

Enabling Language Resources to Expose Translations as Linked Data on the Web

Jorge Gracia, Elena Montiel-Ponsoda,
Daniel Vila-Suero, Guadalupe Aguado-de-Cea

Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Campus de Montegancedo sn Boadilla del Monte 28660 Madrid (Spain)
{jgracia,emontiel,dvila,lupe}@fi.upm.es

Abstract

Language resources, such as multilingual lexica and multilingual electronic dictionaries, contain collections of lexical entries in several languages. Having access to the corresponding explicit or implicit translation relations between such entries might be of great interest for many NLP-based applications. By using Semantic Web-based techniques, translations can be available on the Web to be consumed by other (semantic enabled) resources in a direct manner, not relying on application-specific formats. To that end, in this paper we propose a model for representing translations as linked data, as an extension of the *lemon* model. Our translation module represents some core information associated to term translations and does not commit to specific views or translation theories. As a proof of concept, we have extracted the translations of the terms contained in Terminesp, a multilingual terminological database, and represented them as linked data. We have made them accessible on the Web both for humans (via a Web interface) and software agents (with a SPARQL endpoint).

Keywords: linked data, translations, multilingualism

1. Introduction

Many Language Resources (LRs) such as terminology databases, electronic dictionaries, lexica, etc., are essential in NLP applications. Motivated by the need of maximizing their visibility and reusability, nowadays we are witnessing a growing trend towards making such resources available on the Web. LRs on the Web range from systems available for downloading in isolated websites to LRs that are accessible on the Web by using Semantic Web-based languages and techniques (RDF, ontologies, linked data, etc.). The latter option allows making the data in those resources usable by others in a direct manner, without relying on application-specific formats. This enables also linking the data to other resources in order to expand them with additional linguistic information. It is to that end that the *lemon* model (McCrae et al. 2012) has been proposed. Specifically, such a model is meant for creating lexica and machine readable dictionaries in multiple natural languages as linked data, usually for describing (or accompanying) an ontology.

In this paper, we focus on those LRs whose data are accessible on the Web (either as linked data or, at least, as dereferenceable URIs) in two or more natural languages (i.e., multilingual), and which can be potentially used as a source for translations between different languages. We propose a straightforward extension of the *lemon* model to represent such translations and to put them as linked data on the Web. Semantic search engines can then index such information and enable mechanisms for querying them in a simple manner. Many applications could take advantage of this, such as Machine Translation, cross-lingual instance matching, cross-lingual querying, etc. Having translations as explicit links between ontology lexica

could also contribute to the discovery of links between the ontology entities they describe (Sváb-Zamazal & Svátek, 2008).

In this paper we briefly introduce the *lemon* model in Section 2. In Section 3, we discuss the proposed mechanism to represent translations as linked data. In Section 4 we show a validating example in which a lexical resource available online has been extended with *lemon* to make its translations explicit as linked data. This is followed by some related work in Section 5 and the conclusions and future work in Section 6.

2. Background

The LEXicon Model for ONtologies (*lemon*) is an RDF model of linguistic descriptions. It has been designed to extend the lexical layer of ontologies with as much linguistic information as needed, and to provide it as linked data on the Web.

One of the main features of the model is the independence between the linguistic descriptions and the ontological model they accompany. This means that the lexical elements are defined by pointing to the corresponding semantic objects in the ontology. The model consists of a core set of classes and several modules capturing different types of lexical and terminological descriptions.¹ Linguistic annotations (data categories or linguistic descriptors) are not captured in the model, but have to be specified for each lexicon by dereferencing their URIs as defined in some repositories that contain them (for instance, the ISocat repository²).

¹ An overview of *lemon* and its core classes can be found at <http://www.lemon-model.net/>

² <http://www.isocat.org/>

The bridge between an ontological entity and its lexical descriptions is the `lemon:LexicalSense` class. This is thought to provide the adequate restrictions (usage, context, register, etc.) that make a certain lexical entry appropriate for naming a certain ontology entity in the specific context of the ontology being lexicalized.

3. Representing Translations on the Web as Linked Data

The *lemon* core model supports the representation of multilingual information, in the sense that several lexicon models in different natural languages can be associated to the same ontology or conceptual model. In that case, translation relations could be inferred between terms in different languages when they refer to the same ontology entity. However, additional mechanisms are needed to represent explicit translation relations between or among the terms in different languages, whether they belong to lexicon models that point to the same or different ontologies.

The translation module we propose grounds on a previous attempt for extending *lemon* to support translations (Montiel-Ponsoda et al. 2011). Our current approach, though, has some remarkable differences, being the main one that translation categories have been clearly separated from the model, as we will describe later in this section. In that way, our translation module does not commit to specific views or translation theories and adheres to the design principle in *lemon* of “being descriptive not prescriptive”.

We have decided to provide an explicit representation for translation relations in the *lemon* model, independent/separated from other types of variants that are modelled in what is known as the “variation module” in *lemon*, namely, terminological variants, lexical variants and sense variants, as defined in (Montiel-Ponsoda et al. 2013). The main difference between translation relations and the rest of variants in this module is that the former ones hold between or among terms in different languages (inter-lingual variants), whereas the latter ones typically hold between or among terms in the same language (intra-lingual variants). This does not mean, however, that two terms in different languages cannot be related by a translation relation and, additionally, by a terminological variation relation.

Let us provide an example of such a case. Imagine we want to define the relation between “surrogate mother” (EN) and “mère porteuse” (FR) belonging to two *lemon* lexicons pointing to the same or two different ontologies. Obviously, we will be able to establish a translation relation between them of the type specified below in this section, that is, “direct equivalent”, but also a terminological relation of the type “dimensional variant” to indicate that each language (and culture) approaches the same concept from a different perspective. In fact by combining the two types of variation relations (i.e., translation and terminological variation) we can account for two types of differences, namely, the differences with respect to the existence of semantically equivalent entities

in the two languages and cultures in question, and the pragmatic differences that exist at a term level.

The translation module we propose³ consists essentially of two OWL classes: `Translation` and `TranslationSet`. `Translation` is a reification of the relation between two *lemon* lexical senses that correspond to terms in two languages. The idea of using a reified class instead of a property allows us to describe some attributes of the `Translation` object itself. `Translation` will be the domain of the following OWL properties:

- `translationSource` and `translationTarget`. Their range is unrestricted. In fact, although we encourage to use `lemon:LexicalSense` as source and target of a `Translation`, it could point to any other type of resource that represent senses, or even to the lexical entries or labels themselves (e.g., `lemon:LexicalEntry`, or `skosxl:Label`).
- `translationConfidence`, to assign a confidence value to the translation pair, as in many multilingual resources for translators such as IATE.⁴
- `context`, which is an unrestricted property intended to express/determine, if needed, the specific application context in which a pair of lexical senses are translation of each other.
- `translationCategory`, that points to an external registry of translation categories or types.

We have proposed a set of categories for linguistic translations⁵ (RDF resources with dereferenceable URIs, as they do in ISOcat⁶) to be used in combination with our proposed translation module (by means of the `translationCategory` property):

- `directEquivalent`: Typically, the two terms describe semantically equivalent entities that refer to entities that exist in both cultures and languages. E.g., “surrogate mother” (EN) → “mère porteuse” (FR).
- `culturalEquivalent`: Typically, the two terms describe entities that are not semantically but pragmatically equivalent, since they describe similar situations in different cultures and languages. E.g., “Ecole Normal” (FR) → “Teachers college” (EN).
- `lexicalEquivalent`: It is said of those terms in different languages that usually point to the same entity, but one of them verbalizes the original term by using target language words. E.g., “Ecole Normal” (FR) → “Normal School” (EN).

Although this classification may not be exhaustive for some authors, it is intended to be used in combination with the ontologies the model enriches. In this sense, our objective was not to capture the translation strategy or technique behind the translations (i.e., loan translation, description, metonymy, etc.). Of course, other typologies of translations can be proposed instead to complement or substitute the above one, depending of the necessities of

³ It can be found at <http://purl.org/net/translation>

⁴ <http://iate.europa.eu/>

⁵ Available at <http://purl.org/net/translation-categories>

⁶ For example <https://catalog.clarin.eu/isocat/rest/dc/1297.rdf>

```

@prefix tr: <http://purl.org/net/translation#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix lemon: <http://www.lemon-model.net/lemon#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix trcat: <http://purl.org/net/translation-categories#> .
@prefix terminesp: <http://linguistic.linkeddata.es/data/terminesp/>

terminesp:lexiconES a lemon:Lexicon ;
  lemon:language "es" ;
  lemon:entry terminesp:38756es .

terminesp:lexiconEN a lemon:Lexicon ;
  lemon:language "en" ;
  lemon:entry terminesp:38756en .

terminesp:38756 a skos:Concept ;
  rdfs:seeAlso < http://www.wikilengua.org/index.php/Terminesp:red> .

terminesp:38756es a lemon:LexicalEntry ;
  lemon:sense terminesp:38756es-sense ;
  lemon:form [ lemon:writtenRep "red"@es ] .

terminesp:38756es-sense a lemon:LexicalSense ;
  lemon:reference terminesp:38756 .

terminesp:38756en a lemon:LexicalEntry ;
  lemon:sense terminesp:38756en-sense ;
  lemon:form [ lemon:writtenRep "network"@en ] .

terminesp:38756en-sense a lemon:LexicalSense ;
  lemon:reference terminesp:38756 .

terminesp:38756es-en-TR a tr:Translation ;
  tr:translationSource terminesp:38756es-sense ;
  tr:translationTarget terminesp:38756en-sense ;
  tr:translationCategory trcat:directEquivalent .

terminesp:es-en-transet a tr:TranslationSet ;
  tr:translation terminesp:38756es-en-TR .

```

Figure 1: Example of a translation in RDF turtle syntax

the model's consumer. In that case, correspondent RDF definitions and dereferenceable URIs should be provided.

Finally, there is a class `TranslationSet` that is responsible for grouping a set of translations sharing certain properties. For instance, a translation set could group translations coming from the same language resource, or belonging to the same organisation, etc. We understand this class, which groups translations, as analogous to the class `lemon:Lexicon` that groups lexical entries.

Individuals of both the `Translation` and `TranslationSet` classes can be described using the DCMI Metadata Terms vocabulary⁷ to attach valuable information about provenance, authoring, versioning, or licensing.

4. Terminesp: a validating example

Terminesp⁸ is a terminological database in Spanish created by AETER (Asociación Española de Terminología) by extracting the terminological data from the UNE⁹ documents produced by AENOR (Asociación Española de Normalización y Certificación). It contains the terms and definitions used in the UNE Spanish

technological norms (standards) and amounts to more than thirty thousand terms with equivalences in other languages whenever they are available. These norms, similar to the ISO standards, have been elaborated by Spanish committees composed of experts in different fields.

In order to validate our approach we established an automatic mechanism to extract the lexical entries from Terminesp databases and instantiate a *lemon* lexicon for each available language. In order to have an ontological counterpart for each term in Terminesp, we have instantiated a `skos:Concept` for each of them and related them to their corresponding `lemon:LexicalEntry` by means of an instance of the `lemon:LexicalSense` class. Then, translations between the corresponding lexical senses have been mapped. Finally, such lexicons and translations have been published on the Web¹⁰, following Linked Data best practices (i.e., dereferenceable URIs, HTTP content-negotiation, etc.), and appropriate mechanisms set up to query the extracted information. Additionally, we have made available a simple web user interface¹¹ and a SPARQL endpoint.¹²

⁷ <http://dublincore.org/documents/dcmi-terms/>

⁸ <http://www.wikilengua.org/index.php/Wikilengua:Terminesp>

⁹ UNE stands for Una Norma Española

¹⁰ E.g., <http://linguistic.linkeddata.es/data/terminesp/lexiconES>

¹¹ <http://linguistic.linkeddata.es/terminesp/search>

¹² <http://linguistic.linkeddata.es/terminesp/sparql-editor>

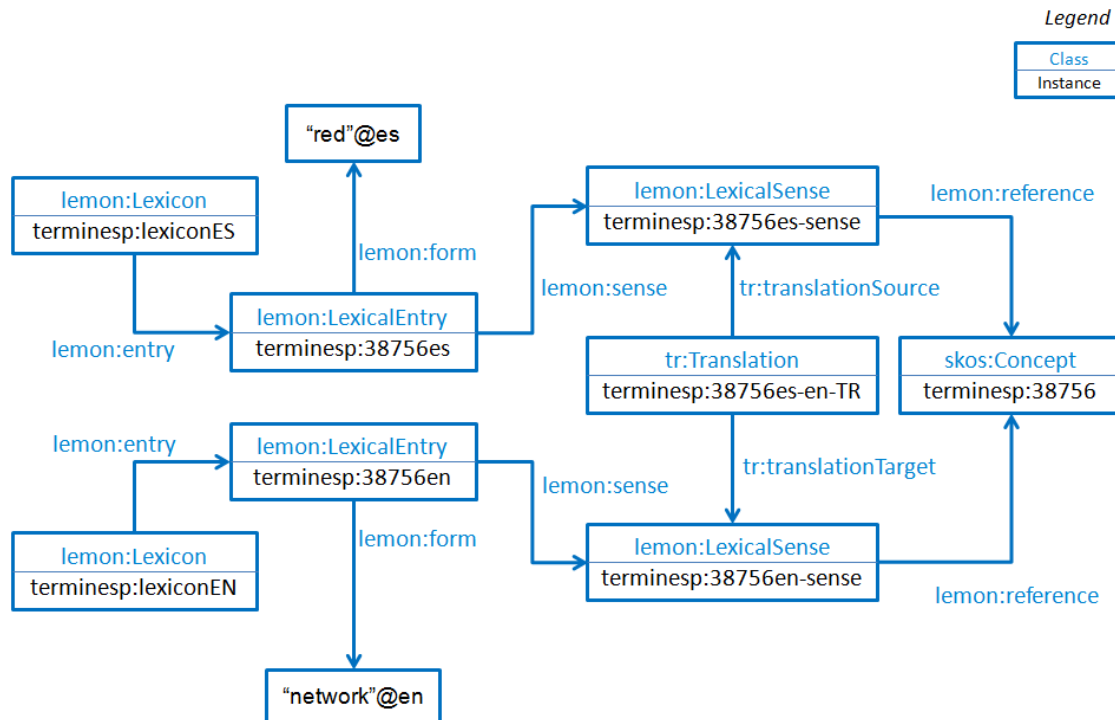


Figure 2: Example of a translation, graphically.

In Figure 1 we illustrate the *lemon*-based representation of the translation between “red” in Spanish and “network” in English, both referring to the same *Terminesp* concept.¹³: “*In networks topology, set of ideal elements and connections of a circuit, considered as a whole*”. Figure 2 depicts the example graphically.

5. Related Work

Several formats and annotation properties have been developed in the Semantic Web to represent natural language descriptions associated to ontologies, such as the *rdfs:label* (Manola & Miller, 2004) or *skos:prefLabel* (Miles & Bechhofer, 2009) properties. These properties enable a simple form of multilingual labelling by using language (e.g., *skos:prefLabel* “network”@en). However, the main limitation of this approach is that no explicit links can be created between or among multilingual labels.

The SKOS-XL (Miles & Bechhofer, 2009b) extension overcomes this problem by introducing a *skosxl:Label* class that allows labels to be treated as first-order RDF resources, and a *skosxl:labelRelation* property that provides links between the instances of *skosxl:Label* classes. In this case, the *skosxl:labelRelation* could be specialised as a translation relation, but no other assertions or statements can be made of this relation (besides OWL annotations).

As it is always the case, depending on the needs of our final application, we will opt for one of these simple representations of translations between ontology entities,

or we will have to rely on more principled ways to do it, as proposed in this contribution.

6. Conclusions and Future Work

In this paper we have proposed a model for representing translations as linked data that allows extracting translations from multilingual language resources and make them available as linked data on the Web. We validated our idea with a real example based on *Terminesp*, a multilingual terminological database.

In the future, more translation resources have to be moved into this representation scheme. We devise a future scenario in which a critical mass of *lemon* lexicons and translations among them are available on the Web. In that context, software agents could assist not only applications but also human translators by selecting translations according to the context given by an ontology and fulfilling certain properties (e.g., it is a preferred translation, it corresponds to a cultural equivalence, etc.).

Currently, a module for representing variations/translations is under discussion in the W3C Ontology-Lexica community group¹⁴. The work presented in this paper served as initial input to that discussion.

Acknowledgements. We are very thankful to AETER and AENOR for making *Terminesp* data available. We also thank Javier Bezos, from FUNDEU, for his assistance with the data. Some ideas contained in this paper were inspired after fruitful discussions with other members of the W3C Ontology-Lexica community group. This work is supported by the FP7 European

¹³ <http://www.wikilengua.org/index.php/Terminesp:red>

¹⁴ <http://www.w3.org/community/ontolex/>

project LIDER (610782), the Spanish national project BabelData (TIN2010-17550) and the Spanish Ministry of Economy and Competitiveness within the Juan de la Cierva program.

References

- F. Manola and E. Miller (2004). RDF primer. W3C Recommendation, Tech. Rep., Available: <http://www.w3.org/TR/rdf-primer/>
- J. McCrae, 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 on the semantic web. *Language Resources and Evaluation*, vol. 46.
- A. Miles and S. Bechhofer (2009). SKOS simple knowledge organization system reference. W3C Recommendation, Available: <http://www.w3.org/TR/skos-reference>
- A. Miles and S. Bechhofer (2009b). SKOS simple knowledge organization system extension for labels (SKOS-XL). W3C Recommendation, Available: <http://www.w3.org/TR/skos-reference/skos-xl.html>
- E. Montiel-Ponsoda, J. Gracia, G. Aguado-de Cea, and A. Gómez-Pérez (2011). Representing translations on the semantic web. In *Proc. of 2nd Workshop on the Multilingual Semantic Web*, at ISWC'11, Bonn, Germany, ISSN 1613-0073, vol. 775. CEUR-WS, pp. 25-37.
- E. Montiel-Ponsoda, J. McCrae, G. Aguado-De-Cea, and J. Gracia (2013). Multilingual variation in the context of linked data. In *Proc. of 10th International Conference on Terminology and Artificial Intelligence (TIA'13)*, Paris (France).
- O. Sváb-Zamazal and V. Svátek (2008). Analysing ontological structures through name pattern tracking. In *Proc. of EKAW 2008*, Acitrezza, Italy. *Lecture Notes in Computer Science*, vol. 5268. Springer, pp. 213-228.