Applying the OntoLex Model to a Multilingual Terminological Resource

Julia Bosque-Gil, Jorge Gracia,
Guadalupe Aguado-de-Cea, and Elena Montiel-Ponsoda

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

Abstract. Terminesp is a multilingual terminological resource with terms from a range of specialized domains. Along with definitions, notes, scientific denominations and provenance information, it includes translations from Spanish into a variety of languages. A linked data resource with these features would represent a potentially relevant source of knowledge for NLP-based applications. In this contribution we show that Terminesp constitutes an appropriate validating test bench for OntoLex and its `vartrans` module, a newly developed model which evolves the *lemon* model to represent the lexicon-ontology interface. We present a first showcase of this module to account for variation across entries, while highlighting the modeling problems we encountered in this effort. Furthermore, we extend the resource with part-of-speech and syntactic information which was not explicitly declared in the original data with the aim of exploring its future use in NLP applications.

Keywords: ontolex, variation, translation, terminological resource

1 Introduction

Recent years have seen growing interest in the publication of language resources (LRs) as linked data, and the presence of machine readable dictionaries, lexicons, and thesauri in the Linguistic Linked Open Data (LLOD) cloud[1] continues to increase. Linking language resources not only enables humans and software agents easier access and querying of structured data collections, but linked multilingual LRs represent a potentially relevant source of knowledge for NLP-based applications developed in the fields of machine translation, content analytics, multilingual information extraction, word sense disambiguation or ontology localization. The inclusion of terminological knowledge from different domains into the LLOD cloud has already been explored with the creation and publication of thesauri, vocabularies and terminology repositories, especially in the environmental and geological domain [10, 1, 2], as well as in the financial [9] and linguistics [6, 3, 4] fields. Converting and publishing terminological dictionaries

---

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

as linked data opens new doors to the reuse of these resources in domain specific NLP applications and machine translation.

Aimed originally at bridging the gap between lexical and conceptual information, the *lemon* model (LExicon Model for ONtologies) [7] is now a widespread representation model for the publication of lexical resources as linked data which has been gradually expanded to include new modules under the umbrella of the W3C Ontology-Lexica Community Group[2], resulting in the newly developed model *OntoLex/lemon*.[3] An extension to *lemon* that accounts for translation relations among lexical senses from the same or different data sets was also developed [5]. In such effort, the Terminesp[4] data served as a validating example of a terminological LR to which the translation module would apply.

Our contribution builds on such previous work [5] and describes a first showcase of the `vartrans` module of OntoLex in order to account for terminological variation and translation relations among entries. In addition to showcasing the `vartrans` module, we extend the resource further by adding components to represent definitions and terminological norms of Terminesp entries. The entries themselves range from simple nouns and adjectives to complex nominal, prepositional and adjectival phrases, which led us to include part-of-speech information for each entry and turn to LexInfo [4] classes to account for that mixed nature as well. We also draw attention to the modeling problems to which these prepositional phrases give rise and which will be tackled in future work.

The structure of this paper is as follows: Section 2 briefly explains the OntoLex `vartrans` module, with special focus on the terminological variation aspects not addressed in previous work. Section 3 introduces the Terminesp database and provides an example of the structure of its data. Section 4 dwells on previous work with Terminesp and describes our approach to model the entries, including definitions and syntactic information, among other aspects. Following, we showcase the `vartrans` module to represent scientific denominations. Lastly, Section 5 discusses some conclusions and future lines of work.

2   The OntoLex `vartrans` Module

OntoLex is the resulting work of the continued efforts made by the W3C Ontology Lexica Community Group during the past three years to build a rich model to represent the lexicon-ontology interface. It is largely based on the *lemon* model [7] and, along with the extensions to it, integrates work of the various Community members.

Broadly, each entry in the lexical database belonging to an `ontolex:Lexicon` is modeled as an `ontolex:LexicalEntry` and mapped to its respective ontology entity. The mappings are established at the sense level through the property `ontolex:reference` and the class `ontolex:LexicalSense`, thereby capturing

---

the fact that a single lexical entry may have different senses, each one referring to a different ontology entity and evoking a particular lexical concept. However, information regarding the realization of a lexical entry (e.g. inflection, pronunciation, etc.) is recorded at the lexical form level via `ontolex:Form`.

The OntoLex `vartrans` module was developed to record variation relations across entries in the same or different languages. The intuition behind it is to capture two kinds of relations: those among senses and those among lexical entries and/or forms. Variation relations among senses are of semantic nature and include terminological relations (dialectal, register, chronological, discursive, and dimensional variation)[5] 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. This last kind of relations are not considered semantic in nature; grammatical meaning encoded in morphological affixes is thus represented at a different layer than lexical meaning (senses), and variation in orthography is thought as a relation between two similar forms (e.g. *analyze*, *analyse*), in contrast to synonymy and antonymy relations between two senses (e.g. *shut, close*) in which the surface forms are not involved. In this paper we only focus on the first kind of relations, variation relations among senses, to represent translations and register (also called *diaphasic*) relations between a term and its scientific denomination.

2.1 Translations

The OntoLex `vartrans` module frames translation relations as a special type of lexico-semantic variation across the entries of different lexica, more specifically, as relations that hold among senses. The translation component goes back to the *lemon* Translation module. In this view, the `vartrans` module turns to a pivot class `vartrans:Translation` to represent translations among lexical entries as relations among `ontolex:LexicalSenses` that point to the same ontology concept. One of the main differences between the OntoLex and the *lemon* translation modules is the conception of a class to encompass variation relations, the `vartrans:LexicoSemanticRelation` class. Its subclasses denote relations that hold among lexical entries (`vartrans:LexicalRelation`) and relations that hold among lexical senses (`vartrans:SenseRelation`). The pivot class `vartrans:Translation` mentioned before is thus a `vartrans:SenseRelation`, and so are the relations among terminological variants as well (`vartrans:TerminologicalVariant`). The Terminesp database contains translation relations of the *directEquivalent* category, but other translation categories (e.g. *culturalEquivalent*, for culture-dependent concepts; *lexicalEquivalent*, for literal translations of the source term, etc.) are supported as well and can be included from an external ontology [5]. In addition to translation relations among their senses, lexical entries in different languages can be directly related through the property `vartrans:translatableAs`.

---

[5] `http://www.w3.org/community/ontolex/wiki/Specification_of_Requirements/Properties-and-Relations-of-Entries`

2.2 Terminological variants in the same language

Terminological variants in the same language are modeled as relations among senses, too. The `vartrans` module allows for the encoding of dialectal (diatopic), register (diaphasic), chronological (diachronic), stylistical (diastratic) as well as dimensional variation among entries. In this way and in the same fashion as translations, two lexical entries are mapped to their respective lexical senses, and these are related through the pivot class `vartrans:TerminologicalVariant`, with the property `vartrans:category` allowing for the specific type of terminological variation at hand to be included as well. For cases in which there is not any directionality involved, that is, there is not any source or target term, the property `vartrans:relates` (similar to the former `tr:translationSense` in *lemon*) links the two senses to the element acting as pivot. In the following example, there is a diachronic variation between the terms *phthisis* and *tuberculosis*, being the latter one the one used nowadays. This shift is captured by representing the two senses as source and target respectively.

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#> .
@prefix dct: <http://purl.org/dc/terms/>.

:tuberculosis a ontolex:LexicalEntry ;
        ontolex:lexicalForm :tuberculosis_form ;
        ontolex:sense :tuberculosis_sense.

:tuberculosis_form ontolex:writtenRep "tuberculosis"@en .

:tuberculosis_sense ontolex:reference
    <http://dbpedia.org/resource/Tuberculosis>.

:phthisis a ontolex:LexicalEntry ;
        ontolex:lexicalForm :phthisis_form ;
        ontolex:sense :phthisis_sense.

:phthisis_form ontolex:writtenRep "phthisis"@en .

:phtisis_sense ontolex:reference
    <http://dbpedia.org/resource/Tuberculosis>;
                dct:subject
                    <http://dbpedia/resource/Medicine> .

:phtisis_diachronic_var a vartrans:TerminologicalVariant;
    vartrans:source :phtisis_sense ;
    vartrans:target :tuberculosis_sense ;
    vartrans:category isocat:diachronic.
```

3 The Terminesp Database

The Terminesp terminological database was created by the *Asociación Española de Terminología* (Spanish Association for Terminology, AETER) by extracting the terminological data from the UNE ('*Una Norma Española*' a Spanish norm) 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. The norms are defined over a range of domains, from aeronautics and electro-technical engineering to fruit nomenclature. An entry in Terminesp consists of the definition of the term, the norm from which the term is extracted, the norm title, and, if available, the translation of the term to one or several different languages, namely German, French, Italian, Swedish, and/or English. A Terminesp entry is presented in Table 1.[6]

Norm	UNE_21302-801
Norm title	Vocabulario electrotecnico: capítulo 801. Acústica y electroacústica
Spanish term	**admitancia cinética**
German term	Bewegungsadmittanz
English term	motional admittance
French term	admittance cinétique, admittance motionnelle
Definition	En un transductor, diferencia entre la admitancia eléctrica en carga y su admitancia eléctrica cuando está mecánicamente bloqueado.
Note	NOTA - Esta definición es válida principalmente para transductores con acoplamiento por transformador.

Table 1. Terminesp entry for the Spanish term *admitancia cinética*

4 Migrating Terminesp to Linked Data

In this contribution we renew the previous work with Terminesp [5] in order to (1) detect errors and inconsistencies in the data before linking the data set to other lexical resources (i.e. LexInfo), (2) provide a validating example of the OntoLex *vartrans* module to account for variation across entries, with emphasis on scientific naming, (3) extend the LD resource with definitions, norms, and

---

[6] En. *Norm*: Electrotechnical vocabulary: Chapter 801. Acoustics and electroacoustics. *Definition*: Referred to a transducer, the difference between the electrical admittance on charge and its electrical admittance when it is mechanically locked. *Note*: NOTE - This definition applies mainly to transducers with transformer coupling.

part-of-speech categories, and (4) create a database of nominal, prepositional, and adjectival phrases with highly specialized content (not covered by other LRs) to be used by NLP applications. Since we have built upon previous work [5] with the lexical database, we have stuck to the resource structure and URI naming strategy that the authors followed in their approach.[7] Being OntoLex still under development, the RDF files resulting from our tests with Terminesp are not published as linked data yet, but they are open and accessible online.[8]

As in [5], we instantiate a `skos:Concept` for any given Terminesp entry in order to ground the terms conceptually. A Terminesp entry, in turn, is modeled as an `ontolex:LexicalEntry` whose `ontolex:LexicalSense` points to the appropriate `skos:Concept`. Translations are included by instantiating a `vartrans:Translation` element with the properties `source` and `target` linking the two translation senses, one for each language. The senses are attached to the `vartrans:Translation` by an additional property too, `vartrans:relates`.

4.1   Definitions, notes, and norms

In addition to the available translations for a given entry that were captured in the previous *lemon* version [5], definitions, notes, norm codes, norm titles and provenance were added as linked data as well.

Specifically, definitions are attached to the `skos:Concept` the `ontolex:LexicalSense` is mapped to. This is done through the property `skos:definition`. Moreover, some of these definitions include a note that provides additional information about the definition content, use cases, etc. and, in order to distinguish this from the definition itself, we use `rdfs:comment` to relate the `skos:Concept` to the string acting as note. In *lemon*, the class `lemon:SenseDefinition` allowed to treat the definition of a term as an object whose property `lemon:value` pointed to the definition string itself. This was particularly well-suited for capturing elements that were not definitions but which were nonetheless related to them, as in the case of notes to definitions, as we have in Terminesp. The class `lemon:SenseDefinition` is not included in OntoLex, in fact, definitions are not encoded at the sense, but at the `ontolex:LexicalConcept` level with the property `ontolex:definition`. A lexical concept in OntoLex aims to reify the concept one or several senses evoke and lexicalize, resembling a *synset* in WordNet. In our view, the approach based on the `skos:definition` property appeared more suitable for the task. That is, an account of Terminesp definitions in terms of LexicalConcept definitions would imply instantiating a lexical concept for any Terminesp entry, which, along with the lexical sense, would bring unnecessary complexity to the representation.

Terminesp entries are all extracted from Spanish UNE documents and, in addition to definitions, the database provides norms and norm code information. The norm each entry comes from is thus captured by `dc:title`, and the norm code included with an instantiation of the `dc:source` property.

4.2 Part-of-speech tags and syntactic phrases

Part-of-speech (POS) tags were not provided in the original data. In order to link each entry to its corresponding syntactic category through `lexinfo:partOf-Speech` at the `ontolex:LexicalEntry` level, TreeTagger[9] was used. The initial idea was to tag the Spanish data in order to obtain part-of-speech information that holds for translations to different languages as well. In other words, given that the terms are highly specialized, a mechanical term in Spanish that is a noun is likely to have a corresponding translation in English or German that is a noun as well. However, the nature of Terminesp entries is mixed: the data set is made up of adjectives, verbs and nouns, along with complex noun phrases (NP), prepositional phrases (PP) and adjective phrases (AP). In order to represent this, we linked multi-word Terminesp entries to `lexinfo:NounPhrase`, `lexinfo:AdjectivePhrase` and `lexinfo:PrepositionPhrase` accordingly. In this way we are encoding the syntax of a prepositional phrase and also stating that it may function as an adjective or as an adverb (via `lexinfo:partOfSpeech`), for instance. Table 2 shows the distribution of the different part of speech tags, and which of them involve a complex phrase structure (NPs, PPs, APs).

lexinfo:PartOfSpeech	Synt.Phrase	Entries
noun	simple	13777
	NP	18552
verb	simple	69
adjective	simple	31
	PP	56
	AP	4
adverb	PP	2
Total		**32491**

Table 2. Distribution of part-of-speech categories and syntactic phrases in Terminesp Spanish entries. The tag *simple* is included here for contrastive purposes: it refers to those entries that do not involve a complex constituent structure.

Interestingly, the PPs are regarded here as independent entries and there is not any information pointing to their syntactic governor, even though there are complex NPs in the data that are formed by a noun and a PP that occurs as lexical entry too. Thus, we find a PP such as *en reposo* 'idle' as a Terminesp entry and other entries (NPs) with that same PP as constituent: *masa en reposo, tinta en reposo, pasador en reposo*. However, the PP and the NP entries are not related in the data. Not only does this contrast with conventional dictionaries, where the preposition is usually accessed through its NP complement (**reposo**, en –) or the whole PP is accessed through its syntactic governor (**masa**, – en reposo),

---

[9] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

but it also prevents us from using OntoLex's syntax and semantics module to encode syntactic behavior, since we cannot access the syntactic governor of the PP or the syntactic frame in which the PP would be fit as argument. Moreover, some of these PPs can also accompany a verb (*estar en reposo*, 'to be idle'; *funcionar en reposo*, 'to work idle'), so that the type of syntactic frame we are dealing with is not always inferable from the PP alone.

Most entries in Terminesp are actually NPs (see Table 2): e.g. *potencia isótropa radiada equivalente* 'equivalent isotropically radiated power'. With an initial random test set of 500 entries, TreeTagger achieved 0.997 precision and 0.995 recall. The reason behind this is the high number of nouns. In cases in which the entry was a complex phrase, TreeTagger tagged every element in it, which allowed us to identify prepositional phrases. The remaining multi-word entries were initially tagged as nouns. An analysis of the errors revealed that deverbal adjectives were used as nouns throughout the data, and that some multi-word entries included their tag (*capacitivo, adj.*, 'capacitive') or even a disambiguation note: *funcionar (para los relés elementales)*, 'function (for elemental relays)'. The tags for the scarce adjectives and verbs were checked and corrected manually, and PPs were tagged as adjectives or adverbs according to their definition and sample uses, if the latter were available.
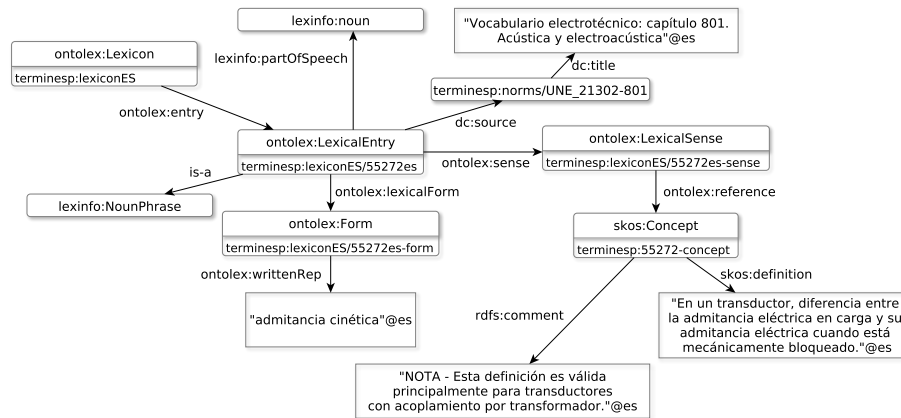

Fig. 1. A Terminesp entry modeled with OntoLex

It is worth mentioning, however, that PPs pose modeling problems still. There seem to be different degrees of lexicalization among them: some of them are fixed both in the specialized domain and in the general language (*en reposo*); others, e.g. *a circuito abierto* 'open-circuited', may admit a certain degree of variation and are not even regarded as a set phrase outside the specialized domain. Being the meaning of these entries compositional (to a certain degree), they could not be considered *idioms* according to the definition of `lexinfo:idiom`, nor are they

collocations in the sense of `olia:Collocation`,[10] and they do not correspond either to the category of prepositional constructions as, for instance, composite prepositions (*in front of*) or prepositional adverbs (*outside*), which are accounted for in linguistic terminology repositories. Furthermore, translations from the Spanish entry (a PP) into other language may be in the form of PPs as well, adjectives, or adverbs, depending on the target language. The part-of-speech that we assigned to the Spanish entry itself is subject to change given that we do not have the syntactic context in which the entry occurs: some of them are eligible for both adjectival and adverbial use. Capturing these nuances, however, was outside the scope of this paper but will be considered for future work on terminological data.

Figure 1 is included as an example of a Terminesp entry in OntoLex with information about the definition, the norm code, the norm title, the part-of-speech category and the syntactic phrase. Translations are not included in this figure, but we refer the reader to Figure 2.
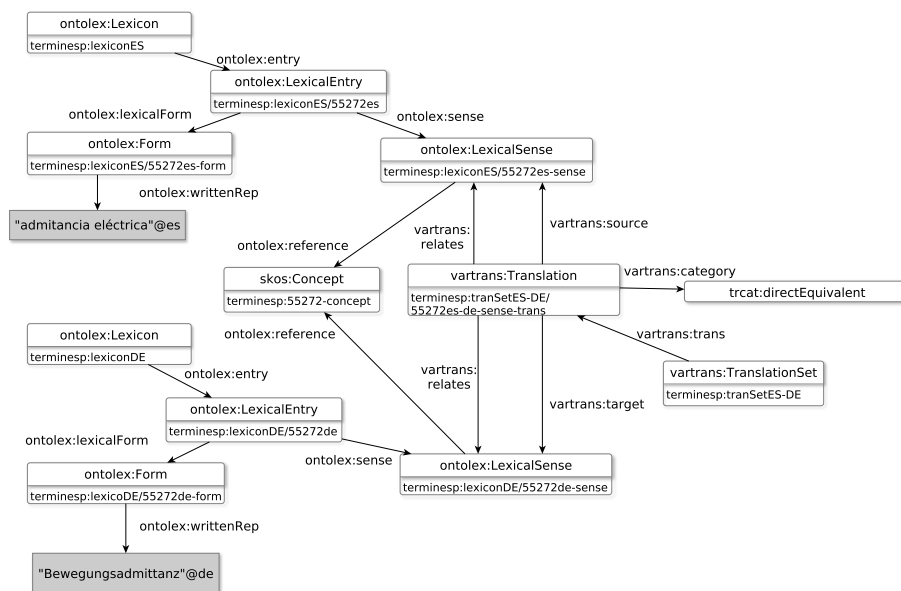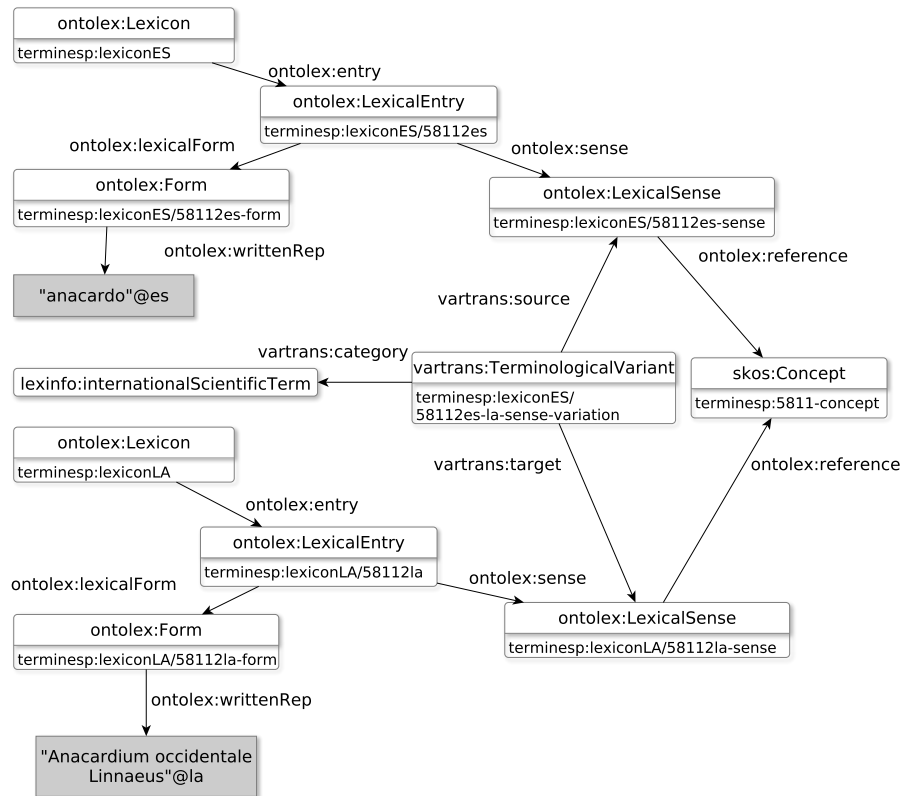

Fig. 2. Modeling of the German translation of the Spanish entry *admitancia cinética* 'motional admittance', *Bewegungsadmittanz.*

4.3 Scientific denominations

Another important issue is the scientific naming of Terminesp terms. The entries that provide a Latin term come mainly from the botanical domain, denoting

---

[10] `http://purl.org/olia/olia.owl#Collocation`

Fig. 3. Modeling of the Spanish entry *anacardo* 'cashew' and its scientific name, *Anacardium occidentale Linnaeus.*

most of them fruits and vegetables. e.g. Sp. *uva* 'grape', Lat. *Vitis vinifera Linnaeus.* These Latin terms could have been modeled as translations from Spanish into Latin, following the approach adopted for all other translations, since, after all, we are dealing with variation across different languages. Nonetheless, and inspired by previous work on this domain, particularly on the LIR (Linguistic Information Repository) model [8], we have decided to identify the Latin entry as the international scientific denomination. In this sense, scientific names are considered a specific type of terminological variants subject to domain and register rather than to any other factor. Also, they are internationally accepted over scientific communities and can appear in texts written in any language, provided that the register and the domain are adequate for their use. Taking this into account, Latin terms are thought here as terminological variants (see Figure 3) and, more specifically, as `lexinfo:InternationalScientificTerm`(-s). Relations among the Latin term and the entries in other languages are not included, since the Spanish lexicon is taken to be the core of the resource, but

we do not discard adding them in future versions. The language tag of Latin entries remains Latin, even though this results in a relation between two senses in different languages that does not use a `vartrans:Translation` element.

5 Conclusions and Future Work

Terminesp has proved to be a suitable testing bench for the OntoLex core and the `vartrans` module. Since OntoLex is still under development, migrating Terminesp to RDF allowed us to carry out a first application of the model on a multilingual and terminological data set to check for any gaps or incongruities in the representation approach. On the one hand, the multilingual nature of Terminesp provided an appropriate use case of the `vartrans` module to account for translations, and on the other hand, the scientific denominations available in some Terminesp entries proved suitable to encode register variation as well. However, we missed the inclusion of a class or property in OntoLex to capture sense definitions or notes to them, since just modeling in terms of `ontolex:LexicalConcept`(-s) would overly complicate the task and did not seem to fit for this particular resource. Lastly, we draw attention to the modeling problems that entries with varying degrees of lexicalization might give rise to in future efforts with terminological data.

As a first future step, the data set will be published as linked data as soon as the OntoLex model is released. This contribution not only aims to report on the first results of applying the model, but the Terminesp RDF data set promises to be a potential significant resource for NLP-based applications. It provides terminological information in several languages, some of them currently under-represented in the LLOD cloud (e.g. Swedish), and translations among the different languages are accessible through the Spanish terms acting as interlingua elements. We will also explore the use of the pool of syntactic phrases discussed in Section 4.2 in multilingual information extraction and language generation tasks.

References

1. Albertoni, R., Martino, M.D., Franco, S.D., Santis, V.D., Plini, P.: EARTh: An Environmental Application Reference Thesaurus in the Linked Open Data cloud. Semantic Web 5(2), 165–171 (2014)
2. Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., Keizer, J.: The AGROVOC Linked Dataset. Semantic Web 4(3), 341–348 (2013)
3. Chiarcos, C., Sukhareva, M.: OLiA – Ontologies of Linguistic Annotation. Semantic Web 6(4) (2015)

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

5. Gracia, J., Montiel-Ponsoda, E., Vila-Suero, D., Aguado-de Cea, G.: Enabling Language Resources to Expose Translations as Linked Data on the Web. In: Proc. of 9th Language Resources and Evaluation Conference (LREC'14), Reykjavik (Iceland). pp. 409–413. European Language Resources Association (ELRA) (May 2014), `http://www.lrec-conf.org/proceedings/lrec2014/pdf/863\_Paper.pdf`

6. Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., Wright, S.E.: ISOcat: Corralling Data Categories in the Wild. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Tapias, D. (eds.) Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (May 2008), http://www.lrec-conf.org/proceedings/lrec2008/

7. McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation 46(4), 701–719 (2012), `http://dx.doi.org/10.1007/s10579-012-9182-3`

8. Montiel-Ponsoda, E., Aguado-de Cea, G., Gómez-Pérez, A., Peters, W.: Enriching ontologies with multilingual information. Natural Language Engineering 17(3), 283 –309 (2010)

9. Neubert, J.: Bringing the "Thesaurus for Economics" on to the Web of Linked Data. In: Bizer, C., Heath, T., Berners-Lee, T., Idehen, K. (eds.) LDOW. CEUR Workshop Proceedings, vol. 538. CEUR-WS.org (2009), `http://dblp.uni-trier.de/db/conf/www/ldow2009.html#Neubert09`

10. Rüther, M., Fock, J., Bandholtz, T., Schulte-Coerne, T.: Linked Environment Data. In: Gerlinde Knetsch, K.J. (ed.) Umweltbundesamt. pp. 89–92. Umweltbundesamt (2010), ISSN 1862-4804