Porting the Latin WordNet onto OntoLex-Lemon

Stefania Racioppa[1], Thierry Declerck[1,2]

[1]German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus D3 2, Saarbrücken, Germany
[2]Austrian Centre for Digital Humanities and Cultural Heritage
Sonnenfelsgasse 19, Wien 1010, Austria
E-mail: stefania.racioppa@dfki.de, declerck@dfki.de

Abstract

In this paper we describe the porting of the Latin WordNet data available at the University of Exeter onto the OntoLex-Lemon model, focusing on the representation of both morphological and conceptual information. In the longer term, we aim at integrating the resulting data set in the Linguistic Linked Open Data (LLOD) infrastructure, linking (or even merging) it to the Latin data sets already published in the LOD framework by the ERC "Linking Latin" (LILA) project. We discuss some lessons learned, as it turned out that such a transformation and linking exercise can lead to an improved consistency and accuracy of the original data.

Keywords: Latin; WordNet; Morphology; OntoLex-Lemon

1. Introduction

In our work, we are concerned with the transformation of heterogeneous digital lexical resources, available in a multitude of formats, into a harmonised representation in the context of the OntoLex-Lemon model,[1] which is briefly introduced in Section 3 of this paper. Besides mainstream language resources, we are also dealing with ancient and low-resourced languages, as we are aiming at contributing to the improved access to such resources, which could further support the deployment of language technologies in the broader field of digital humanities.

Our first steps in dealing with Latin language data consisted in mapping the Latin WordNet available at the University of Exeter[2] onto the OntoLex-Lemon model. We set the main focus on the semantic representation of both the morphological and conceptual information encoded in the Latin WordNet. In this context, we are also starting a cooperation with the ERC project "LiLa" (*Linking Latin Building a Knowledge Base of Linguistic Resources for Latin*)[3] on the harmonisation of the semantic representation of Latin language data for their optimal publication on the Linguistic Linked Open Data cloud.

In the following sections, we introduce first the Latin WordNet data of the University of Exeter, and describe then briefly the Linguistic Linked Open Data cloud as well as the OntoLex-Lemon representation model for lexical data. We continue with the presentation of the first results of the mapping of the Latin WordNet data onto OntoLex-Lemon, comparing them with the Latin data already ported to the Linked Open Data by the LILA project. We close with the discussions of some lessons learned.

2. Latin WordNet at the University of Exeter

The Latin WordNet initiative at the University of Exeter "builds on, and extends, the original Latin WordNet developed as part of the Fondazione Bruno Kessler's

[1] https://www.w3.org/2016/05/ontolex/. See also (Cimiano et al., 2016).

[2] https://latinwordnet.exeter.ac.uk/. See also (Fedriani et al., 2020).

[3] https://lila-erc.eu/. See also (Mambrini & Passarotti, 2019) and the Latin Lemma Bank Query Interface of the LiLa project, available at https://lila-erc.eu/query/.

MultiWordNet project”[4] and is developed in the context of a cooperation, among others, with the University of Genoa[5] and the LiLa project.[6]

Periodically updated versions of the Latin WordNet are available in two formats (JSON and CSV) in a GitHub repository.[7] The data is distributed over distinct files for different categories, from which we considered the files displaying information on the lemmas, literal senses and synsets.

After working on the CSV data set of January 2020,[8] we communicated some issues we found in the source data to the resource developer. In a second step, we worked on the CSV data set of October 2020.[9] Also in this case, the communication with the developer was essential to solve the remaining open issues.

Concerning the lemma information associated with the synsets, both data sets differ slightly in their layout. The January version included the following columns in the files containing the lemmas: *ID*, *URI*, the lemma itself, *part of speech*, *morphological information*, *principal parts*, *irregular forms*, *alternative forms*, *IPA phonetic representation*, *prosody*, and *validation id*. In the October data set, the irregular and alternative forms, as well as the phonetic representation have been dropped, and the column order was changed. The information included in the distinct files are described in detail in Listings 2.1, 2.2 and 2.3.

In Listing 2.1, displaying the lemma “abdicatio” (in the version of 2020.10.10), we can see the lexical and morphological information associated with the lemma, that needs to be represented in OntoLex-Lemon. The lemma is included in the file “lemma_0.csv” with the ID “19117”. This ID is present three times in the file “literalsense_0.csv”, indicating that the lemma has 3 senses pointing to the synsets “136508”, “136706” and “104057”, which are included in the file “synset_10.csv”, and “synset_0.csv”.

Listing 2.3 displays the information associated with the synsets, where the glosses are from the Princeton WordNet,[10] while in Listing 2.2 we can see how the synsets are related to the lemmas by the use of their respective IDs.

```
id,uri,lemma,pos,validated,morpho,principal_parts,prosody
19117,a0031,abdicatio,n,1,n–s——fn3−,abdication,abdicatio
```

Listing 2.1: The entry *abdicatio* in the 2020-10-10 data set

```
id,lemma,synset,period,genre,notes
2,19117,136508,,,
3,19117,136706,,,
4,19117,104057,,,
```

Listing 2.2: The literal senses for the lemma *abdicatio* in the 2020-10-10 data set

[4] Quoted from https://latinwordnet.exeter.ac.uk/. See also (Fedriani et al., 2020).

[5] This cooperation is documented, for example, in (Fedriani et al., 2020).

[6] The “Linked Latin” (LILA) is a project funded by the European Research Council (ERC). See https://lila-erc.eu/ for more details. See also (Passarotti et al., 2019; Mambrini & Passarotti, 2019).

[7] https://github.com/latinwordnet/latinwordnet-archive/tree/master/csv/.

[8] https://github.com/latinwordnet/latinwordnet-archive/tree/master/csv/2020-01-31.

[9] https://github.com/latinwordnet/latinwordnet-archive/tree/master/csv/2020-10-10.

[10] https://wordnet.princeton.edu/. See also (Fellbaum, 1998).

```
id,offset,pos,language,gloss,semfield
136508,05385235,n,10,refusal to acknowledge as one's own,
136706,05414335,n,10,a verbal act of renouncing,
104057,00134568,n,10,the act of renouncing,
```

Listing 2.3: The synsets to which the literal senses for the lemma *abdicatio* are pointing to (in the 2020-10-10 data set)

3. OntoLex-Lemon

The OntoLex-Lemon model, which results from a W3C Community Group,[11] was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the labels, definitions or comments of ontology elements are equipped with an extensive linguistic description.[12] This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as their syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialised vocabularies.

The main organising unit for those linguistic descriptions is the *LexicalEntry* class, which enables, among other things, the representation of morphological patterns for each entry (a multi-word expression, a word or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the *ontolex:denotes* property or is mediated by the *LexicalSense* or the *LexicalConcept* classes, as this is represented in Figure 1, which displays the core module of the model.

OntoLex-Lemon builds on and extends the preceding *lemon* model (McCrae et al., 2012). A major difference is that OntoLex-Lemon includes an explicit way to encode conceptual hierarchies, using the SKOS[13] standard. As can be seen in Figure 1, lexical entries can be linked, via the *ontolex:evokes* property, to such SKOS concepts, which can represent WordNet synsets. This structure aligns the relation between lexical entries and ontological resources, which is implemented either directly by the *ontolex:reference* property or mediated by the instances of the *ontolex:LexicalSense* class.

More recently, OntoLex-Lemon has been used also as a de facto standard in the field of digital lexicography and is being applied for example in the European infrastructure project ELEXIS (European Lexicographic Infrastructure).[14]

4. Representation of the Latin WordNet Lemmas in OntoLex-Lemon

The modelling of the linguistic information from the Latin WordNet data within OntoLex-Lemon took into consideration the recent morphology module, currently under (advanced) discussion within the W3C "Ontology-Lexica" Community Group[15], in which

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

[12] See (Cimiano et al., 2016).

[13] SKOS stands for "Simple Knowledge Organization System". SKOS provides "a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary" (https://www.w3.org/TR/skos-primer/).

[14] See http://www.elex.is/ for more detail.

[15] See (Klimek et al. (2019))

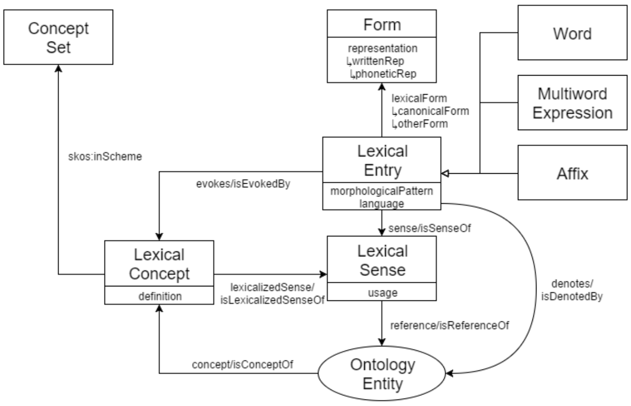

Figure 1: The core modules of OntoLex-Lemon. Graphic taken from https://www.w3.org/2016/05/ontolex/

(among others) also members of the LiLa developer team are actively involved. However, as the discussion is still ongoing, we cannot exclude discrepancies with the most recent model definition.

In a first step, we performed a light data clean-up, i.e. merging split entries, and separating the elements in the column *principal parts* consistently with blanks. In the "cleaned" files, we looked for potential duplicates. While identical entries can just be dropped in the generated OntoLex-Lemon compliant output, in some cases we detected lemmas with the same part of speech, but different genders, or declension groups. As we cannot decide if these are actually errors, homographs with different senses, or if the lemmas really allow for different inflections, we shared our findings with the developers and are currently iteratively adapting and running our transformation process from the updated CSV data onto OntoLex-Lemon.

Analysing the available data in the lemma category of both January and October data sets, we found out that the required morphological features were represented in a quite structured form for each lemma, which includes in the "morpho" column an abbreviated information, i.e. `n-s---fn3-`. This indicates the values for, respectively, part of speech (here: *noun*), adjective degree, number (*singular*), verb tense, mood, and voice, gender (*feminine*), case (*nominative*), declension group (*3rd*), and stem variations (where applicable, e.g. in *abnept-abneptis*: `n-s---fn3i`).

The morphological information was not changed in the latest Latin WordNet version, so that we were able to map the *morpho* value of both January and October data sets into an OntoLex-compatible format using a simple Python script. As a side effect, the script also helped us to highlight and remove the very few errors in the original data.

For the further processing, we split the data by part of speech and converted it in a "readable" CSV/Pandas format, as shown in Listing 4.1 below:[16]

```
,base ,forms ,pos ,number ,gender ,case ,group ,fonipa ,stem ,degree ,person ,tense ,
mood,voice
,abdicatio ,abdication ,noun ,singular ,feminine ,nominative ,3 ,− ,,,,,,

,base ,forms ,pos ,number ,gender ,case ,group ,stem ,degree ,person ,tense ,mood,
voice
```

[16] The phonetic transcription (value *fonipa*) of the January dataset is not displayed in this example.

```
,abdicatio ,abdication ,noun ,singular ,feminine ,nominative ,3 ,,,,,,
```

Listing 4.1: CSV/Pandas output for the entry *abdicatio*
from the 2020-01-31 and 2020-10-10 data set

As the readers can notice, some values — as well as the meaning of the used "codes" — depend partially on the part of speech of the corresponding lemma, and the part of speech is listed separately in a dedicated column. Above this, the Latin declensions can (mostly) be recognised by the ending of the lemma. All these factors helped us in this phase to detect several inconsistencies in the original data, such as a wrong gender or declension groups, or even inconsistent part of speech information.

After processing the January data set, we forwarded our findings to the Latin WordNet developers, and some corrections were implemented in the following versions. The inconsistencies found in the October data set are currently under revision. A first feedback from the developer confirmed that some items were indeed mis-tagged, although the "morpho" fields are mostly correct. However, examining the apparently "duplicated" entries might be more complicated. Some highlighted items seem to be morphological variants, which need to be checked also with respect to the semantic distance between the items. While "real" variants can be merged, it is possible that others mean something different, in which case it would be reasonable to keep them as distinct lemmas.

Also, the *prosody* column plays a relevant role in the lemma disambiguation (e.g. *scŏpa* vs. *scōpa*), and it might be worth including this piece of information in a future OntoLex version of this resource. In general, the Latin WordNet can be seen as "work in progress", so that besides this, further changes might be made in the future.

As the Latin WordNet does not include full forms or the declension tables corresponding to the defined groups, we decided to represent the lemma inflection not as full-form reference, but using the morphological patterns principle described in the OntoLex Morphology Module, which explicitly recommends linking to external sources for such purposes. We found a detailed description of the Latin declension groups in Wikipedia[17] and mapped the declension tables listed there into the OntoLex-Lemon format.

This work resulted in the generation of 73,949 entries (19,999 adjectives, 38,135 nouns, 60 prepositions, 4902 adverbs, 10,854 verbs) from the January data set, and 73,945 entries (19,999 adjectives, 38,130 nouns, 60 prepositions, 4,901 adverbs, 10,855 verbs) from the October data set, as well as 1,219 morphological patterns (192 for nouns, 192 for adjectives and 835 for verbs). However, as the possible inconsistencies we mentioned above are currently under review, the final figures might change in the future.

Listing 4.2 displays the OntoLex-Lemon lemma for "abdicatio" and its forms. The representation is the same for both data sets. However, the IPA phonetic representation was dropped in the latest version.

```
:lex_abdicatio a ontolex:LexicalEntry ;
    lexinfo:gender lexinfo:feminine ;
    lexinfo:partOfSpeech lexinfo:noun ;
    morph:inflects :la—noun_3 ;
    ontolex:canonicalForm :form_abdicatio ;
    ontolex:evokes :a0031 ;
```

[17] https://en.wikipedia.org/wiki/Latin_declension

```
    ontolex:otherForm :form_abdicatio_root .

:form_abdicatio a ontolex:Form ;
    lexinfo:case lexinfo:nominative ;
    lexinfo:number lexinfo:singular ;
    ontolex:writtenRep "abdicatio"@la .

:form_abdicatio_root a ontolex:Form ;
    ontolex:writtenRep "abdication"@la .
```

Listing 4.2: The OntoLex-Lemon representation for *abdicatio* including the “canonical” and the “other” forms

The corresponding morphological pattern and some associated “rules” are displayed in Listing 4.3. In the examples, we can see the entries for the accusative forms, singular (*abdicationem*) and plural (*abdicationes*). The inflections are represented in the “replacement” value as a pattern, using the syntax of regular expressions.

The feature *generates* lists the morphological information related to each inflection. Alternative values (*lexinfo:feminine*, *lexinfo:masculine*) indicate the allowed morphological information, which is disambiguated by the corresponding value in the “main” entry.

```
:la-noun_3 a morph:paradigm ;
    rdfs:comment "Latin 3rd noun declension" .

:la-noun_3_acc_m-f_pl a morph:rule ;
    morph:generates [ lexinfo:case lexinfo:accusative ;
            lexinfo:gender lexinfo:feminine ,
                lexinfo:masculine ;
            lexinfo:number lexinfo:plural ] ;
    morph:paradigm :la-noun_3 ;
    morph:replacement [ morph:source "$" ;
            morph:target "es" ] .

:la-noun_3_acc_m-f_sg a morph:rule ;
    morph:generates [ lexinfo:case lexinfo:accusative ;
            lexinfo:gender lexinfo:feminine ,
                lexinfo:masculine ;
            lexinfo:number lexinfo:singular ] ;
    morph:paradigm :la-noun_3 ;
    morph:replacement [ morph:source "$" ;
            morph:target "em" ] .
```

Listing 4.3: The *la-noun_3* paradigm and some of the associated rules

This way, we are making the morphological information available in a declarative manner.

5. The OntoLex-Lemon Representation of the Synsets of Latin WordNet and their Relations to the Lemmas

The original Latin WordNet corpus includes 107,687 synsets, which are taken from Princeton WordNet. The mapping from the original conceptual data in CSV format onto OntoLex-Lemon was simpler to achieve as for the lexical and morphological data, as there was no need to design paradigms or rules to be included in the target representation format.

Listing 5.1 displays an example of a synset, encoded as an instance of the *LexicalConcept* class, and the way it is related to the instances of the *LexicalEntry* class that "evokes" it.

```
:LexicalConcept_134535
  skox:definition "a line drawn on a map connecting points of equal height" ;
  ontolex:isEvokedBy :lex_conputatio ;
  ontolex:isEvokedBy :lex_configuratio ;
  ontolex:isEvokedBy :lex_computatio ;
  ontolex:isEvokedBy :lex_idolon ;
  ontolex:isEvokedBy :lex_efformatio ;
  ontolex:isEvokedBy :lex_sinus ;
  ontolex:isEvokedBy :lex_circumcaesura ;
  ontolex:isEvokedBy :lex_spectrum ;
.
```

Listing 5.1: The OntoLex-Lemon representation for the original synset with id *134535* – including the Princeton WordNet definition and the links to the lexical entries realising the lexical concept

We noticed that many synsets have not yet been related to a Latin word (or lemma). We also discovered that some synsets are on the contrary linked to a multitude of lemmas, like the example in Listing 5.2, which clearly points to an issue in the granularity of the relations between synsets and lemmas in the current version of the data set.

```
:LexicalConcept_134565
  skox:definition "a symbol used to represent a number:
    'he learned to write the numerals before he went to school'" ;
  ontolex:isEvokedBy :lex_numerus ;
  ontolex:isEvokedBy :lex_dessignatio ;
  ontolex:isEvokedBy :lex_plurimus ;
  ontolex:isEvokedBy :lex_auditus ;
  ontolex:isEvokedBy :lex_simplum ;
  ontolex:isEvokedBy :lex_conplus ;
  ontolex:isEvokedBy :lex_carnuficina ;
  ontolex:isEvokedBy :lex_caudex ;
  ontolex:isEvokedBy :lex_compactura ;
  ontolex:isEvokedBy :lex_penecostas ;
  ontolex:isEvokedBy :lex_connubium ;
  ontolex:isEvokedBy :lex_flexio ;
  ontolex:isEvokedBy :lex_quoteni ;
  ontolex:isEvokedBy :lex_reuolutio ;
  ontolex:isEvokedBy :lex_chilias ;
  ontolex:isEvokedBy :lex_aditio ;
  ontolex:isEvokedBy :lex_offa ;
  ontolex:isEvokedBy :lex_cybus ;
  ontolex:isEvokedBy :lex_simulacrum ;
  ontolex:isEvokedBy :lex_infrequentia ;
  ontolex:isEvokedBy :lex_plurimum ;
  ontolex:isEvokedBy :lex_frenus ;
  ontolex:isEvokedBy :lex_binio ;
  ontolex:isEvokedBy :lex_trias ;
  ontolex:isEvokedBy :lex_compar ;
  ....
  ....
```

Listing 5.2: The OntoLex-Lemon representation for the original synset with id *134565* (in the January version) – with a very high number of lemmas that are referred to

6. Linguistic Linked Open Data Cloud

The Linguistic Linked Open Data (LLOD) cloud is an initiative started in 2012 by a working group of the Open Knowledge Foundation.[18] The aim of the initiative was to break the data silos of linguistic data and thus encourage Natural Language Processing (NLP) applications that make use of data from multiple languages and modalities (e.g., lexicon, corpora, etc.). Technologies for representing language data in the LLOD include tools for the discovery, transformation and linking of language data sets which can be applied to both data and metadata, in order to provide multi-portal access to heterogeneous data repositories.

Looking at the current state of the LLOD, displayed in Figure 2, the reader can see that the data sets published in this cloud are classified along the lines of six categories:

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

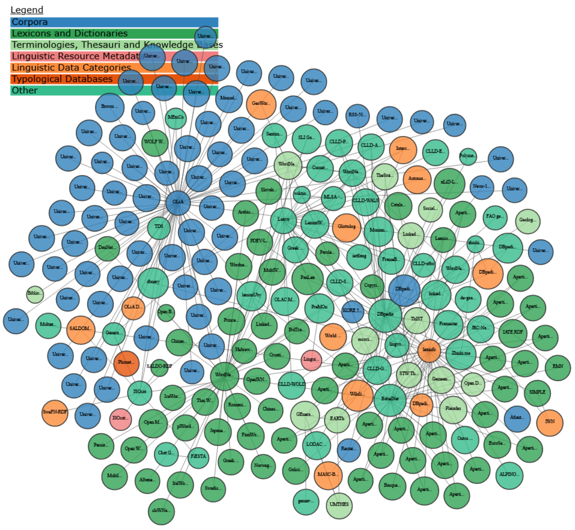
Legend
Corpora
Lexicons and Dictionaries
Terminologies, Thesauri and Knowledge Bases
Linguistic Resource Metadata
Linguistic Data Categories
Typological Databases
Other

Figure 2: The Linguistic Linked Data Cloud

The final goal of our work is to publish as many language data as possible in the LLOD cloud, and to do this, a representation of the data with the Resource Description Framework (RDF) is a prerequisite.

The research community involved in the development of the LLOD cloud aims at increasing the uptake of language technologies also in the broader field of digital humanities and cultural heritage. Dealing with historical languages and porting them to RDF is therefore an important achievement.

[18] See https://linguistic-lod.org/llod-cloud and (McCrae et al., 2016).

The encoding of the Latin in WordNet in RDF and OntoLex-Lemon also allows to establish more precise comparisons with the Latin data already available in the Linked Data framework, resulting from the work pursued by the "Linked Latin" (LiLa) project.[19]

Apart from the different naming of the single features and values, the OntoLex-Lemon representation of our example "abdicatio" (displayed above in Listing 2.1) and the corresponding LiLa lemma (https://lila-erc.eu/data/id/lemma/86857, displayed below in *turtle* format) indeed show a large degree of compatibility: Both have in the "main" entry dedicated values for part of speech and gender definition, as well as a written representation of the lemma itself and a reference to the inflection class.

The main difference between both corpora is how the inflected forms of the lemma are handled. While the OntoLex-Lemon representation just builds a plain reference to the canonical and the "other" form(s) (*abdicatio*, *abdication*), the LiLa representation offers a better analysis of the lemma, because it labels its constituent elements - prefix, radix, and suffix (*ab-*, [base], *-(t)io(n)*). Above this, LiLa adds a reference to the lemma group "dico", which is inflected similarly. Finally, the synset value is a specific feature of the Latin WordNet corpus.

```
<data/id/lemma/86857> a lila:Lemma ;
     rdfs:label "abdicatio" ;
     lila:hasBase <data/id/base/8> ;
     lila:hasGender lila:feminine ;
     lila:hasInflectionType lila:n3 ;
     lila:hasPOS lila:noun ;
     lila:hasPrefix <data/id/prefix/1> ;
     lila:hasSuffix <data/id/suffix/2> ;
     ontolex:writtenRep "abdicatio" .

<data/id/base/8> a lila:Base ;
     rdfs:label "Base of dico" .

<data/id/prefix/1> a lila:Prefix ;
     rdfs:label "a(b)−" .

<data/id/suffix/2> a lila:Suffix ;
     rdfs:label "−(t)io(n)" .
```

Listing 6.1: LiLa lemma representation for "abdicatio" in *turtle* format

For this reason, both data sets could be put in relation by using the OntoLex-Lemon element they have in common: the value of the *ontolex:writtenRep* property. It would also be straightforward to establish a mapping between the LiLa properties expressing the morphological information and the corresponding properties of the LexInfo vocabulary,[20] which are used in OntoLex-Lemon. This way, we could detect which elements are only in one of the data sets, or if inconsistencies are present in describing one and the same phenomenon.

Last but not least, we could suggest the merging of (compatible) pieces of information. Just to mention a few examples, we could share the value of the associated synset from the

[19] Repeating a former footnote for the convenience of the reader: https://lila-erc.eu/. See also (Mambrini & Passarotti, 2019) and the Latin Lemma Bank Query Interface of the LiLa project, available at https://lila-erc.eu/query/.

[20] See https://lexinfo.net/ontology/3.0/lexinfo.

OntoLex-Lemon entry (expressed in the property *evokes*) with the LiLa representation of the same lemma. On the other hand, as mentioned above, LiLa offers a more detailed analysis of the lemma decomposition (i.e. the values *hasPrefix* and *hasSuffix*), which would complete the shallow representation of alternative forms in OntoLex-Lemon (i.e. the simple value *otherForm* and its written representation).

While this is work we have ahead of us, it shows perspective for cross-linked or event unified resources for the Latin language.

7. Lessons Learned

Our work on porting the Latin WordNet onto a Linked Data-compliant format has reinforced our conviction that the encoding in such a format is an added value, as the information contained in the original data set is made available in a declarative way, which supports its linking to other sources of information. Here we see particularly the possibility to cooperate with the LiLa project, as the data encoding is really interoperable.

Another added value lies in the fact that such (automated) transformation work helps to detect potential inconsistencies in the original data. We experienced this in both morphological and conceptual aspects of the CSV data we were working with. The new versions of the Latin WordNet could also benefit of the feedback given to the developer. A simple example of small errors in the conceptual domain is the missing of correct data in a column of the CSV file. Something very difficult to find manually, but which causes an error message when running the Python code to generate the OntoLex-Lemon representation.

8. Conclusions

We presented the current state of our work consisting in mapping the Latin WordNet data onto the OntoLex-Lemon model, in order to support its publication in the Linguistic Linked Open Data cloud. This way this type of language resources can be made directly accessible to NLP applications in the field of eLexicography and digital humanities.

The next steps in our work will be directed at a close cooperation with the LiLa project, towards the best possible semantic representation of Latin language data for their consumption on the Web of Linguistic Linked Data. Thereby we will aim at linking to both encyclopaedic resources, DBpedia[21] and Wikidata,[22] in order to link the Latin language data to additional extra-linguistic information.

Our data set and the algorithms for generating the OntoLex-Lemon representation will be made freely available, either at the GitHub repository of the Latin WordNet or within the LOD presence of the LiLa project.

9. Acknowledgements

Contributions by the German Research Center for Artificial Intelligence (DFKI GmbH) were supported in part by the H2020 project Prêt-à-LLOD with Grant Agreement number

[21] https://wiki.dbpedia.org/.

[22] https://www.wikidata.org/wiki/Wikidata:Main_Page.

825182. Contributions by the Austrian Centre for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences were supported in part by the H2020 project “ELEXIS” with Grant Agreement number 731015. The work described in this paper was also pursued in part in the larger context of the COST Action CA18209 - NexusLinguarum “European network for Web-centred linguistic data science”. We also thank the anonymous reviewers for their insightful comments.

10. References

Cimiano, P., McCrae, J.P. & Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report.

Fedriani, C., Felice, I.D. & Shorth, W.M. (2020). The Digital Lexicon Translaticium Latinum: Theoretical and Methodological Issues. In C. Marras, M. Passarotti, G. Franzini & E. Litta (eds.) *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica*. Associazione per l'Informatica Umanistica e la Cultura Digitale, pp. 106–113.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge, MA: MIT Press.

Klimek, B., McCrae, J., Bosque-Gil, J., Ionov, M., Tauber, J. & Chiarcos, C. (2019). Challenges for the Representation of Morphology in Ontology Lexicons. In *Proceedings of eLex 2019*. URL https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_20.

Mambrini, F. & Passarotti, M. (2019). Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 13th Linguistic Annotation Workshop*. Florence, Italy: Association for Computational Linguistics, pp. 71–80. URL https://www.aclweb.org/anthology/W19-4009.

McCrae, J., de Cea, G.A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6), pp. 701–709.

McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A. & Pool, J. (2016). The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In N.C.C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. ELRA, 9, rue des Cordelières, 75013 Paris: ELRA.

Passarotti, M.C., Cecchini, F.M., Franzini, G., Litta, E., Mambrini, F. & Ruffolo, P. (2019). The LiLa Knowledge Base of Linguistic Resources and NLP Tools for Latin. In *LDK*.