TERM VARIANTS IN ONTOLOGIES

GUADALUPE AGUADO DE CEA
ELENA MONTIEL-PONSODA
Universidad Politécnica de Madrid, Madrid

RESUMEN

Uno de los problemas de la representación de conocimiento en terminología es la variación terminológica, ya que los conceptos se pueden lexicalizar mediante unidades terminológicas diferentes. En esta contribución, tras analizar la tipología de las variantes terminológicas propuestas por diferentes autores, nos centramos en cómo se pueden representar las variantes terminológicas con relación a un modelo conceptual. Este enfoque permite atender por un lado a las variantes que apuntan al mismo concepto y se consideran sinónimas, por otro, a las que reflejan una "distancia semántica" pero se refieren al mismo concepto, y finalmente, a las variantes que están relacionadas mediante un enlace conceptual. Estos casos se ejemplifican mediante lemon*, un modelo de lexicón para ontologías.*

Palabras clave: variación terminológica, ontologia, modelo *lemon*, sinonimia, relaciones conceptuales y semánticas

ABSTRACT

When representing knowledge in terminology, one of the problems encountered is terminological variation, as concepts can be lexicalized by means of different terminological units. In this contribution, after analyzing the typology of variants proposed by different researchers, we focus on how term variation can be represented with respect to a conceptual model. This approach allows us to account for those variants that are considered synonyms and point to the same concept, those that reflect a "semantic distance", but still refer to the same concept, and those related by means of a

conceptual link. We will exemplify this by relying on lemon, a lexicon model for ontologies.

1. INTRODUCTION

When representing knowledge in terminology, one of the problems encountered is terminological variation, as concepts can be lexicalized by means of different terminological units. Despite the efforts of the traditional theory of terminology, (Wüster, 1979) whose aim was to achieve univocity in specialized subject communication, corpus-based studies have shown that term variation is present in many specialized domains. Thus, term variation has attracted the attention of scholars (Cabré, 2008; Daille et al., 1996; Daille, 2005; Freixa, 2002) in an attempt to delimit the problem as well as to identify the causes of variability in terms and the dynamics of term evolution (Kageura, 2002). Moreover, term variation is a relevant issue in many terminology-related fields, such as term detection (Cabré et al 2001), term extraction (Vivaldi, 2004), ontology-driven translation (Budin 2003), machine learning (Mooney, 2004), and ontology building, to mention just a few.

In this paper, after analyzing the typology of variants proposed by different researchers, we focus on how term variation can be represented with respect to a conceptual model. As a result of this new perspective, we propose a slightly modified classification of term variants. Then, we concentrate on those variants that are considered synonyms, those that reflect a "semantic distance" but refer to the same concept, and those related by means of a conceptual link. Since our purpose is to analyze terminology variants with respect to a given ontology that captures a certain conceptualization of a domain agreed by a community of experts, the ontology will already constrain or restrict the set of term variants that can be related to it.

For this aim, we rely on an external linguistic model that has been developed to linguistically enrich ontologies, and that can capture terminological variation, among other linguistic properties (lexical variation, morphological decomposition, decomposition of phrase structures, syntactic frames, multilingualism, etc.). This model is

called *lemon* (McCrae et al., 2011), and provides the necessary mechanisms to represent term variants. After introducing the model, we will provide some examples to illustrate how the different types of terminology variation can be accounted for in such a model.

2. DEFINING TERM VARIANTS

A term variant has been defined as "an utterance which is semantically and conceptually related to an original term" (Daille et al., 1996). The same author expands this definition by explaining what is meant by "utterance", "original term", and "semantically and conceptually related terms" (Daille, 2005). An utterance is an attested form encountered in a text. It is considered to be a variant with respect to an authorised term, i.e., a term listed in an authoritative terminological resource and accepted by a community of experts. This term variant can be related to the original term in three forms: 1) by a synonymy relation, 2) by reflecting a "semantic distance from the reference term", or 3) by referring to "another term linked to the authorised term by a conceptual link".

Interestingly enough, the same author claims that the typology of term variants proposed in the studies she analyses are dependent on the final application for which they have been identified (information retrieval, machine-aided text indexing, scientific and technological watch, or controlled terminology for computer-assisted translation systems). For example, in the case of machine-aided text indexing, where the goal is to provide access to relevant documents from a query, the author maintains that semantic variants (*benign mouse skin tumours* and *benign neoplasms*) are particularly relevant because they widen and enhance searches. But, what is understood here by *semantic variants*? In the specific study for text indexing by Jacquemin (2001) mentioned in Daille (2005), semantic links are provided by thesauri such as AGROVOC[1] or the WordNet lexicon (Fellbaum 1998). The type of semantic relations captured in these resources is quite different. Whereas thesauri make use of fuzzy "conceptual" relations such as *broader*, *narrower* or *related to*, the WordNet lexicon includes lexical relations like *hyponymy-hyperonymy*, *synonymy*, *antonymy* or *meronymy*. This work points to the fact that, for certain purposes, the

definition of semantic variants is quite wide and includes terms that are rather linked by conceptual relations.

Daille also identifies a core of term variation types that appear in all studies:

a) Inflectional (*conservations de produit* and *conservations de produits*)
b) Shallow syntactic (*fixation azote* and *fixation d'azote*)
c) Morphosyntactic (*hand function* and *function of the hand*)
d) Paradigmatic (*un procédé alimentaire* and *procédé de conservation alimentaire; spotting telescope* and *spotting scope*)

It could be claimed that the first three core variant types refer to the same concept, whereas the last type accounts for *semantically or conceptually related terms*. Again, although the borderline between semantically or conceptually related terms is not so clear according to the examples provided in Daille (2005), we believe that making this difference may be of relevance for our purposes.

In this regard, by analyzing the causes that produce this variation, we may identify the type of variant at hand. Freixa (2006) refers to five potential causes of term variation:

- dialectal, caused by different origins of the authors
- functional, caused by different communicative registers
- discursive, caused by different stylistic and expressive needs of the authors
- interlinguistic, caused by contact between languages
- cognitive, caused by different conceptualizations and motivations.

We are particularly interested in those cognitive causes of term variation, since the rest of the causes will most probably derive in term variants related by synonymy, i.e., pointing to the same concept. But, what if term variants are originated because of different conceptualizations? As Freixa (2006: 65) puts it:

> *In terminology it has not always been accepted that the knowledge of reality is diverse; this diversity is explained by the different structures, experiences and objectives through which an individual or group approaches the comprehension of reality. A different segmentation and structuring of reality leads, in the process of knowledge, to different categories,*

since the activity of categorization is not unique. These different categorisations can lead to different mental representations of these categories and therefore to different conceptualizations.

These differences in conceptualizations caused by the perspective taken when observing a certain reality, the motivation behind it, or the way in which denomination makes explicit a selection of semantic features of a concept over others, may be the reason for those term variants that cannot be considered synonyms, but *partial synonyms* (Cabré, 2008). This phenomenon has also been termed *multidimensionality* of terms (Bowker, 1997; Rogers, 2004). As explained in Fernández-Silva et al., (2011):

(…) multidimensionality occurs when a concept can be seen from more than one perspective and can therefore be classified and designated in more than one way based on the different characteristics that it possesses.

Thus, the question here is to find out if this *partial synonymy* between or among term variants can be accounted for by referring to one and the same concept, or if each of the highlighted senses or semantic features point to different concepts, though having many features in common.

In this regard, and basing our proposal on Cabré's classification (2008), we define three types of term variants:

1. Term variants that are *semantically coincident but formally different*, also referred to as synonyms or terminological units that totally correspond to the same concept;
2. Term variants that are *semantically and formally different*, but still refer to the same ontological concept. In this case, each variant may highlight one facet of the same concept or serve a certain purpose (style, register, level of specialization), but this is not conceptually relevant or it is not explicitly manifested in the conceptualization taken as reference; and,
3. Term variants that are *semantically and formally different* and point to two related, but also different, ontological concepts, which means that they are also *conceptually different*.

We argue that the distinction between 2 and 3 will depend on the ontological model we take as reference, and the granularity of the

conceptual distinctions made there, as we will further explain in section 2. It could also be discussed, if we still want to consider term variants, those terms that are semantically, formally and conceptually different, but this will be out of the scope of this paper.

3. EXAMPLES OF TERM VARIANTS

Based on previous classifications of terminology variation already commented in section 1, we have identified three main groups of term variants that include the following types:

Group 1. Synomyms or terminological units that totally correspond to the same concept:

- graphical and orthographical variants (*localization* and *localisation*);
- inflectional variants (*cat* and *cats*);
- morphosyntactic variants (*nitrogen fixation* and *fixation of nitrogen*).

Group 2. Partial synonyms or terminological units that highlight different aspects of the same concept:

- stylistic or connotative variants (*man* and *bloke*)
- diachronic variants (*tuberculosis* and *phthisis*)
- dialectal variants (*gasoline* vs. *petrol*)
- pragmatic or register variants (*headache* and *cephalalgia*; *swine flu* and *pig flu* and *H1N1* and *Mexic pandemic flu*)
- explicative variants (*immigration law* and *law for regulating and controlling immigration*)

Group 3. Terminological units that highlight different features of the same concept and that belong to different conceptualizations, or variants that refer to two conceptually related concepts. According to the final purpose of the application, conceptual relations can be restricted to "subtype of" relations (*hyponymy-hyperonymy*), or can include other types of conceptual relations (*meronymy*).

- variants with conceptual consequences (*neoplasm* and *tumour*; *residuos hospitalarios* and *residuos biosanitarios*)

The relation between *neoplasm* and *tumour* is a "subtype of" relation, whereas *residuos hospitalarios* and *residuos biosanitarios* would most probably appear in different conceptualizations of the same domain produced by users coming from different backgrounds or with different purposes.

4. REPRESENTING TERM VARIANTS IN LEMON

In the context of our research, we are able to capture the three sorts of terminological variants in a complex model of lexical descriptions that is to be published with ontologies, namely, the *lemon* model (McCrae, 2011).
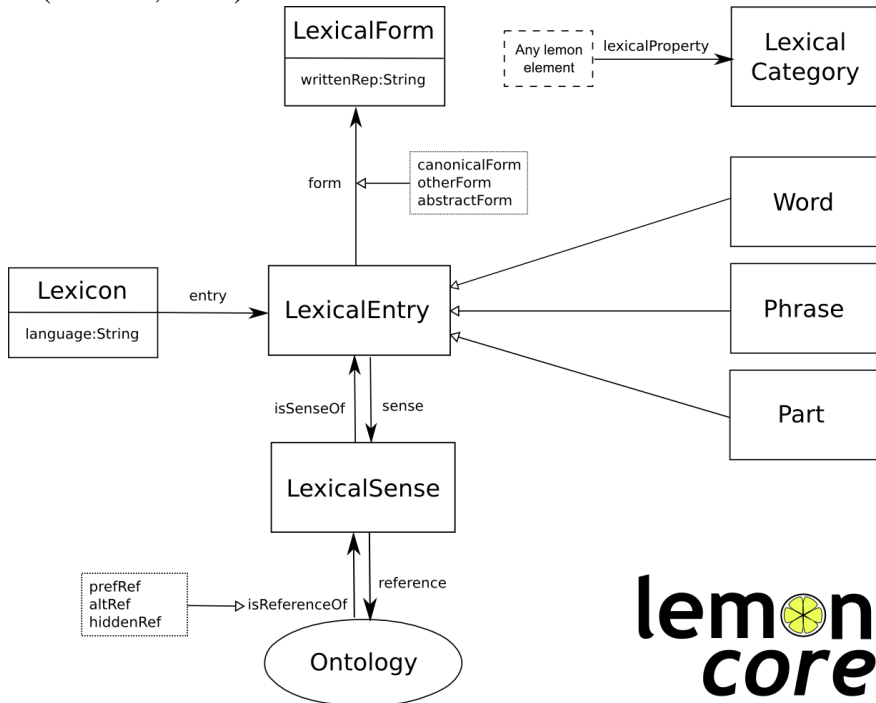
Figure 1. Core classes of the *lemon* model

In *lemon*, concepts are represented by the ontology, and terms are associated with concepts by means of a principled link represented by the class *LexicalSense*. It is this intermediate class that allows us to capture those semantic properties of term variants that make them

semantically distinct. The core classes of the *lemon* model are the ones that make up the main path between the ontology and the lexical entry, its forms and written representations, as can be seen in Figure 1. Since concepts as defined in ontologies, and lexical entries as defined in lexicons cannot be said to overlap (Hirst, 2004), the *LexicalSense* class provides the adequate restrictions (usage, context, register, etc.) that make a certain lexical entry appropriate for naming a certain concept in the specific context of the ontology being lexicalized. This class will be a key factor in making a distinction between those term variants included in Group 1 and the ones included in Group 2. Essentially, the main difference is that those term variants considered semantically coincident but formally different will be pointing to the same lexical sense, whereas those considered semantically and formally different will be linked to different lexical senses, which, in its turn, are pointing to the same ontology element. Finally, the term variants that make up Group 3 will be pointing to different lexical senses, and also to different ontology concepts. Let us illustrate this with some examples.
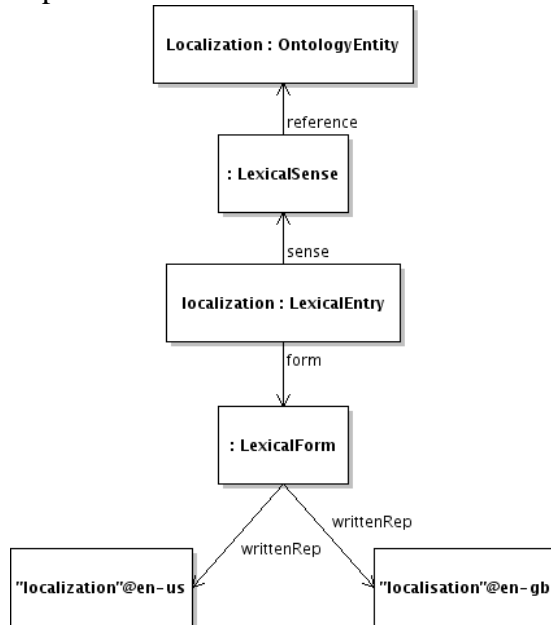
Figure 2. Example of ortographical variants

In Figure 2, we have included an example of the so-called graphical or orthographical variants. They are represented as two different written representations of the same lexical form, associated to the same lexical entry and pointing to the same lexical sense and ontology concept.
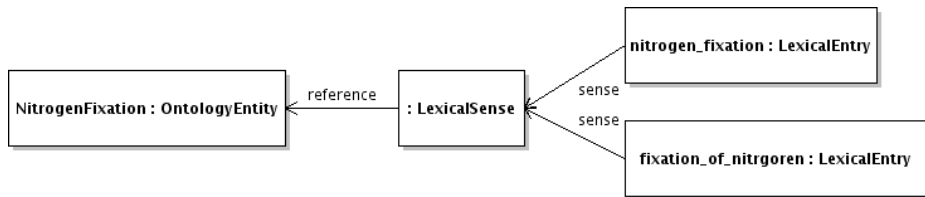

Figure 3. Example of morphosyntactic variants

Figure 3 represents two different lexical entries (*nitrogen fixation* and *fixation of nitrogen*) that are associated to the same lexical sense, as their differences in format do not have any meaning or pragmatic consequences, but further represent the same meaning in the context of the ontology. In Figure 4, we aim to illustrate one example of term variants which are semantically and formally different, in that they are used in different geographical settings. With the aim of capturing that restriction, we associate each lexical entry to a different lexical sense, and account for that usage restriction.
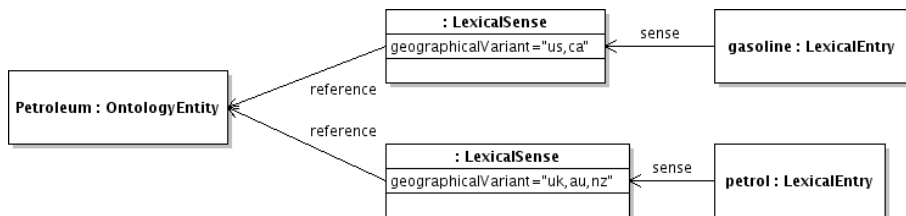

Figure 4. Example of dialectal variants

Finally, in Figure 5, we represent the two lexical entries *neoplasm* and *tumour* linked to two different lexical senses, which, in its turn, point to two concepts in the ontology related by means of the "subclassOf" relation.
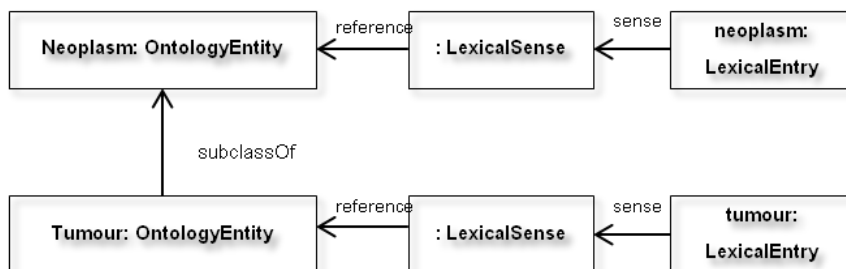
Figure 5. Example of variants with conceptual consequences

4. CONCLUSIONS

In this paper, we discuss how to represent term variants in a given ontology, by means of a lexicon ontology model. After analyzing the typology of term variants suggested by different authors we propose a threefold classification of term variation. We first focus on those variants that refer to the same concept and are seen as synonyms. Then, we deal with those that reflect a semantic distance because they stress different aspects, but denote the same concept, and finally, on those related by means of a conceptual link, be they a subtype of or part of relation. We claim that *lemon*, a model to represent and share lexical information in ontologies, can account for all these problems, as can be seen in the examples provided in this paper.

ACKNOWLEDGMENTS

NOTES

[1] http://aims.fao.org/standards/agrovoc

REFERENCES

Bowker, L. 1997. You say "flatbed colour scanner", I say "colour flatbed scanner": A descriptive study of the influence of multidimensionality on term formation and use with special reference to the subject field of optical scanning technology. *Terminology* 4(2):275-302.

Budin, G. 2005. Ontology-driven translation management, in *Knowledge systems and translation*, Dam, H.V., Engberg, J. Gerzymisch H. (eds.) Mouton de Gruyter.

Cabré, M. T. 2008. El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología (I). *IBÉRICA* 16:9-36.

Daille, B. 2005. Variations and application-oriented terminology engineering. *Terminology* 11(1):181-197.

Daille, B. Habert, B. Jacquemin, C. and Royauté, J. 1996. Empirical observation of term variations and principles for their description. *Terminology* 3(2):197-257.

Fellbaum Ch. (ed.) 1998. *An Electronic Lexical Database*. Cambridge, Mass. The MIT Press.

Fernández-Silva, S. Freixa, J. Cabré, M.T. 2011. "A proposed method for analysing the dynamics of cognition through term variation". *Terminology* 17(1): 49-73.

Freixa, J. 2006. "Causes of denominative variation in terminology. A typology proposal", *Terminology* 12(1): 51-77.

Freixa, J. 2002. *La variació terminològica: anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient.* PhD Thesis. Universitat Pompeu Fabra: Barcelona.

Hirst, G. 2004. "Ontology and the Lexicon". In S. Staab and R. Studer (Eds.), *Handbook on Ontologies*, 209-230.

Jacquemin, C. 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, MA: MIT Press

Kageura, K. 2002. *The dynamics of terminology*, John Benjamins.

McCrae, J. , G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner. Interchanging Lexical Resources on the Semantic Web. In *Journal of Language Resources and Evaluation*, 2011 (en prensa).

Mooney, 2004, Machine learning, in *The Oxford Handbook of Computational Linguistics*, Mitkov, R. (ed.), 376-394.

Rogers, M. 2004. Multidimensionality in concepts systems: A bilingual textual perspective. *Terminology* 10(2): 215-240.

Vivaldi, J. 2004. *Extracción de Candidatos a Término Mediante Combinación de Estrategias Heterogéneas*, Barcelona: IULA, Sèrie Tesis 9.

Wüster, E. (1979). *Introduction to the General Theory of Terminology and Terminological Lexicography*. Vienna: Springer.