Representing term variation in lemon

Elena Montiel-Ponsoda
Guadalupe Aguado-de-Cea
Ontology Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
{emontiel, lupe}@fi.upm.es

John McCrae
Semantic Computing Group
CITEC
Universität Bielefeld
jmccrae@cit-ec.uni-bielefeld.de

Abstract

In this contribution our objective is to define term variation, analyze the state of the art, and propose a new classification of term variants according to our representation purposes in lemon, a lexicon-ontology model to enrich ontologies with linguistic descriptions.

1 Introduction

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 certain community. And it 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 a conceptual link.

According to Daille (2005), the adopted definition of term variation depends on the purpose of the final application. For instance, in information retrieval the term variants usually handled are morpho-syntactic variants (histamine of the wine vs. wine histamine1) or variants related by a conceptual link (printer vs. laser printer).

In this contribution we concentrate on those variants that are considered synonyms and on those that reflect a "semantic distance" but that refer to the same concept. In doing so, we will not be dealing with those terminology variants related by means of a conceptual link. The reason for this is that we aim to analyze terminology variants with respect to an ontology or conceptual model, and we argue that conceptual relations will be already available in the knowledge model. However, we also foresee some mechanisms for the case that conceptually related variants are to be represented outside the ontology.

We understand synonym related variants as those term variants that are semantically coincident but formally different, as defined in Cabré (2008). With regard to variants that reflect a semantic distance, we include those variants that are semantically and formally different (Freixa, 2002; Cabré, 2008) but still refer to the same ontological concept.

In section 2 we propose a classification based on state of the art works and provide examples of each type of term variant. Then, in section 3 we describe how we aim at representing terminology variation in lemon, an ontology-lexicon model proposed in the framework of the Monnet project in order to linguistically enrich ontologies with lexical, terminological and syntactic information.

2 Typologies of variants revisited

Based on previous classifications of terminology variation (Freixa, 2002; Daille, 2005; Cabré, 2008) we identify two main groups of term variants: 1) term variants that are semantically coincident but formally different, and 2) term variants that are semantically and formally dif-

---

1Some examples have been extracted from Daille (2005) and Cabré (2008)
Different. This has representation consequences as will be shown in section 3.

Group 1) would include,
- **graphical and orthographical variants** (localization vs. localisation);
- **inflectional variants** (cat vs. cats);
- **morpho-syntactic variants** (nitrogen fixation vs. fixation of nitrogen).

Regarding group 2), here we are dealing with terms that correspond to one and the same concept, but whose usage reflects a different aspect of the concept or a different intention on the side of the user, thus the semantic and formal distinction. This shows the pragmatic aspects necessary to be considered in scientific communication. It means that the use of one term or the other is conditioned by a certain cognitive intention and highlights certain dimensions or features of the concept that will make its use more appropriate in certain situations. This phenomenon has been termed multidimensionality (Broker, 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". In Cabré (2008) these term variants are also referred to as partial synonyms.

According to these definitions, we consider that the following term variants belong to this group:
- **stylistic or connotative variants** (man vs. bloke)
- **dialectal variants** (gasoline vs. petrol)
- **pragmatic or register variants** (headache vs. cephalalgia)
- **diachronic variants** (tuberculosis vs. phthisis)
- **domain or concept dimension variants** (swine flu vs. pig flu vs. H1N1 vs. Mexican pandemic flu; MRSA (as Methicillin-resistance Staphylococcus aureus) vs. HA-MRSA vs. CA-MRSA)
- and what we dubb **explicative variants** (immigration law vs. law for regulating and controlling immigration).

It could also be argued that the term variants belonging to group 2) refer to different concepts, or, at least, to concepts belonging to different ontologies or to ontologies in the same domain created with different purposes. However, we claim that since they are pointing to the same concept or object in the world, they can be represented as term variants for that concept. In the context of our research we are able to capture these terminological variants in a complex model of lexical descriptions that is to be published with domain ontologies, namely, the lemon model (McCrae, 2011).

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 and formally distinct. In the next section, we aim at illustrating the representation of term variants in lemon.

1 **Terminology variation in lemon**

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, 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 key 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 terms variants considered se-
manically coincident but formally different will be pointing to the same LexicalSense, 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. Let us illustrate this with some examples.

In Figure 2, we have included an example of the so-called graphical or orthographical variants. There we see that they are represented as two different written representations of the same LexicalForm, associated to the same LexicalEntry and pointing to the same LexicalSense and ontology concept. As these differences are only due to orthographical rules and not reflected in the spoken language, we consider them to be the same form of the entry.

In Figure 3 we represent two different lexical entries (nitrogen fixation and fixation of nitrogen) that are associated to the same LexicalSense, 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.

Finally, 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 LexicalEntry to a different LexicalSense, and account for that usage restriction.

Acknowledgments

This work is supported by the EU project Monnet (FP7-248458), the Spanish project BabelData (TIN2010-17550), and the CITEC excellence initiative funded by the EU and the Deutsche Forschungsgemeinschaft.

References


