitive Models, and Meaning Construction.* Oxford University Press, Oxford.

Pamela Faber. 2015. Frames as a framework for terminology. In Hendrik Kockaert and Frieda Steurs, editors, *Handbook of Terminology*, pages 14–33. John Benjamins, Amsterdam / Philadelphia.

Gilles Fauconnier. 1994. *Mental spaces: aspects of meaning construction in natural language*. Cambridge University Press, Cambridge, New York.

Charles Fillmore. 1982. Frame Semantics. In The Linguistic Society of Korea, editor, *Linguistics in the morning calm. Selected papers from SICOL-1981*, pages 111–137. Hanshin Publishing Company, Seoul.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 Task 4: Free Paraphrases of Noun Compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 138–143.

Istvan Kecskes. 2014. *Intercultural Pragmatics*. Oxford University Press, Oxford, New York.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Ková, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Judith Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.

Marie-Claude L'Homme and Janine Pimentel. 2012. Capturing Syntactico-semantic Regularities among Terms : An Application of the FrameNet Methodology to Terminology. In *Proceedings of LREC 2012*, pages 262–268, Istanbul.

Marie-Claude L'Homme. 2012. Adding syntactico-semantic information to specialized dictionaries: an application of the FrameNet methodology. *Lexicographica*, 28(1):233–252.

Marie-Claude L'Homme. 2016. Terminologie de l'environnement et Sémantique des cadres. In *Congrès mondial de linguistique française (CMLF 2016)*, Tours, France.

Preslav Nakov and Marti A. Hearst. 2006. Using Verbs to Characterize Noun-Noun Relations. *Artificial Intelligence Methodology Systems and Applications*, 4183:233–244.

Preslav Nakov. 2013. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19(03):291–330.

Diarmuid Ó Séaghdha and Ann Copestake. 2013. Interpreting compound nouns with kernel methods. *Natural Language Engineering*, 19(3):331–356.

Mick O'Donnell. 2008. Demonstration of the UAM CorpusTool for text and image annotation. In *Proceedings of the ACL-08:HLT Demo Session*, number June, pages 13–16, Columbus, Ohio.

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

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, and Jan Johnson, Christopher R. Scheffczyk. 2010. *Framenet II: Extended theory and practice*.

Antonio San Martín. 2016. *La representación de la variación contextual mediante definiciones terminológicas flexibles*. Ph.D. thesis, University of Granada.

Eric SanJuan, James Dowdall, Fidelia Ibekwe-SanJuan, and Fabio Rinaldi. 2005. A symbolic approach to automatic multiword term structuring. *Computer Speech and Language*, 19(4):524–542.

Lucien Tesnière. 1976. *Eléments de syntaxe structurale*. Klincksieck, Paris.