Amparo Alcina

REPRESENTING COLLOCATIONS USING ONTOLOGIES

1.   Introduction

Dictionaries, lexicographic or terminographic, sometimes collect some typical uses of some words or terms. But methodology for preparing dictionaries rarely includes the systematic detection of contexts of use. Furthermore, computer tools have not been developed to facilitate the systematic or timely collection of this type of information.

Therefore, it is of great interest to develop and apply models and structures that facilitate the collection of contextual information on terms and their formal representation, that can be useful for both humans and machines, so that terminological resources offer access to this information.

2.   Databases and knowledge bases

Research on the lexicon often requires storing data and processing it in various ways in databases. These systems facilitate the work of storing, maintaining and managing linguistic data; process its ordering by different criteria; manage the filtering, consultation, modification and updating of data; and its display on the screen or its printing in various ways (Cabré Castellví 1999; Sager 1990). They are used in the investigation of the different linguistic levels (phonology, morphology, lexicon, syntax, semantics).

In the last twenty years, these databases have improved in terms of technical aspects (storage capacity, friendlier interfaces), interaction with other systems (automatic extraction of terms, assisted and automatic translation) and in terms of developing standards, such as LMF, TMF, TBX, to ensure lexical data exchange and reuse (Francopoulo (ed.) 2013; Francopoulo et al. 2007; Melby 2015).

However, sometimes, databases do not allow us to model the data in the way we would like, to develop improvements in the access and query of the data or in the organization or to diversify the presentation of the data structures. Some authors point at lexical networks to represent lexical (Lamb 1999; Moerdijk 2008; Moerdijk/Tiberius/Niestadt 2008; Polguère 2014) and terminological (Buendía Castro/Montero Martínez/Faber 2014; Faber 2012) information and relations.

The limitations that the databases present come largely from technical limitations, derived from the fact that their base structure, records and fields, generates a rigid organization of the data in the form of a table.

To achieve more flexible representations of lexical data, it is necessary to abandon the rigidity of the record structures and fields that databases currently employ and move to represent the lexical data with more flexible systems, such as knowledge bases or ontologies.

Ontologies, through their structure of classes, properties and individuals, allow to establish data relationships that result in network structures, much closer to what the representation of the lexicon requires.

3. Representing collocations in an ontology model

In our previous research, the processing of lexical data through ontologies has allowed us to advance in the hierarchical organization of concepts and their detailed analysis in characteristics and attributes. This has been of great help in an onomasiological access to the terms, the visual representation of the structure of their relationships and the elaboration of lexicographically correct definitions (Alcina 2020; Alcina/Valero Doménech 2018).

More recently, we have added to this ontological model the ability to represent the collocations of terms. For this, we have implemented the terms of industrial ceramics and their collocations. From a theoretical point of view, the lexical functions of Explanatory and Combinatorial Lexicography (Jousse/Bouveret 2003; L'Homme 2020; Mel'čuk/Clas/Polguere 1995; Polguère 2003) have helped us to establish the binary relationships between terms.

In this ontological model, the terms have been represented as individuals, and they have been classified as class instances. Linguistic concepts (such as Term, Grammar Category, or Concept Type) are implemented as Classes. Collocations have been implemented as properties that link two terms (individuals). We can create subtypes of collocations using restrictions for these properties following the caracterizations of lexical functions in Meaning Text Theory (such as Oper, Real, etc.), or other lexical semantics theories. In Figure 1, you can see the collocations of the term *absorber* in the ceramic industry terminology.
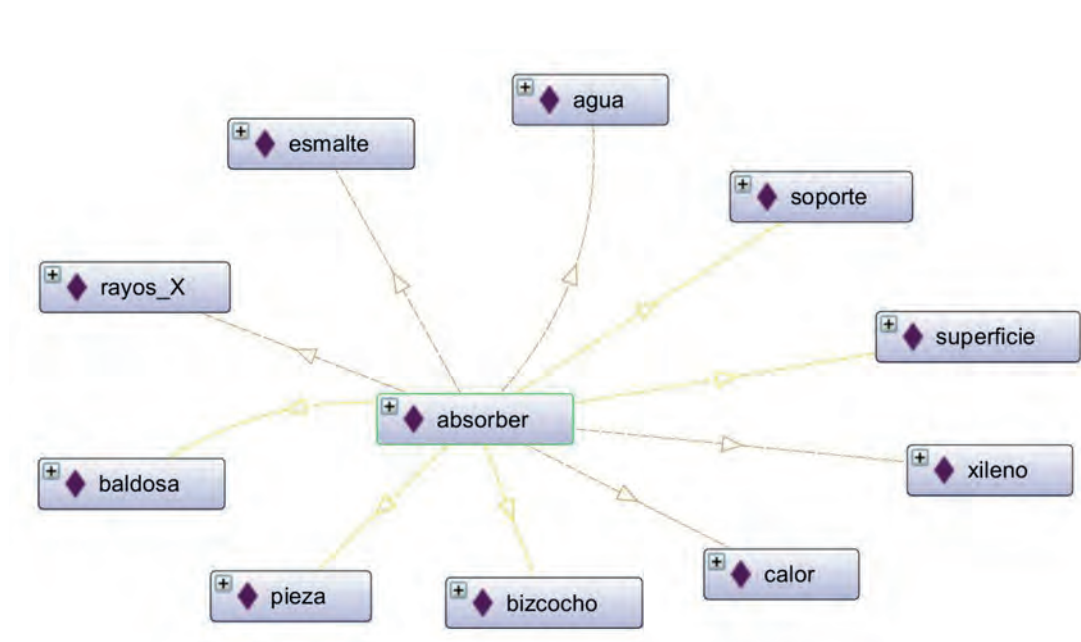
Figure 1: Collocations of 'absorber' in Protégé

We have developed the research using the Protégé ontology editor (Musen 2015). We will present how we have used the tool: the class configuration, properties used to model the collocations of the industrial ceramics language.

References

Alcina, A. (2020): La representación de relaciones conceptuales en una ontología. In: Ibáñez Rodríguez, M. (ed.): Enotradulengua. Vino, lengua y traducción. Berlín.

Alcina, A./Valero Doménech, E. (2018): Description of the terminological concept in an ontology. In: TOTH 2017: Terminologie & Ontologie: Théories et Applications. Chambéry, pp. 161–179. http://hdl.handle.net/10234/188438 (last access: 05-04-2022).

Buendía Castro, M./Montero Martínez, S./Faber, P. (2014): Verb collocations and phraseology in EcoLexicon. In: Yearbook of Phraseology 5 (1), pp. 57–94.

Cabré Castellví, M. T. (1999): Informática y terminología. In: Blecua, J. M./Clavería, G./Sánchez, C./Torruella, J. (eds.): Nuevas tecnologías en los estudios filológicos. Barcelona.

Faber, P. (ed.) (2012): A cognitive linguistics view of terminology and specialized language. Berlin.

Francopoulo, G. (ed.) (2013): LMF lexical markup framework. London.

Francopoulo, G./Bel, N./George, M./Calzolari, N./Monachini, M./Pet, M./Soria, C. (2007): Lexical Markup Framework: ISO Standard for Semantic Information in NLP Lexicons. Tübingen.

Jousse, A. L./Bouveret, M. (2003): Lexical functions to represent derivational relations in specialized dictionaries. In: Terminology 9 (1), pp. 71–98.

L'Homme, M. C. (2020): Lexical semantics for terminology: an introduction. Amsterdam.

Lamb, S. M. (1999): Pathways of the brain: the neurocognitive basis of language. Amsterdam.

Mel'čuk, I./Clas, A./Polguere, A. (1995): Introduction à la lexicologie explicative et combinatoire. Louvaine la Neuve.

Melby, A. K. (2015): TBX: A terminology exchange format for the translation and localization industry. In: Kockaert, H. J./Steurs, F. (eds.): Handbook of terminology, pp. 393–424.

Moerdijk, F. (2008): Frames and semagrams. Meaning description in the General Dutch Dictionary. In: Bernal, E./DeCesaris, J. A. (eds.). Proceedings of the XIII EURALEX International Congress. Barcelona, pp. 561–569.

Moerdijk, F./Tiberius, C./Niestadt, J. (2008): Accessing the ANW dictionary. In: Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008). Manchester, pp. 18–24.

Musen, M. A. (2015): The Protégé project: A look back and a look forward. In: AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence 1 (4), pp. 4–12.

Polguère, A. (2003): Lexicologie et sémantique lexicale. Montreal.

Polguère, A. (2014): From writing dictionaries to weaving lexical networks. In: International Journal of Lexicography 27 (4), pp. 396–418.

Sager, J. C. (1990): A practical course in terminology processing. Amsterdam/Philadelphia.

Contact information

Amparo Alcina
TecnoLeTTra Research Group, Universitat Jaume I
alcina@uji.es

Acknowledgements