LDK 2023

Formalizing Translation Equivalence and Lexico-Semantic Relations Among Terms in a Bilingual Terminological Resource

Giulia Speranza, Maria Pia di Buono and **Johanna Monti**
University of Naples "L'Orientale"
{gsperanza, mpdibuono,jmonti}@unior.it

Abstract

In this paper we investigate the feasibility of applying the Semantic Web formalisms, in particular the OntoLex-Lemon model, to represent bilingual terminological resources, both from a conceptual and a lexico-semantic point of view. As a proof of concept for our study we select a bilingual Italian-English terminological resource in the specialized domain of archaeology, in order to identify possible modelling solutions as well as potential challenges.

1 Introduction

Recent years have witnessed a significant increase in the conversion and development of lexical resources into RDF, following the Linguistic Linked Open Data (LLOD) principles[1]. Indeed, there is a growing recognition of the importance of the interoperability, reuse and accessibility of data, also in the field of language resources Khan et al. (2022). The employment of Semantic Web formalisms, such as the OntoLex-Lemon model, allows the enrichment of linguistic and terminological resources with structured semantic information, making them easily integrated with other semantic resources, such as ontologies, linked datasets and semantic knowledge bases, thus preventing the so-called data-silos. The rich semantic information that can be easily represented in a resources by means of the Semantic Web formalisms is also beneficial in many applicative scenarios where Natural Language Processing (NLP) is concerned.

Among several formalisms that have been proposed for the formalization of such resources, the OntoLex-Lemon model allows to represent in detail the meaning of terms, the semantic relationships between them, and other related linguistic information, enabling a complete and accurate representation of the entries in terminological resources.

Furthermore, the Ontolex-Lemon model is flexible and easily extendable, offering several representation possibilities to meet different formalization needs.

These achievements are also due to the efforts, experiments, and proposals of a community of researchers and scholars of the W3C Ontology-Lexica Community Group[2] and the Nexus Linguarum COST Action[3], who collaborate on the systematization of models and modules that continue to evolve in order to meet the needs of the LLOD community.

The LLOD principles are being applied to the formalization of several types of resources. Indeed, the analysis carried out by di Buono et al. (2022) about the existing resources and their metadata used to represent them within the LOD Cloud and AnnoHub, which resulted in the creation of METASHARE Enriched LLD (MELLD)[4], a new enriched metadata resource, show that out of the 666 total LLOD resources, 315 are Corpora, 303 are Lexicons and Dictionaries and only 30 are catalogued as Terminologies, Thesauri and Knowledge Bases.

Furthermore, the comprehensive survey by Gromann et al. (forthcoming) sheds light on the linguistic description levels represented in the LLOD resources available and reports several studies focused on the description of the Translation and Terminology level.

Finally, for the description and representation of the terminologies some proposals are also emerging and being discussed such as the TermLex (Martín-Chozas and Declerck, 2022), an extension module for the OntoLex-Lemon model.

In order to contribute to the discussion we investigate the feasibility of applying the Semantic

[1] https://en.wikipedia.org/wiki/Linguistic_Linked_Open_Data

[2] https://www.w3.org/community/ontolex/

[3] https://nexuslinguarum.eu/

[4] https://github.com/unior-nlp-research-group/melld

Web formalisms, in particular the OntoLex-Lemon model, to represent bilingual terminological resources, both from a conceptual and a lexical point of view. As a proof of concept for our study we select a bilingual Italian-English terminological resource in the specialized domain of archaeology, in order to identify possible modelling solutions as well as potential challenges.

2 Case Study

As case study for our modelling experiment we select a bilingual Italian-English terminological resource (TR) in the specialized domain of archaeology.
The TR has been created by means of a semi-automatic extraction process based on appositional constructions from a parallel domain corpus (Speranza et al., 2021, 2022). The TR is composed of 300 terms in each language in the form of single and multi-word units (MWUs) terms.
Furthermore, by means of the terminology extraction methodology previously applied to create the TR, we were also able to enrich it with other information such as Part of Speech (PoS), terminological variants, examples of terms in the context of a sentence and reformulations of technical terms. The inclusion of lay reformulations of technical terms, retrieved hinging on appositional constructions structures, can be a useful information in a TR to be employed for the simplification and exemplification of technicisms in different communicative scenarios involving experts and non-experts.

Starting from our case study our representation needs concern the following information:

- **Terminological entry**: Single and multi-word terms, Syntactic and grammatical information (PoS, gender and number), Context (Example sentence)
- **Lexico-semantic relations**: diaphasic and synonymous variants and taxonomical and translation equivalence relations

3 Modelling Strategy

In order to formalize the TR according to the Linked Open Data principles applied to Linguistics (Cimiano et al., 2020), we choose to adopt the OntoLex-Lemon core model, including some of its specific modules (see table 1), such as the Variation and Translation Module (vartrans), the Decomposition Module (decomp) as well as the LexInfo.

Furthermore, since we also need to represent the Conceptual level of the entries we use the Skos Models.

Prefix	Namespaces
ontolex	http://www.w3.org/ns/lemon/ontolex#
vartrans	http://www.w3.org/ns/lemon/vartrans#
decomp	http://www.w3.org/ns/lemon/decomp#
lexinfo	http://www.lexinfo.net/ontology/2.0/lexinfo#
skos	http://www.w3.org/2004/02/skos#

Table 1: Models and modules' prefixes and namespaces

In particular, we use the Ontolex-Lemon core model to formalize the terminological entries and we use the LexInfo Model as the Data Category Ontology for the representation of grammatical information about the terms.

Furthermore, the decomp module is used for representing the internal structure of MWUs terms, since in our TR many MWUs are endocentric MWUs which present a fixed head, which is usually post-modified by prepositional phrases or through adjectival post-modification as in *anfora a piramide, anfora da trasporto, anfora punica*.

In addition, we use skos for reporting an example of sentence containing the term, which can also be useful for the user of a TR.

Then, we use the vartrans module for representing both the monolingual lexico-semantic relations in Italian or English and the translation equivalence relations between the two languages. Indeed, the vartrans module has been developed to record "lexico-semantic relations across entries in the same or different languages" (Montiel-Ponsoda et al., 2015). In addition, translation relations in Ontolex-Lemon are intended as a special type of lexico-semantic variation (Bosque-Gil et al., 2015) or a special case of a sense relation (McCrae et al., 2017).

Finally, in order to provide for each terminological entry in the resource a conceptual scheme, we use the SKOS Core Vocabulary[5]. SKOS is in fact used for expressing the basic structure of concept schemes i.e., thesauri, taxonomies, terminologies, glossaries and other types of controlled vocabulary.

[5] https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/

3.1 Conceptual Level

Following the Ontolex-Lemon module Specifications[6], SKOS and Ontolex-Lemon can be used in conjunction to provide more detailed information about the "labels". As a consequence, by means of the `skos:concept` property we choose to link each lexical entry to the conceptual schema proposed in the Italian *Istituto Centrale per il Catalogo e la Documentazione* (ICCD) Thesaurus of Archaeological Finds in the SKOS version. The ICCD's Thesaurus is indeed organized according to a hierarchical classification which provides general categories (macro-categories) and specific categories (sub-categories) to conceptually organize the archaeological terms.

For example, the archaeological find *amuleto* (amulet) is a term listed under the macro-category (I° Level) *Strumenti, Utensili e Oggetti d'Uso* (Tools); more precisely belonging to the sub-category (II° level) *Amuleti e Oggetti per uso cerimoniale, magico e votivo* (Magic and votive supplies) (Di Buono, 2015).

In the SKOS version of the ICCD's Thesaurus Felicetti et al. (2013) converted the 10 macro-categories of the taxonomic hierarchy of the ICCD's Thesaurus into different corresponding URIs distinguished by different identifiers from 001 to 010, representing different macro-categories (i.e., *Abbigliamento e Ornamenti personali* (Clothing and Accessories) (001), *Arredi* (Furnishing) (002), *Edilizia* (Building) (003), etc.), linked by means of the `skos:hasTopConcept` property.

In such a way, the Italian lexical entry *anfora da trasporto* can be connected to the conceptual level by means of the `ontolex:sense` property and the lexical sense can point to the `skos:Concept` by means of the `ontolex:reference` property, thus reusing previously set URIs to uniquely identify the concepts in our TR (see figure 1).

Linking each lexical entry to an ontology entity in the CIDOC Conceptual Reference Model (CRM) (Doerr, 2003), which is the reference ontology for Cultural Heritage domain, even if the OntoLex-Lemon module easily allows this operation by means of the `ontolex:denotes` property, would only provide us with a single class for linking our terms in the archaeological domain, namely `E22 Human-Made Object`, since all of our terms conceptually belong to the class of objects made by humans (Human Made Objects).

```
:anfora_da_trasporto_lex a
    ontolex:lexicalEntry, ontolex:
    MultiWord;
       ontolex:sense
           :anfora_da_trasporto_sense

:anfora_da_trasporto_sense a
    ontolex:LexicalSense;
   ontolex: reference
<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/009.
005.000.005.002>

<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/009.
005.000.005.002> a
    ontolex:LexicalConcept;
  skos:inScheme
<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/>;
```

Figure 1: RDF serialization of the conceptual level of the term *anfora da piramide*

3.2 Terminological Entry Level

In order to test the representation of the grammatical information of the terms, we report in Figure 2 the formalization of the Italian lexical entry *anfora da trasporto*.

By means of the Ontolex-Lemon core model we are able to represent different information such as the type of forms a lexical entry can have: a canonical form (*anfora da trasporto*) and another form (*anfore da trasporto*). With `LexInfo` we can further specify some grammatical and syntactic information such as the number (singular and plural), the gender (masculine or feminine) and the PoS about the term.

The decomposition of the MWU terms is realized resorting to the property `decomp:constituent` that relates a lexical entry to its components, as in figure 3.

Moreover, by means of the property `decomp:correspondsTo` we are also able to link the single components of the MWU to the corresponding lexical entries, enabling, as a consequence, the further specification of the linguistic information connected with the lexical entries. Finally, in order to specify the order of the components, it is possible to use the RDF properties `rdf:_1, rdf:_2, etc.`

In addition, in our TR we also provide for each entry an example sentence containing the term extracted from the parallel corpus. We formalize this information resorting to the `skos` module which

[6] https://www.w3.org/2016/05/ontolex/

```
:anfora_da_trasporto_lex a
    ontolex:lexicalEntry, ontolex:
    MultiWord;
       ontolex: canonicalForm
            :form_anfora_da_trasporto_sn;
       ontolex: otherForm
            :form_anfore_da_trasporto_pl;

:form_anfora_da_trasporto_sn a
    ontolex:Form;
       ontolex:writtenRep
       "anfora da trasporto"@it;
       lexinfo:partOfSpeech lexinfo:noun;
       lexinfo:gender lexinfo:feminine;
       lexinfo:number lexinfo:singular.

:form_anfore_da_trasporto_pl a
    ontolex:Form;
       ontolex:writtenRep
       "anfora da trasportoe"@it;
       lexinfo:partOfSpeech lexinfo:noun;
       lexinfo:gender lexinfo:feminine;
       lexinfo:number lexinfo:plural.
```

Figure 2: RDF serialization of the term *anfora da trasporto*

offers the possibility to use the `skos:example` property, as in the figure 4 but it could also be represented resorting to the OntoLex module for Frequency, Attestations, and Corpus-Based Information (OntoLex-FrAC) (Chiarcos et al., 2022), as example sentences are, in our case, corpus attestations.

3.3 Lexico-semantic Relations

3.3.1 Diaphasic variations

As far as the monolingual terminological variation in each language is concerned, the OntoLex-Lemon model Specifications include the diatopic, diaphasic, diachronic, diastratic and dimensional variants as examples of terminological relations.

In our TR, we mainly need to represent the diaphasic relations, especially when Latin or Greek origin terms coexist with the target language variants and are employed in different communicative registers, namely in different communicative situations (Montiel-Ponsoda et al., 2013). In this case, both terminological variants share the same conceptual meaning by pointing to the same `Skos:concept`, while changing their respective surface forms. Therefore, by means of the class `vartrans:TerminologicalVariants` and the property `vartrans:category :diaphasic` we are able to frame this kind of terminological relation between functional variants as in the example of the term *foculo* and its Latin origin variant *foculum* in Figure 5.

```
:anfora_da_trasporto_lex a
    ontolex:LexicalEntry;
   decomp:constituent :anfora_component;
      rdf:_1 :anfora_component;
   decomp:constituent :da_component;
      rdf:_2 :da_component;
   decomp:constituent
       :trasporto_component;
      rdf:_3 :trasporto_component;

:anfora_component a decomp:Component;
   decomp:correspondsTo :anfora_lex.
:da_component a decomp:Component;
   decomp:correspondsTo :da_lex;
:trasporto_component a decomp:Component;
   decomp:correspondsTo :trasporto_lex.
```

Figure 3: RDF serialization of the MWU decomposition of the term *anfora da trasporto*

```
:anfora_da_trasporto_sense a
    ontolex:LexicalSense;
   ontolex: reference
<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/009.
005.000.005.002>

<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/009.
005.000.005.002> a
    ontolex:LexicalConcept;
  skos:inScheme
<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/>;
  skos:example
   rdf:value "Significativa anche la
       presenza di un’anfora da
       trasporto di produzione greca"
       @it
```

Figure 4: RDF serialization of the context sentence example for the term *anfora da trasporto*

3.3.2 Taxonomic relations

In the modelling phase we are also confronted with the need of representing the semantic relation of hypernymy/hyponymy, which can be represented with the `vartrans` module in combination with the `LexInfo` categories (`LexInfo:hypernym` or `LexInfo:hyponym`). We use the property `vartrans:senseRelation`, which connects together two lexical entries' senses and allows the declaration of the `category:hypernym` and the indication of the relation direction from the `source` to the `target` term. In Figure 6 we report the example of the formalization of the relation

```
:foculo_lex a ontolex:LexicalEntry;
      ontolex:lexicalForm :foculo_form;
      ontolex:sense :foculo_sense.
:focylo_form ontolex:writtenRep
    "foculo"@it.
:foculo_sense ontolex:reference
<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/009.
000.000.011>

:foculum_lex a ontolex:LexicalEntry;
      ontolex:lexicalForm :foculum_form
          ;
      ontolex:sense :foculum_sense.
:foculum_form ontolex:writtenRep
    "foculum"@it .
:foculum_sense ontolex:reference
<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/009.0
00.000.011

:foculo_foculum_relation a
      vartrans:TerminologicalRelation;
      vartrans:source :foculo_sense;
      vartrans:target :foculum_sense;
      vartrans:category :diaphasic.
```

Figure 5: RDF serialization of the diaphasic terminological relation between the entries *foculo* and *foculum*

```
:rython_lex a ontolex:LexicalEntry;
      ontolex:sense :rython_sense;
      ontolex:canonicalForm
          :rython_form.
:rython_sense ontolex:reference
<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/009.
005.000.046>
:rython_form ontolex:writtenRep
    "rython"@it .

:coppa_lex a ontolex:LexicalEntry ;
      ontolex:sense :coppa_sense ;
      ontolex:canonicalForm :coppa_form.
:coppa_sense ontolex:reference
<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/009.
005.000.046>
coppa_form ontolex:writtenRep
    "coppa"@it .

:senseRelation a vartrans:SenseRelation;
      vartrans:source :coppa_sense;
      vartrans:target :rython_sense;
      vartrans:category
          lexinfo:hypernym.
```

Figure 6: RDF serialization of the hypernymyc relation between the terms *rython* and *coppa*

between the term *rython* which is a hyponym and *coppa* (cup) which is its hypernym, namely a more generic term.

3.3.3 Synonymous reformulations

By means of the methodology applied to extract bilingual terms from the parallel corpus which is based on a special kind of linguistic constructions between brackets named appositional constructions, we were able to retrieve from our parallel corpus terms and their exemplifications or simplifications in Italian (a) and English (b). Technically speaking, we were able to retrieve *anchors* and *supplements*, which are the two elements composing the appositional construction (Huddleston and Pullum, 2005) as in the example (1).

(1) a. **rhyton** (una coppa a forma di corno)
    b. **rhyton** (a horn-shaped cup)

In a terminological resource it could be useful to also include this kind of synonymous reformulation of technical terms.

In this specific case, the `skos:definition` property is not taken into consideration since what we need to formalize is not a canonical definition as intended by the ISO 1087:2019[7]: "Representation of a concept (3.2.7) by an expression that describes it and differentiates it from related concepts" which normally are much more complex and articulated (Magris, 1998).

This kind of reformulation could be intended as a very short descriptive definition of the term in plain language with the aim of simplify and explain the technical concept. From this point of view, they can not obviously include a fine-grained and nuanced level of definition.

3.3.4 Translation equivalence relations

Finally, since we need to formalize a bilingual TR, among the several possibilities provided in the OntoLex-Lemon model Specifications, we choose to represent equivalent translations by means of the `vartrans:Translation` class and the properties `vartrans:source` and `vartrans:target`, which also enable the explicit indication of the translation direction. The two lexical entries in the two languages (Italian and English) can be connected to the conceptual level by means of the `ontolex:sense` property, pointing to the `skos:Concept`. Since the two entries in the two languages share the same concept, they can be linked together in a relationship

[7] https://www.iso.org/obp/ui/#iso:std:iso:1087:ed-2:v1:en

of translation equivalence at sense level by means of the `vartrans` module, by even specifying the translation direction from the Italian `source` (*anfora da trasporto*) to the English `target` (transport amphora) (see figure 7).

```
:anfora_da_trasporto_sense
    ontolex:reference
<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/009.
005.000.005.002>
:transport_amphora_sense
    ontolex:reference
<https://dati.beniculturali.it/lodview/v
ocabularies/reperti_archeologici/def/009.
005.000.005.002>

:trans a vartrans:Translation ;
      vartrans:source
          :anfora_da_trasporto_sense ;
      vartrans:target
          :transport_amphora_sense .
```

Figure 7: RDF serialization of relation of translation equivalence between the lexical entry *anfora da trasporto* and *transport amphora*

4 Conclusions and Future Works

In this paper we tried to formalize a bilingual terminological resource in Italian and English using the vocabularies offered by the Semantic Web Formalisms.

OntoLex-Lemon model with its modules in conjunction with `LexInfo` and `SKOS` resulted to be detailed and flexible enough for covering all the representation needs of our specific TR both from the monolingual and the bilingual point of view.

During the modelling phase we were, nevertheless, confronted with the challenge of representing special kinds of synonymous reformulations extracted from the corpus that we wanted to include in the TR. Possible modelling solutions are offered by the `Lexinfo` category `synonym` which "Indicates the the terms have the same meaning lexicographically"[8] or by the `Lexinfo` category `gloss`, which according to the TEI is "A phrase or word used to provide a gloss or definition for some other word or phrase."[9]. Nonetheless, these options might be limiting from one perspective, since they do not account for the actual status of linguistic reformulations of terminology in plain language.

Future works might therefore be needed to meet specific necessities related to particular representations as long as further information about terms such as reformulations of technical terms or very short descriptive definitions are needed to be included and addressed in a TR more directly.

Finally, in terms of applicability, terminological resource formalized with OntoLex-Model can also be easily converted in other formats which are also widely employed for the representation, storing and sharing of terminological resources, such as the TBX, which can be used in CAT-Tools for translation purposes.

[8] https://lexinfo.net/index.html

[9] https://tei-c.org/release/doc/tei-p5-doc/it/html/examples-gloss.html

References

Julia Bosque-Gil, Jorge Gracia, Guadalupe Aguado-de Cea, and Elena Montiel-Ponsoda. 2015. Applying the ontolex model to a multilingual terminological resource. In *European Semantic Web Conference*, pages 283–294. Springer.

C. Chiarcos, Elena Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. Modelling frequency, attestation, and corpus-based information with ontolex-frac. In *International Conference on Computational Linguistics*.

Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. Linguistic linked open data cloud. In *Linguistic Linked Data*, pages 29–41. Springer.

Maria Pia Di Buono. 2015. Information extraction for ontology population tasks. an application to the italian archaeological domain. *International Journal of Computer Science: Theories and Applications*, 3(2):40–50.

Maria Pia di Buono, Hugo Gonçalo Oliveira, Verginica Barbu Mititelu, Blerina Spahiu, and Gennaro Nolano. 2022. Paving the way for enriched metadata of linguistic linked data. *Semantic Web*, vol. 13(6):1133–1157.

Martin Doerr. 2003. The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75.

Achille Felicetti, Tiziana Scarselli, Maria Letizia Mancinelli, and Franco Niccolucci. 2013. Mapping iccd archaeological data to cidoc-crm: the ra schema. In *CRMEX@ TPDL*, pages 11–22.

Dagmar Gromann, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Verginica Mititelu, Liudmila Mockiene, Michael Rosner, et al. forthcoming. Multilinguality and llod: A survey across linguistic description levels. *Semantic Web*.

Rodnry Huddleston and Geqffrry Pullum. 2005. The cambridge grammar of the english language. *Zeitschrift für Anglistik und Amerikanistik*, 53(2):193–194.

Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena Gonzalez-Blanco Garcia, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P McCrae, et al. 2022. When linguistics meets web technologies. recent advances in modelling linguistic linked open data. *Semantic Web Journal*.

Marella Magris. 1998. La definizione in terminologia e nella traduzione specialistica. *EUT-Edizioni Università di Trieste*.

Patricia Martín-Chozas and Thierry Declerck. 2022. Representing multilingual terminologies with ontolex-lemon. In *1st International Conference on "Multilingual digital terminology today. Design, representation formats and management systems", June 16 – 17, Padova, Italy*.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Elena Montiel-Ponsoda, Julia Bosque-Gil, Jorge Gracia, Guadalupe Aguado de Cea, and Daniel Vila-Suero. 2015. Towards the integration of multilingual terminologies: an example of a linked data prototype. In *TIA*, pages 205–206.

Elena Montiel-Ponsoda, John P McCrae, Guadalupe Aguado de Cea, and Jorge Gracia del Río. 2013. Multilingual variation in the context of linked data. In *Proceedings of 10th International Conference on Terminology and Artificial Intelligence (TIA'13)*.

Giulia Speranza, Maria Pia Di Buono, and Johanna Monti. 2021. Terms and appositions: What unstructured texts tell us. In *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ*, pages 219–230. Springer.

Giulia Speranza, Maria Pia Di Buono, and Johanna Monti. 2022. Tailoring terminological resources to the users' needs: a corpus-based study on appositive constructions in italian and english. In *CEUR Workshop Proceedings: 1st International Conference on "Multilingual Digital Terminology Today. Design, representation formats and management systems", 16 - 17 June 2022, Padua, Italy*.