Towards the Integration of Multilingual Terminologies: an Example of a Linked Data Prototype

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

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

Abstract

Many language resources are nowadays available in machine readable formats, but still contained in isolated silos. Current Semantic Web-based techniques enable the transformation and linking of those resources to become a navigable graph of linked language resources, which can be directly consumed by third-party applications. The prototype we have developed builds on a web user interface and SPARQL endpoint initially developed to query a single terminological database (Terminesp), now extended to navigate a set of multilingual terminologies. The vocabulary used to represent these terminologies into the linked data format is *lemon-ontolex*, a *de facto* standard for representing lexical information relative to ontologies and for linking lexicons and machine-readable dictionaries to the Semantic Web.

1. Introduction

The Linguistic Linked Open Data (LLOD) cloud[1] is a sub-cloud of linguistic resources provided in an interoperable way (using the Resource Description Framework or RDF data model), freely accessible and linked with each other. In its current state, the LLOD Cloud contains monolingual and multilingual dictionaries, lexicons, thesauri and even corpora. English is the best represented language, and some languages are underrepresented or not present at all.

With Terminesp (a multilingual terminological database created by the Spanish Association for Terminology, AETER), we aimed at validating the *lemon-ontolex* model as a representation scheme for lexical resources, specifically, the so-called `vartrans` module, a dedicated module that accounts for terminological variation and translation relations among entries (Bosque-Gil et al., 2015). Building on that experience, we have now transformed additional multilingual terminological resources, namely, a set of freely available terminology databases[2] from the Catalan Terminological Centre, TERMCAT, into linked data (LD) using *lemon-ontolex* as underlying data format, and aim to showcase the benefits of integrating terminological resources.

In this paper, we focus on the design decisions taken in the transformation and linking steps, and on the impact they have in the search and navigation of the resulting linked terminological data.

In Section 2, we introduce *lemon-ontolex* and the `vartrans` module. In section 3, we describe the design decisions taken in the transformation process. In section 4, we refer to the benefits of browsing and navigating linked multilingual terminologies.

2. lemon-ontolex

The *lemon-ontolex* model is the resulting work of the efforts made by the W3C Ontology Lexica Community Group since 2011 to build a rich model to represent the lexicon-ontology interface. It is largely based on the *lemon* model (McCrae et al., 2012) and consists of a core set of classes and several modules[3]. The `vartrans` module has been developed to record lexico-semantic relations across entries in the same or different languages (Fig. 1.): those among senses and those among lexical entries and/or forms. Lexico-semantic relations among senses are of semantic nature and include

[1] http://linguistic-lod.org/

[2] http://www.termcat.cat/es/terminologiaoberta/

[3] See *lemon-ontolex* final model specifications at http://www.w3.org/community/ontolex/wiki/Final_Model_Specification

terminological relations (dialectal, register, chronological, discursive, and dimensional variation) and translation relations. In contrast, relations among lexical entries and/or forms concern the surface form of a term and encode morphological and orthographical variation, among other aspects.

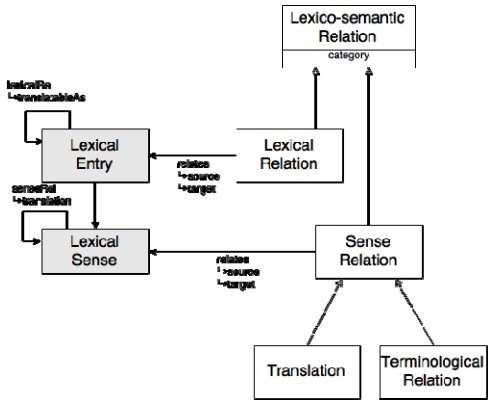

Fig. 1. Classes and properties in *vartrans*

3. Migration and linking of the resources

For the transformation of TERMCAT terminology repertoires to the LD format and linking to Terminesp we followed these steps: data exploration, URI naming strategy, data modeling, RDF generation and linking (Vila-Suero et al., 2014).

Data exploration. TERMCAT terminology repertoires are divided by domain. Each database consists of a list of entries in Catalan and their translations into Spanish, English, French, etc., along with the term type (full form or abbreviation), references to associated terms, synonyms, and, sometimes, definitions. Data for part-of-speech, gender and number in nouns, and subcategorization of verbs, is also available.

URI naming strategy. Inspired by the work in the Apertium dictionaries[4], the term itself, its part of speech and the language of the term are part of the URI of the lexical entry. For lexical senses, the domain is included in the URI.

Modeling. For the modeling process, we regard each term in a set of translations as a specific sense of a lexical entry, a sense that is mapped to a concept in a particular domain. This allows us to have a unique lexical entry *red* (network), for instance, which occurs both in the lexicon *Internet i societat de la informació* as in the lexicon *Indústria electrònica i dels materials elèctrics*, with different senses that we extract from each domain lexicon. This results in a number of RDF lexica that matches the number of languages available in TERMCAT data, and each lexical entry will have a different number of senses depending on its use across domains. In this way, the lexical entry :red-n-es will be mapped to a sense :red-n-es-Internet-sense, as well as to a :red-n-es-Industria-sense, etc. Each of these senses refers to a skos:Concept with a particular definition and domain. Regarding translations, the `vartrans` module represents them as relations across lexical senses of the entries of each lexicon. Parts of speech, subcategorization, gender and number are accounted for as well.

Generation and linking. For the transformation we used the data cleaning and transformation tool OpenRefine[5] with its extension for LD. We linked to lexinfo[6] to cover morphosyntactic information, and to Terminesp at the lexical entry level. Linking to DBpedia is also planned as a next step.

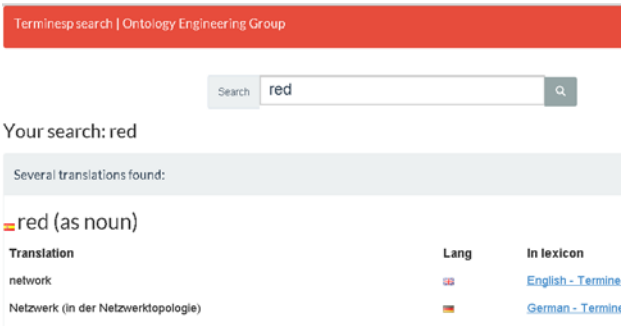

Fig. 2. Web user interface

4. Browsing multilingual terminologies

We reuse the Terminesp web user interface (see Fig. 2.) and SPARQL endpoint to browse and query this set of integrated terminologies[7]. Benefits are related to easy access and reuse of linguistic data by end users (translators, terminologists) and semantic-aware software agents.

Acknowledgements. This work is supported by the FP7 EU project LIDER (610782), and the Spanish 4V project (TIN2013-46238-C4-2-R).

References

J. Bosque-Gil et al. (2015). Applying the OntoLex Model to a Multilingual Terminological Resource. In Proc. of ESWC 2015. Springer.

J. McCrae et al. (2012). Interchanging lexical resources on the semantic web. Language Resources and Evaluation, vol. 46.

D. Vila-Suero et al. (2014). Publishing Linked Data on the Web: the Multilingual Dimension. In P. Cimiano & P. Buitelaar (Eds.) Towards the Multilingual Semantic Web. Springer.

[4] http://linguistic.linkeddata.es/apertium/

[5] http://openrefine.org/index.html

[6] http://lexinfo.net/

[7] http://linguistic.linkeddata.es/terminesp/