Multilingual Variation in the context of Linked Data

Elena Montiel-Ponsoda
Ontology Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
elena.montiel@upm.es

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

Guadalupe Aguado-de-Cea
Ontology Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
lupe@fi.upm.es

Jorge Gracia
Ontology Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
jgracia@fi.upm.es

Abstract

In this paper we present a revisited classification of term variation in the light of the Linked Data initiative. Linked Data refers to a set of best practices for publishing and connecting structured data on the Web with the idea of transforming it into a global graph. One of the crucial steps of this initiative is the linking step, in which datasets in one or more languages need to be linked or connected with one another. We claim that the linking process would be facilitated if datasets are enriched with lexical and terminological information. Being that the final aim, we propose a classification of lexical, terminological and semantic variants that will become part of a model of linguistic descriptions that is currently being proposed within the framework of the W3C Ontology-Lexica Community Group to enrich ontologies and Linked Data vocabularies. Examples of modeling solutions of the different types of variants are also provided.

1 Introduction

In the same way that hyperlinks enable the creation of connections between documents, current semantic web technologies enable the establishment of connections or *links* between or among pieces of data, information, and knowledge, in what is known as the ***Linked Data*** paradigm[1], with the goal of better exploiting them in linked data-driven Web applications (Hausenblas, 2009).

In the context of this new paradigm, we believe that terminology has much to contribute to this field. In the past, terminology work was extensively applied to the identification of terms and relations for their subsequent transformation into concepts and conceptual relations in ontologies (Velardi et al., 2001; Aussenac-Guilles and Sörgel, 2005; Maynard et al., 2008; to mention just a few). Currently, works on terminological variation may play a significant role in the Linked Data linking step.

Linked Data refers to a set of best practices for publishing and connecting distributed data on the Web with the idea of transforming it into a *global graph*. For this purpose, data must be previously structured according to graph-based models in the form of ontologies, using the standard RDF (Resource Description Framework) syntax. Moreover, these data or information units have to use URIs (Uniform Resource Identifiers) as their names on the Web, and follow the HTTP (Hypertext Transfer Protocol) schema so that users can look up those names and find the information related to them. Finally, data have to be connected to similar data, so that users can explore those data and discover additional data.

---

[1] http://linkeddata.org/

Thus, the more links an RDF dataset has to other datasets, the more useful it will be.

The linking step is the key one, but also the one that involves greater difficulties. As stated in (Heath and Bizer, 2011), it is common practice to use the property or relation `owl:sameAs` to state that one data source in an "RDF dataset A" provides the same information as another data source in an "RDF dataset B". But, is it easy to identify two data sources in different RDF datasets that mean the same? Can it be done automatically, or does it require an expert to analyze and compare the datasets? Ideally, taking into account the number of RDF datasets currently published as Linked Data[2], this task should be performed automatically.

Moreover, although the initiative of transforming data into the linked data format was initially led by English speaking countries, nowadays we find an increasing amount of RDF datasets in languages other than English that need to be linked to similar or related datasets in other languages (Gómez-Pérez et al., 2013).

The analysis of terminological variation, a cornerstone in communicative and cognitive approaches to terminology (Cabré 1995, Daille 2005, Temmerman, 2000) could contribute to the identification of terms that refer to the same ontological concept, thus attempting to integrate univocity, defended by the traditional theory (Wüster, 1979) with variation in real situations. The result of such an analysis could be used in the automatic identification of concepts that *mean the same* or that hold a certain type of relation. It could also contribute to the definition of the reasons that caused that variation, and propose alternatives to the `owl:sameAs` property to capture more fine-grained relations between data sources. Finally, from a multilingual perspective, it could also help to establish cross-lingual relations between RDF datasets in various languages.

For all these reasons, we believe that Linked Data datasets should be enriched with terminological variants, as well as with other types of lexical and linguistic information as proposed in (McCrae et al., 2011; Gracia et al., 2012), so that further processes in the Linked Open Data Cloud construction – specifically the linking step – become smoother and more reliable. Such enrichment could also be very profitable due to its potential exploitation by linked data-driven Web applications.

In this contribution we provide a classification of lexical, terminological and semantic variants that has been proposed within the framework of the W3C Ontology-Lexica Community Group[3] to enhance a model of linguistic descriptions intended to enrich domain ontologies and RDF datasets. The model being designed in this framework relies on previous computational models of linguistic description, such as LMF (Francopoulo, 2013; ISO 24613), SKOS (Miles et al., 2005), or, fundamentally, the *lemon* model (McCrae et al., 2011).

Basing on works that analyse the causes of denominative variation in communicative approaches to terminology, in section 2 we revisit previous classifications of terminological variants in the light of the Linked Data paradigm. In other to justify the proposed classification, in section 3 we provide examples of modelling solutions for the different types of variants (lexical, terminological and semantic variants). We compare the mechanisms provided by available models (SKOS) to represent such variants, in contrast to the richer, more complex model of linguistic descriptions that is being proposed in the W3C Ontology-Lexica Community Group and that takes as starting point the *lemon* model. Finally, in section 4, we present some concluding remarks and discuss some further lines of research.

2   Revisited classification of term variation

As suggested in Cabré (2008), term variants that refer to one and the same concept can be divided into two types: (1) Term variants that are *semantically coincident but formally different*, i.e., terms that mean the same but are expressed by different lexical forms, generally known as synonyms (e.g., eczema vs. skin rash); and (2) Term variants that are *semantically and formally different*, since each one is highlighting one facet or dimension of the same concept (e.g., hospital waste vs. biomedical waste), so that they *do not mean exactly the same*, but refer to the same concept or real world entity. The same author refers to the latter variants as *partial synonyms* and leaves open the question of whether the two terms should point to the same concept or each

20

---

term should point to a different concept, with many assumed commonalities between the two.

This discussion becomes highly significant in view of a model that is designed to associate complex linguistic descriptions to conceptual structures (ontologies, RDF datasets), because it informs how lexical and terminological descriptions of the concepts are represented. If the conceptual structure is already given and contains that conceptual difference (let us say that it makes a distinction between *biosanitary waste*, in general, and *hospital waste*, only for the waste produced in hospitals), the two terms will most probably be associated to two different concepts. Conversely, if only one concept is represented in the ontology, we may still want to account for both terminological variants in the linguistic model, and explicitly state the motivation behind each denomination. In this way, we would also facilitate the linking of this data source to another data source contained in a different dataset and to which only the term *biosanitary waste* has been associated.

The classification we propose is motivated by the causes that provoked the variation, and has been inspired for the terminological part on the work by Freixa on denominative variation in terminology (2006). In this case, we distinguish between lexical, terminological and semantic or cognitive variants. Each type of variant will be devoted a sub-section below.

2.1 Lexical variants

For the purposes of this work, lexical variants are defined as those variants that are *semantically coincident but formally different*, and which are mainly motivated by grammatical requirements, style (*Wortklang*), and linguistic economy (helping to avoid excessive denominative repetition and improving textual coherence)[4]. As Freixa (2006: 61) maintains for acronyms and reductions of terms, this lexical variation has a high level of conceptual equivalence. Also, the use of one variant over the other does not really change the intention of the message, but it is rather caused by formal aspects of the text.

The following types can be mentioned:

- Orthographic variants

21

[4] This type of variants have been thoroughly analysed by Jacquemin (1997) mainly for French.

- o Diatopic variants (e.g., localize vs. localise)
- o Diachronic variants (e.g., different scripts for languages such as Azeri)
- o Ideographic variants (e.g., in Japanese both "寿司" and "鮨" are used for sushi)

- Affixal variants

- o Derivational variants (e.g., adjective -> adverb variation: quick vs. quickly; verb -> noun: activate vs. activation)
- o Inflexional variants (e.g., adjective agreement: *rojo, roja, rojos, rojas*)

- Morphosyntactic variants

- o Compounds (e.g., ecological tourism vs. eco-tourism)
- o Abbreviations (including acronyms, among others. E.g., peer to peer and p2p; WYSWYG, FAO, UNO, etc.)
- o Rephrasing variants (e.g., immigration law vs. law for regulating and controlling immigration)

2.2 Terminological variants

As for terminological variants, we understand those variants that are *not only formally, but also semantically different*, and this difference is intentionally caused. As stated in Diki-Kidiri (2000:29 and ff), in order to better understand this type of variants, it may be useful to make a distinction between concept and meaning or sense (*le signifiant, le signifié* and *le concept*), since we could say that these terminological variants refer to the same concept but they represent "the multiple specific perceptions of the same object".

In this type of terminological variants, the denomination or term itself is a clear indicator of the reasons or causes for variation. As mentioned in Freixa (2006), these reasons can be the origins of the authors, in the case of *diatopic* variants; the different communicative registers, in the case of *diaphasic* variants (also termed functional

variants); the stylistic or expressive needs of the authors, as for the so-called *diastratic* variants; and the different conceptualizations, approaches or perspectives underlying them, in what we have termed *dimensional* variants (dubbed cognitive variants in Freixa (2006)).

Regarding the latter ones, we would like to emphasize that it is more common than not to find different conceptualizations of the same domain when different groups approach the same area of knowledge from different perspectives or with different needs. Because of that, some terms may highlight certain properties of a concept, which are not so relevant for other users. This is even more obvious in a multilingual context, in which different geographical, cultural and social groups comprehend reality in different ways. In this sense, we have included a subtype of dimensional variants called *cross-lingual dimensional variants*.

Finally, we would like to refer to the cross-lingual variants. It could be argued that these are not terminological variants strictly speaking, but translations. However, we have decided to consider them a subtype of variants with the aim of covering those scenarios in the Linked Data context in which datasets in different natural languages have to be linked and this linkage becomes essential in this new paradigm.

We have identified two types of cross-lingual variants. First, we include translations, in the general sense. It is widely accepted that original and target cultures have segmented and categorized a bit of reality in a very similar way and have a similar concept and equivalent term to refer to it. We do not account for the reasons for this similarity (it may be that one culture has imported not only the concept but also the term by providing a loan translation, etc.). However, we account for the case in which the target culture has no equivalent concept and describes the concept of the original culture and/or directly reuses the foreign term.

Finally, it is worth mentioning that we could also have *cross-lingual diaphasic variants*, if one language uses the scientific term in all registers, and the other language has two terms: one for an expert-to-expert communication situation and another term for an expert-lay communication situation (e.g., huesos metacarpianos in Spanish vs. *Ossa metacarpi* and *Mittelhandknochen* in German). The same could happen in the case of diachronic and diastratic variants.

In any case, it is important to consider that in a multilingual scenario these terminological variants would be pointing to the same concept or conceptual structure, or even share the same conceptualization. This is one of the main differences compared to the variants in section 2.3, namely, the so-called semantic or cognitive variants.

So, we have classified terminological variants as follows:

- Diatopic (dialectal or geographical variants) (e.g., gasoline vs. petrol)

- Diaphasic (register) (e.g., headache and cephalalgia; swine flu and pig flu and H1N1 and Mexican pandemic flu)

- Diachronic (or chronological variants) (e.g., tuberculosis and phthisis)

- Diastratic (discursive or stylistic variants) (e.g., man vs. bloke)

- Dimensional variants: the terms point to the same concept but highlight a different property or dimension of the concept (e.g., biosanitary waste vs. hospital waste; Novel Coronavirus vs. Middle East Respiratory Syndrome Coronavirus; obsolete technology vs. dangerous technology[5])
    - o Cross-lingual dimensional variants: the concept exists in both cultures, but the terms highlight different aspects of the concept or approach it from different perspectives (e.g., *madre de alquiler* (lit. rental mother) in Spanish vs. *mere porteuse* (carrier mother) in French vs. surrogate mother in English.

- Cross-lingual variants
    - o Translations (e.g., the translation *nogomet* instead of the loanword *fudbal* for soccer in Serbo-Croatian)

---

[5] From Freixa (2006: 68)

o Descriptions or glosses (when the concept does not exist in the target language and a literal translation or gloss is used) (e.g., École normal and French Normal School, Panetone vs. Panetone, Italian Christmas cake)

2.3 Semantic or Cognitive Variants

Semantic or cognitive variants are mainly caused by different conceptualization and/or motivations. We could say that these term variants are *semantically and formally* different, as in the case of terminological variants, but they usually point to two closely related, but different, ontological concepts, which means that they are also *conceptually different* (Aguado-de-Cea and Montiel-Ponsoda, 2012).

Such variants are commonly found at a multilingual level, but we can also found them in monolingual contexts. Let us imagine the case of an ontology or dataset that contains the concept *religious building*, and another ontology that contains the concept of *mosque*. At the linguistic level we could say that religious building and mosque are in a relation of hypernymy-hyponymy (one concept is subsumed by the other, but they are referring to two different concepts included in two different conceptual structures that have a different granularity level).

Here we also distinguish between vertical (general-specific) or horizontal variants. Vertical variants are defined as those variants that refer to concepts that share most properties, but one is more specific than the other (they are not at the same level in a classification tree, but one is more general and the other more specific. See the example of river in English vs. *rivière* and *fleuve* in French). In the case of horizontal variants, we refer to those terms that point to concepts that share most properties, but one includes properties that the other does not, and vice versa. As a result of these divergences, terms will be pointing to two different concepts in two different conceptual structures at the same level of specificity in a classification tree, but including unequal properties. Therefore, we can consider them counterparts or closest equivalents.

Within this group we find the following types: 23

- Vertical (general-specific) variants (e.g., benign neoplasms vs. benign mouse skin tumours)
  - o Cross-lingual vertical variants (e.g., river in English vs. *rivière* and *fleuve* in French; *testamento* in Spanish vs. testament and last will in English)

- Horizontal variants (counterparts or closest equivalents):
  - o Cross-lingual horizontal variants (e.g., Prime Minister in English vs. *Presidente del Gobierno* in Spanish)

3 Modelling examples

Incorporating all this terminological knowledge in ontologies is important if we aim at optimizing the linking process. Thus, distinguishing different forms of term variation turns into a key issue, when we model terminology for practical applications. In the context of linked data this means that we will model the data by means of an existing model such as SKOS or *lemon*. In this section we present practical modelling examples for kind of term variation.

3.1 Lexical variants

Lexical variants are modelled by either the multiple forms of the same entry or by means of relationships between lexical entries. For example in *lemon*, we would model *orthographic variants* as different representations of the same form[6]:

```
:myExampleLexicon a lemon:Lexicon ;
    lemon:language "en" ;
    lemon:entry :theatre_lexicalentry .
example_ontology:Theatre
    lemon:isReferenceOf [
      lemon:isSenseOf :theatre_lexicalentry].

:theatre_lexicalentry a lemon:LexicalEntry ;
    lemon:canonicalForm [
        lemon:writtenRep "theater"@en-us ;
        lemon:writtenRep "theatre"@en-gb ] .
```

In the example above, a monolingual lemon lexicon is defined which contains a lexical entry *:theatre_lexicalentry*. The concept "Theatre" in a

---

[6] We are using turtle RDF notation in our examples.

certain ontology is the reference of such lexical entry, which has two associated written representations: "theater" and "theatre" for American and British English respectively.

Alternatively, two different lexical entries could have been defined for each different representation. In that case, a relation can be defined between the lexical entries in *lemon* in this way:

```
:theatre_lexicalentry a lemon:LexicalEntry ;
  lemon:form [
    lemon:writtenRep "theatre"@en-gb ] .

:theater_lexicalentry a lemon:LexicalEntry ;
  lemon:form [
    lemon:writtenRep "theater"@en-us ] .

:theater_lexicalentry :diatopicVariant
  :theatre_lexicalentry .

:diatopicVariant rdfs:subPropertyOf
  lemon:lexicalVariant  .
```

In this example "theater" and "theatre*"* are associated to two different lexical entries. They are linked by a new relation *:diatopicVariant*, which is defined as a subtype of a *lemon* lexical variant.

Notice that *:diatopicVariant* does not exist in the *lemon* model as such, but it can be defined as in the example or, alternatively, an external category could be used, such as *"diatopical"* included in ISOCAT[7].

Morphosyntactic variants are also represented as links between lexical entries, as there may be differences in the syntactic properties of the entries. Let us consider the term "peer to peer" and its abbreviated form "p2p":

```
:p2p a lemon:LexicalEntry ;
  lemon:form [ lemon:writtenRep "P2P"@en ] .

:peer_to_peer a lemon:LexicalEntry ;
  lemon:form [
    lemon:writtenRep "Peer-to-peer"@en ] .

isocat:fullFormFor rdfs:subPropertyOf
  lemon:lexicalVariant .
isocat:initialismFor rdfs:subPropertyOf
  lemon:lexicalVariant .

:peer_to_peer isocat:fullFormFor :p2p .
:p2p isocat:initialismFor :peer_to_peer .
```

In the previous example *:p2p* and *:peer_to_peer* are lexical entries with their respective written representations. Then, ISOCAT categories are defined as lexical variants and used to relate or link both lexical entries[8].

We can also use SKOS for representing lexical variants. In that case, we can show two preferred labels for different dialects of a language as follows:

```
:theatre skos:prefLabel "theater"@en-us  ,
                        "theatre"@en-gb.
```

However, to represent morphosyntactic variants it is necessary to use the extended label model (SKOS-XL) as follows, but, otherwise, it is similar to *lemon,* where we define a named label entity for each label and represent the link between them as a triple.

```
example_ontology:P2P
  skosxl:prefLabel :p2p_label ,
  skosxl:altLabel :peer_to_peer_label .

:p2p_label skosxl:literalForm "p2p"@en .

:peer_to_peer_label skosxl:literalForm
  "peer to peer"@en .

:P2P_label lexinfo:abbreviationFor
  :peer_to_peer_label .
```

3.2   Terminological variants

As terminological variants maintain the meaning of the term while changing the surface form, it is necessary to distinguish between the syntactic and semantic level of the term. For this reason, we characterize terminological variants as links between different senses not between lexical entries. In *lemon* this is easy to model as can be seen in the following example:

```
biontology-icd:011
  lemon:prefRef :tuberc_sense ,
  lemon:altRef :phthisis_sense .

:tuberculosis_sense lemon:isSenseOf
  :tuberculosis_lexicalentry .
:phtisis_sense lemon:isSenseOf
  :phthisis_lexicalentry .

:tuberc_sense lexinfo:dating lexinfo:modern .
:phthisis_sense lexinfo:dating lexinfo:old .
```

---

[7] http://www.isocat.org/rest/dc/3669

[8] For readability, we have substituted the original identifiers of the ISOCAT categories by descriptive ones. The originals are: http://www.isocat.org/datcat/DC-321 and http://www.isocat.org/datcat/DC-329

In the example above, an entity of *biontology-icd*[9] has been associated with two *lemon* senses. Such senses constitute the bridge between the ontology entities and their respective lexical entries. Then, the temporal dimension ("modern" vs. "old") can be established as an attribute of the senses (by using the LexInfo vocabulary (Cimiano *et al*, 2011) in our example).

Alternatively, a relation at the level of senses can be established in *lemon* in this way:

```
:diachronicVariant rdfs:subpropertyOf
  lemon:senseRelation .
:phthisis_sense  :diachronicVariant
  :tuberculosis_sense .
```

However, there is no way to make this distinction in SKOS and this will lead inevitably to confusion about the syntactic and semantic layer.

3.3   Semantic or cognitive variants

Cognitive variants are distinct but closely related meanings of a word. So, we can model the variation not only as a relationship between words but also as described by a semantic model, i.e., an ontology. As such, we would state OWL axioms to describe the relationship, as illustrated in the example below. There, Chancellor of Germany and Prime Minister of Spain are both subclasses of the concept Head of government.

```
dbpedia:Chancellor_of_Germany rdfs:subClassOf
  dbpedia:Head_of_government .
dbpedia:Prime_Minister_of_Spain
  rdfs:subClassOf dbpedia:Head_of_government.
```

At the lexicon level we could also establish a relation of horizontal variants between the two terms. This relation is established because we know that the two terms "Chancellor of Germany" and "Prime Minister of Spain" are not equal but can be considered similar (or counterparts) in the two cultural settings, as they have a close antecedent concept, "head of government".

4   Conclusions

As mentioned in this paper, the Linked Data initiative needs to find ways of linking the huge amount of structured datasets found on the Web in the same or in different languages. We believe

that although ontologies aim at achieving univocity in as much as traditional terminology did, the more sociolinguistic cognitive approaches to terminology can also contribute to enrich the current computational models of linguistic descriptions. With this purpose, we have revisited previous classifications of term variants in the light of the Linked Data initiative so as to facilitate the process of recognition of terminological variation. We have proposed a classification of term variants in three wide groups: lexical variants, terminological variants and semantic or cognitive variants. We have also illustrated this classification with the corresponding examples at the ontology level by resorting to different ontology representation models, such as *lemon*. With the solutions proposed we also aim to enrich the linguistic ontology models as well as to make them more reliable when applied to Linked Data.

References

Aussenac-Gilles, N. and D. Soergel. 2005. Text analysis for ontology and terminology engineering. In Journal of Applied Ontology (1), 35 – 46.

Cabré, M.T. 1995. On diversity and terminology. Terminology 2(1), 1-16.

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.

Cimiano, P., P. Buitelaar, J. McCrae, and M. Sintek. 2011. LexInfo: A declarative model for the lexicon-ontology interface. Web Semantics: Science, Services and Agents on the World Wide Web, 9(1): 29-51.

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

Diki-Kidiri, M. 2000. Une aproche culturelle de la terminologie. Terminologies nouvelles 21 (Terminologie et diversité culturelle): 27-31.

Francopoulo, G. (ed.) 2013. LMF Lexical Markup Framework, Wiley- ISTE.

Gómez-Pérez, A. Vila-Suero, D. Montiel-Ponsoda, E. Gracia, J. and G. Aguado-de-Cea. 2013. Guidelines for multilingual linked data. In Proceedings of the 3rd International Conference on Web Intelligence,

---

[9] http://purl.bioontology.org/ontology/ICD9CM/

Mining and Semantics (WIMS'13), ACM, New York.

Gracia, J. Montiel-Ponsoda, E. Cimiano, P. Gómez-Pérez, A. Buitelaar, P. and J. McCrae. 2011. Challenges for the multilingual Web of Data. In Web Semantics: Science, Services and Agents on the World Wide Web 11: 63-71.

Heath, E. and Ch. Bizer. 2011. Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, (1)1: 1-136. Morgan & Claypool.

ISO 24613 - LMF- Lexical Markup Framework, Language Resource Management. 2006. Available from
http://lirics.loria.fr/doc_pub/LMF%20rev9%2015
March2006.pdf

Jacquemin, C. 1997. Variation terminologique: reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Memoire d'habilitation a diriger des recherches en informatique fondamentale, Universite de Nantes.

Maynard, D., Y. Li, and W. Peters. 2008. NLP Techniques for Term Extraction and Ontology Population. In Buitelaar, P. and P. Cimiano, (eds.), Ontology Learning and Population: Bridging the Gap between Text and Knowledge, IOS Press, Amsterdam: 171-199.

McCrae, J., G. Aguado de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez- Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. 2012. Interchanging Lexical Resources in the Semantic Web. Language Resources and Evaluation 46, (4): 701-719.

Miles, A., B. Matthews, D. Beckett, D. Brickley, M. Wilson, and N. Rogers. 2005. SKOS: A language to describe simple knowledge structures for the web. In Proceedings of the XTech Conference.

Temmerman, R. 2000. Towards new ways of terminology description. The Sociocognitive Approach. John Benjamins, Amsterdam.

Velardi, P. Missikoff, M. and R. Basili. 2001. Identification of relevant terms to support the construction of Domain Ontologies. In Proceedings of ACL-01 Workshop on Human Language Technologies, Toulouse, France, 2001.

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