Variation and Translation Modules

By Elena Montiel-Ponsoda, Jorge Gracia, Guadalupe Aguado de Cea

1. Introduction

In order to propose a **Translation Module** in *lemon*, we previously analyzed (or revisited) the ways in which to represent variants, which would belong to the **Variation Module** (which Philipp named Terminology Module in his e-mail), for which we also have some proposals.

According to the classification provided in the Ontolex wiki for the "Relations and Properties of Lexical Entries", we had identified three types of "relations":

(see: http://www.w3.org/community/ontolex/wiki/Specification_of_Requirements/Properties-and-Relations-of-Entries):

1. Relations between LEXICAL variants
2. Relations between TERMINOLOGICAL variants
3. Relations between SEMANTIC variants

In the following document we present a twofold proposal:

- how to represent variants in *lemon*
- how to represent translations in *lemon*

The representation proposals we provide in this document are the ones we would propose as default in a sort of lemon-ontolex cookbook. However, there would be alternative ways of modelling it according, some of which were already captured in http://www.w3.org/community/ontolex/wiki/Specification_of_Requirements/Properties-and-Relations-of-Entries

2. VARIATION MODULE

2.1 LEXICAL variants

Lexical variants: lexical variants were 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)

a) Orthographic variants

    1. Diatopic variants (e.g., localize vs. localise)

2. Diachronic variants (e.g., different scripts for languages such as Azeri)

3. Ideographic variants (e.g., in Japanese both “寿司” and “鮨” are used for sushi)

b) Affixal variants

1. Derivational variants (e.g., adjective -> adverb variation: quick vs. quickly)

2. Inflexional variants (e.g., adjective agreement: rojo, roja, rojos, rojas)

c) Morphosyntactic variants

1. Compounds (e.g., ecological tourism vs. eco-tourism)

2. Abbreviations (including ac-ronyms, among others. E.g., peer to peer and p2p; WYSWYG, FAO, UNO, etc.)

3. Rephrasing variants (e.g., immigration law vs. law for regulating and controlling immigration)

For this type of variants we propose the following representation: 1 LexicalEntry with 1…n LexicalForms (and 1 LexicalSense pointing to the ontology element), as show in the code below.

```
:fao a lemon:LexicalEntry ;
    lemon-ontolex:sense [ lemon-ontolex:reference dbpedia:fao ] ;
    lemon-ontolex:lexicalForm :fao_form .
:fao_form lemon-ontolex:writtenRep "FAO"@en  .

:food_agriculture_organization  a lemon:LexicalEntry ;
    lemon-ontolex:sense [ lemon-ontolex:reference dbpedia:fao ] ;
    lemon-ontolex:lexicalForm :food_agriculture_organization_form .
:food_agriculture_organization_form lemon-ontolex:writtenRep "Food and
Agriculture Organisation"@en  .
isocat:fullFormFor rdfs:subPropertyOf lemon-ontolex:lexicalVariant .
isocat:initialismFor rdfs:subPropertyOf lemon-ontolex:lexicalVariant .

:food_agriculture_organization_form isocat:fullFormFor :fao_form  .
:fao_form  isocat:initialismFor :food_agriculture_organization_form .
```

The properties of each LexicalForm would be expressed as properties (gender, number, @en-us vs. @en-gb, abbreviation vs. full form, etc.)

The other option would be to represent them as several LexicalEntries, with one canonical form each, and also with one LexicalSense per each LexicalEntry. We think, however, that this second more complex option would not be necessary as the formal differences between lexical variants do not require to be justified, or accounted for, at the LexicalSense level.

2.2 TERMINOLOGICAL variants

Terminological variants: variants that are not only formally, but also semantically different, and this difference is intentionally caused. In this type of terminological variants, the denomination or term itself is a clear indicator of the reasons or causes for variation. 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.

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

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

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

d) Diastratic (discursive or stylistic vari-ants) (e.g., man vs. bloke)

e) 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; madre de alquiler (rental mother) vs. vientre de alquiler (womb mother), in Spanish).

For this type of variants we propose the following representation: 1 LexicalEntry per variant with 1…n LexicalForms, and one LexicalSense per each LexicalEntry.

This representation permits to account for the differences at the LexicalSense level, by means of LexicalSense properties, which, in our opinion, would need to be revisited or re-defined to cover the type of terminological variants identified (register, style, dimension, dialect, time…).

2.3 SEMANTIC variants

Semantic variants: variants that 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.

a) Vertical (general-specific; broader-narrower) variants (e.g., benign neoplasms vs. benign mouse skin tumors)

b) Horizontal variants (counterparts or closest equivalents): (e.g., matrimonio in Peninsular Spanish vs. marriage in Argentinean Spanish)

We would propose to represent this type of variants in the same way as Terminological variants, namely: 1 LexicalEntry per variant (with 1…n LexicalForms), and one LexicalSense per each LexicalEntry. The only different with the previous variants would be that LexicalSenses would "most probably" be pointing to two different ontology terms.

This representation permits to account for the differences at the LexicalSense level, by means of LexicalSense properties. For this purpose, several subtypes of the senseRelation were already defined in the lemon model (see p. 20 of the lemon cookbook), namely, equivalent, incompatible, broader and narrower, which could, if needed, be redefined.

3. TRANSLATION MODULE

The translation module we propose in this document can be understood as a "small extension" or "complementary or additional module" of the Variants Module explained in the section above.

This module creates the class Translation as a reification of the relation between two LexicalSenses. This reification allows us to provide some properties of the Translation class. The properties are:

- TranslationSource
- TranslationTarget
- TranslationConfidence
- Context (?)
- TranslationCategory

The Range of TranslationSource and TranslationTarget is unrestricted (it could point to a SKOSXL:label), but they will tipically point to Lexical Senses. The Domain will be the class Translation.

The TranslationConfidence class would include a numeric value for the degree of confidence.

Context is still to be defined.

The last property, TranslationCategory, would be pointing to an external registry of translation categories or types, for example: culturalEquivalence, translation, closeEquivalence. The exhaustive list is still to be determined and properly defined.

See attached files for the Translation Module and the Translation Categories.