Porting Multilingual Morphological Resources to OntoLex-Lemon

Thierry Declerck
German Research Center
for Artificial Intelligence
Saarbrücken, Germany
thierry.declerck@dfki.de

Stefania Racioppa
German Research Center
for Artificial Intelligence
Saarbrücken, Germany
stefania.racioppa@dfki.de

Abstract

We describe work consisting in porting various morphological resources to the OntoLex-Lemon model. A main objective of this work is to offer a uniform representation of different morphological data sets in order to be able to compare and interlink multilingual resources and to cross-check and interlink or merge the content of morphological resources of one and the same language. The results of our work will be published on the Linguistic Linked Open Data cloud.

1 Introduction

A significant number of linguistic resources have been published on the Linguistic Linked Open Data cloud (LLOD),[1] and projects like ELEXIS[2] and Prêt-à-LLOD[3] are contributing to its further extension. But we notice that only very few, if any, specific morphological resources are included in the LLOD cloud. Available morphology information is mostly contained in lexical or dictionary entries. Our aim is to make also specialized morphological resources available in this cloud. With this step we want to support the interlinking of such resources with other types of linguistic data, in a multilingual fashion, extending work described in (Gromann and Declerck, 2019), which is linking synsets of the Princeton WordNet (PWN) (Fellbaum, 1998) that are associated with plural forms to full lexical descriptions.

A first step of our current work consisted in mapping the MMorph[4] set of multilingual morphological resources to the OntoLex-Lemon model.[5]

In the next sections, we first describe briefly the Linguistic Linked Open Data cloud, and one of its salient component, the OntoLex-Lemon model. Following this summary, we describe the (multilingual) morphological resources we selected for mapping to OntoLex-Lemon. We present the result of such mappings and conclude with a description of the next steps of our work, aiming at supporting the cross-lingual comparison of morphological resources, and the cross-checking, correcting and merging of different morphological resources for one and the same language.

2 The Linguistic Linked Open Data Cloud

The LLOD initiative had its inception in 2012 at a workshop co-located with the 34th Annual Conference of the German Linguistic Society (DGfS). The workshop was organized by members of the Open Knowledge Foundation,[6] and the contributions to this workshop are available in (Chiarcos et al., 2012). The workshop has been a point of focal activity for several research and infrastructure projects, as well as for the "Ontology Lexica" W3C Community Group.[7] Those developments are described in (McCrae et al., 2016). A major result of those activities is the development of the OntoLex-Lemon model, which is described in more details in Section 3.

We adopted OntoLex-Lemon for the representation of morphological resources, as this model was shown to be able to represent both classical lexicographic description (McCrae et al., 2017) and lex-

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

[2] https://elex.is/. See also (Krek et al., 2018) for a general overview of ELEXIS and (Declerck et al., 2018) for a focused description of the role of LLOD in the context of the project.

[3] https://www.pret-a-llod.eu.

[4] See (Petitpierre and Russell, 1995).

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

[6] See https://okfn.org/.

[7] See for more details https://www.w3.org/community/ontolex/.

Proceedings of Recent Advances in Natural Language Processing, pages 233–238,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_027

ical semantics networks, like WordNet (McCrae et al., 2014), to which we want to link full morphological descriptions.

3 OntoLex-Lemon

The OntoLex-Lemon model was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the description of ontology elements are equipped with an extensive linguistic description.[8] This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialized vocabularies.

The main organizing unit for those linguistic descriptions is the lexical entry, which enables the representation of morphological patterns for each entry (a MWE, a word or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the `denotes` property or is mediated by the `LexicalSense` or the `LexicalConcept` properties, as this is represented in Figure 1, which displays the core module of the model.

OntoLex-Lemon builds on and extends the *lemon* model (McCrae et al., 2012b). A major difference is that OntoLex-Lemon includes an explicit way to encode conceptual hierarchies, using the SKOS standard.[9] 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 is paralleling 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 recent developments of the model have been described in (McCrae et al., 2017). Currently two extension modules are being discussed: a lexicographic and a morphology module.[10] Our work can also be seen as preparing the field for a detailed representation of morphological components of lexical data by first porting morphological resources to the core module of OntoLex-Lemon, displayed in 1, before applying the representation guidelines of the morphology extension module, which is not yet in a stable and final state.

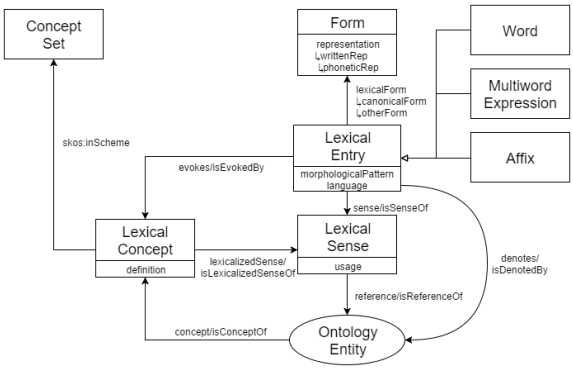

Figure 1: The core module of OntoLex-Lemon: Ontology Lexicon Interface. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

4 The Morphological Resources

We considered two types of morphological data sets. One is an updated version of the multilingual MMorph resource (Petitpierre and Russell, 1995), covering 5 languages. And we also mapped two monolingual data sets, one for German and one for Italian. We will use those additional data sets for the comparison, cross-checking and merging of monolingual morphological resources, using the uniform representation of the data in OntoLex-Lemon.

4.1 MMorph

MMorph was originally developed by ISSCO at University of Geneva in the past MULTEXT project.[11] For our purposes, we used the extended MMorph version developed at DFKI LT Lab (*MMorph3*). This version includes huge morphological resources for English, French, German, Italian and Spanish.

We choose this resource as it provides already in its original format a largely unified representation of the morphological data in the different lan-

[8] See (McCrae et al., 2012a), (Cimiano et al., 2016) and also `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`.

[9] 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/)

[10] See respectively `https://www.w3.org/community/ontolex/wiki/Lexicography` and `https://www.w3.org/community/ontolex/wiki/Morphology`.

[11] See `https://www.issco.unige.ch/en/research/projects/MULTEXT.html` for more details on the resulting `MMorph 2.3.4` version.

guages, with only few differences across the distinct sources.

Very generally, the MMorph tool relates a word to a morphosyntactic description (MSD) containing free-definable attribute and values. The MMorph lexicon used to realize such MSD consists of a set of lexical entries and structural rules. For example, the following rule creates in English a noun plural concatenating the singular form and the noun suffix "s" (Petitpierre and Russell, 1995):

```
"s" noun_suffix[number=plur]

noun[number=plur gender=$gen]
    <- noun[number=sing gender=$gen]
       noun_suffix[number=plur]
```

Note how the rule ensures that the gender does not change in the plural form. Further *adjustment rules* are defined to catch the orthographic features of a specific language (e.g. *box+s = boxes* in English).

The MMorph lexica can be dumped to full form lists for the usage in further programs:

Listing 1: The MMorph entry for the German noun "Aachener" (*inhabitant of Aachen*)

```
"aachener" = "aachener" Noun[gender=masc
    number=singular case=nom|dat|acc]
"aachener" = "aachener" Noun[gender=masc
    number=plural case=nom|gen|acc]
"aachenern" = "aachener" Noun[gender=masc
    number=plural case=dat]
"aacheners" = "aachener" Noun[gender=masc
    number=singular case=gen]
```

As the reader can observe in Listing 1, the nominal entries are completed by appropriate features describing *case*, *gender*, and *number*. Multiple values of a feature are expressed by "|". The user can freely define language- and word class-specific features (e.g. *clitics* for verbal entries or *rection* of prepositions). As the example above demonstrates, the dumped lexica are ideally suited for the mapping into the OntoLex-Lemon format, as they present their data in a well structured fashion.

Our German version of MMorph contains over 2.630.000 full-forms. Compared to the original version, it has specifically improved the coverage of compounds.

To transform the MMorph data into OntoLex-Lemon we used a Python script including the `rdflib` module[12], which supports the generation of RDF-graphs in `rdf:xml`, `turtle` syntax and other relevant formats.

In Listing 2 we show the resulting OntoLex-Lemon representation of the German noun "Aachener".

Listing 2: The OntoLex-Lemon entry for *Aachener*

```
:lex_aachener a ontolex:LexicalEntry ;
   lexinfo:gender lexinfo:masculine ;
   lexinfo:partOfSpeech lexinfo:noun ;
   ontolex:canonicalForm :form_aachener ;
   ontolex:otherForm
      :form_aachener_dat_plural ,
      :form_aachener_gen_singular ,
      :form_aachener_nom-gen-acc_plural .

:form_aachener a ontolex:Form ;
   lexinfo:case lexinfo:accusative ,
      lexinfo:dative ,
      lexinfo:nominative ;
   lexinfo:number lexinfo:singular ;
   ontolex:writtenRep "Aachener"@de .

:form_aachener_dat_plural a
      ontolex:Form ;
   lexinfo:case lexinfo:dative ;
   lexinfo:number lexinfo:plural ;
   ontolex:writtenRep "Aachenern"@de .

:form_aachener_gen_singular a
      ontolex:Form ;
   lexinfo:case lexinfo:genitive ;
   lexinfo:number lexinfo:singular ;
   ontolex:writtenRep "Aacheners"@de .

:form_aachener_nom-gen-acc_plural a
      ontolex:Form ;
   lexinfo:case lexinfo:accusative ,
      lexinfo:genitive ,
      lexinfo:nominative ;
   lexinfo:number lexinfo:plural ;
   ontolex:writtenRep "Aachener"@de .
```

The reader can observe how the relations between a lemma (an instance of the class `LexicalEntry`) and its different morphological forms (instances of the class `Form`) is made explicit by the use of named properties. Another feature of our work is the re-use of established vocabularies, for example the LexInfo vocabulary[13] to represent the morpho-syntactic features.

In Listing 3, we show examples of the resulting data for the lemma "cura" in Spanish.

Listing 3: The OntoLex-Lemon entry for *cura*

```
:lex_cura_1 a ontolex:LexicalEntry ;
    lexinfo:gender lexinfo:feminine ;
    lexinfo:partOfSpeech lexinfo:noun ;
    ontolex:canonicalForm :form_cura ;
    ontolex:otherForm :form_cura_plural .
```

[12] See `https://github.com/RDFLib/rdflib` for more details.

[13] See `https://lexinfo.net/` and (Cimiano et al., 2011) for more details

```
:lex_cura_2 a ontolex:LexicalEntry ;
    lexinfo:gender lexinfo:masculine ;
    lexinfo:partOfSpeech lexinfo:noun ;
    ontolex:canonicalForm :form_cura ;
    ontolex:otherForm :form_cura_plural .

:form_cura a ontolex:Form ;
    lexinfo:number lexinfo:singular ;
    ontolex:writtenRep "cura"@es .

:form_cura_plural a ontolex:Form ;
    lexinfo:number lexinfo:plural ;
    ontolex:writtenRep "curas"@es .
```

As the reader can observe, we have two lexical entries for the entry "cura", as this is suggested by the lexicographic module of Ontolex-Lemon.[14] "Cura" in feminine means *cure* or *healing*, while in masculine it refers to a *cure*. But one can also propose a unique entry for "cura" and add in each of the associated senses a usage restriction indicating the gender of the corresponding `ontolex:Form`.

The reader can also see the harmonized representation of morphological resources across languages (here German and Spanish). This is an important feature that will allow to link various lemmas (or senses) from different languages to a unique reference point in external information sources, like WordNet(s)[15] or knowledge Graphs, like DBpedia[16] or Wikidata[17].

The transformation of nominal entries from MMorph to the OntoLex-Lemon format resulted in 67778 instances of the class `LexicalEntry` for German, 17313 for Spanish, 21085 for Italian, 29959 for English and 13525 for French. The English nominal data in OntoLex-Lemon include 59108 instances of the class `Form`, while the German data consists of 224449 such forms. This largely depends on the maintenance state of the original resources, but gives nevertheless a good idea on the difference of morphological variations in the distinct languages.

[14] See the discussion on this case at https://www.w3.org/community/ontolex/wiki/Lexicography.

[15] Concerning the linking to WordNets, we started linking the French, Italian and Spanish morphological data to their counterparts in the Open Multilingual WordNet initiative. See http://compling.hss.ntu.edu.sg/omw/ and (Bond and Paik, 2012) for more details

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

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

4.2 Two Monolingual Resources

Other data sets we are considering are "Morph-it!"[18] for Italian and the "DE_morph_dict data"[19].

An entry in Morph-it! has the form displayed in Listing 4:

Listing 4: The Morph-it! entry for "abbassamento" (*lowering* or *reduction*)

abbassamento	abbassamento	NOUN-M: s
abbassamenti	abbassamento	NOUN-M: p

The corresponding OntoLex-Lemon encoding is displayed in Listing 5. While this representation looks much more complex than the original Morph-it! one, it represents the relations in an explicit and declarative way and at the same time it gives a full "autonomy" to the form variants, which are now represented as instances of the class `ontolex:Form` and equipped with an URI, so that they can be accessed independently of their corresponding headword.

Listing 5: The OntoLex-Lemon representation for "abbassamento" (*lowering* or *reduction*)

```
:lex_abbassamento a
        ontolex:LexicalEntry;
    lexinfo:gender lexinfo:masculine ;
    lexinfo:partOfSpeech lexinfo:noun ;
    ontolex:canonicalForm
:form_abbassamento ;
    ontolex:otherForm
      :form_abbassamento_m_p .

:form_abbassamento a ontolex:Form ;
    lexinfo:number lexinfo:singular ;
    ontolex:writtenRep "abbassamento"@it.

:form_abbassamento_m_p a ontolex:Form ;
    lexinfo:number lexinfo:plural ;
    ontolex:writtenRep "abbassamenti"@it.
```

In Listing 6 below, we display an original entry of the DE_morph_dict resource, in this example the word "Abgang" (*departure*, *leaving*, *dispatch*, etc.). The reader can immediately see the difference to the Italian entry in Listing 4, as in German there are four cases and three genders, a fact which leads to a high number of morphological form variants. Also this entry is including obsolete forms of the word, which is adding an additional line in the original encoding.

[18] https://docs.sslmit.unibo.it/doku.php?id=resources:morph-it. See also (Zanchetta and Baroni, 2005)).

[19] https://github.com/DuyguA/german-morph-dictionaries. See also for this resource the companion morphological analyser in (Altinok, 2018)).

Listing 6: The DE_morph_dict entry for "Abgang" (*departure* or *leaving* etc)

```
Abgang
Abgang NN,masc,acc,sing
Abgang NN,masc,nom,sing
Abgang NN,masc,dat,sing
Abgange
Abgang NN,masc,dat, ing,old
Abganges
Abgang NN,masc,gen,sing
Abgangs
Abgang NN,masc,gen,sing
Abgänge
Abgang NN,masc,nom,plu
Abgang NN,masc,acc,plu
Abgang NN,masc,gen,plu
Abgängen
Abgang NN,masc,dat,plu
```

The mapping of this entry to OntoLex-Lemon results in a representation that is by now familiar, and which is given in Listing 7.

Listing 7: The OntoLex-Lemon representation for "Abgang" (*departure* or *leaving* etc)

```
:lex_abgang a ontolex:LexicalEntry ;
    lexinfo:gender lexinfo:masculine ;
    lexinfo:partOfSpeech lexinfo:noun ;
    ontolex:canonicalForm :form_abgang ;
    ontolex:otherForm
        :form_abgang_dat_plu,
        :form_abgang_dat_sing,
        :form_abgang_gen_sing,
        :form_abgang_nom-gen-acc_plu .

:form_abgang a ontolex:Form ;
    lexinfo:case lexinfo:accusative,
        lexinfo:dative,
        lexinfo:nominative ;
    lexinfo:number lexinfo:singular ;
    ontolex:writtenRep "Abgang"@de .

:form_abgang_dat_plu a ontolex:Form ;
    lexinfo:case lexinfo:dative ;
    lexinfo:number lexinfo:plural ;
    ontolex:writtenRep "Abgängen"@de .

(etc.)
```

With those resources represented in OntoLex-Lemon, which are duplicating the German and Italian resources we have already from MMorph, we aim at discovering possible inconsistencies or similarities within resources for one language, which could lead to both a improvement and a merging of the original resources.

We are in a sense extending a former experiment on automatically merging Italian morphological resources in the context of a finite automata environment, and which is described in (Declerck et al., 2012). The new work is not only a multilingual extension, but is aiming at a broad interoperability of morphological resources by using a defacto standard developed by a W3C Community Group and publishing the results in an accessible subset of the Linked Data cloud.

5 Conclusion

We described our current work consisting in porting a number of (multilingual) morphological resources to OntoLex-Lemon, in order to harmonize those and to support their interlinking, cross-checking, but also their linking with other data source in the Linguistic Linked Open Data, as for examples WordNets, or with data sets included in knowledge graphs, like DBpedia or Wikidata.

As a final goal of our work, we see the possibility to interlink or merge those morphological resources in the Linguistic Linked Open Data cloud.

Acknowledgments

Work presented in this paper has been supported in part by the H2020 project "Prêt-à-LLOD" with Grant Agreement number 825182. Contributions by Thierry Declerck are also supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015.

References

Duygu Altinok. 2018. Demorphy, german language morphological analyzer. *CoRR*, abs/1803.00902.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71, Matsue.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors. 2012. *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata*. Springer.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. Lexinfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report.

Thierry Declerck, John McCrae, Roberto Navigli, Ksenia Zaytseva, and Tanja Wissik. 2018. Elexis - european lexicographic infrastructure: Contributions to and from the linguistic linked open data. In *Proceedings of the 2nd GLOBALEX Workshop. GLOBALEX (GLOBALEX-2018), Lexicography & WordNet, located at 11th Language Resources and Evaluation Conference (LREC 2018), May 8, Miyazaki, Japan*, pages 17–22. ELRA.

Thierry Declerck, Stefania Racioppa, and Karlheinz Mörth. 2012. Automatized merging of italian lexical resources. In *Proceeding of the LREC 2012 Workshop on Language Resource Merging*, Paris. ELRA, ELRA.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Dagmar Gromann and Thierry Declerck. 2019. Towards the detection and formal representation of semantic shifts in inflectional morphology. In *2nd Conference on Language, Data and Knowledge (LDK)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 21:1–21:15. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Simon Krek, Iztok Kosem, John P. McCrae, Roberto Navigli, Bolette S. Pedersen, Carole Tiberius, and Tanja Wissik. 2018. European lexicographic infrastructure (elexis). In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 881–891, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asuncion Gomez-Perez, Jorge Garcia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012a. Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719.

John McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, D ennis Spohr, and Tobias Wunner. 2012b. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–709.

John P. McCrae, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex 2017*, pages 587–597. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o.

John P. McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora, and Jonathan Pool. 2016. The open linguistics working group: Developing the linguistic linked open data cloud. In *The 10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016, Slovenia, Portorož*.

John P. McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3 rd Workshop on Linked Data in Linguistics*.

Dominique Petitpierre and Graham. Russell. 1995. MMORPH: The Multext morphology program. Multext deliverable 2.3.1, ISSCO, University of Geneva.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the italian language. In *Proceedings of Corpus Linguistics 2005*. University of Birmingham.