Enriching Open Multilingual Wordnets with Morphological Features*

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

Thierry Declerck
German Research Center
for Artificial Intelligence
Saarbrücken, Germany
&
Austrian Centre for Digital Humanities
Austrian Academy of Sciences
Vienna, Austria
thierry.declerck@dfki.de

Abstract

English. In this article, we describe our work on porting Open Multilingual Wordnet resources into the OntoLex-Lemon model, in order to establish an interlinking with corresponding morphological resources, such as the MMorph resource set. For this purpose, the morphological resources were also ported onto OntoLex-Lemon. We show how the "lemmas" contained in the Wordnet resources can be enriched with morphological features using the lexical representation and linking features of OntoLex-Lemon, which support, among others, the formulation of restrictions in the usage of such expressions. Our work will result in an improved lexical resource combining Wordnet senses and full morphological descriptions in a single ontological framework, as specified in the OntoLex-Lemon model.

1 Introduction

WordNets are well-established lexical resources with a wide range of applications. For more than twenty years they have been elaborately set up and maintained by hand, especially the original Princeton WordNet of English (PWN) (Fellbaum, 1998). In recent years, there have been increasing activities in which open WordNets for different languages have been automatically extracted from various resources and enriched with lexical semantics information, building the so-called Open Multilingual Wordnet (OMW) (Bond and Paik, 2012). These WordNets were linked to PWN via shared synset IDs (Bond and Foster, 2013; Bond et al., 2016). The resources in OMW are of different coverage and contain not always the same amount of information, as for example many resources are lacking definitions (or "glosses"), contrary to the PWN resource, or example sentences.

The work described in the present article is an extension of previous experiments done with English (Gromann and Declerck, 2019) and more recently with German lexical semantics resource, as we wanted to consider languages with a complex morphology.[1] In the present article we focus on Romance languages, especially Italian.

Our current work deals primarily with the morphological enrichment of OMW resources for Italian, i.e. "ItalWordNet".[2] The first morphological resource we took into consideration for this purpose is an updated version of the MMorph morphological analyser (Petitpierre and Russell, 1995).

As a representation mean we chose OntoLex-Lemon (Cimiano et al., 2016)[3], as this model has proven to be able to represent both a classical lexicographic description (McCrae et al., 2017) as well as lexical semantics networks like WordNet (McCrae et al., 2014).

OntoLex-Lemon is a further development of the "Lexicon Model for Ontologies" (*lemon*) (McCrae et al., 2012). Following the Guidelines[4] for mapping Global WordNet formats onto *lemon*-based RDF[5], some WordNets have already been

[1] This work will be published soon in the proceedings of the Global Wordnet Conference 2019.

[2] See (Pianta et al., 2002; Toral et al., 2010). But we also made similar experiments with French and Spanish.

[3] See also https://www.w3.org/2016/05/ontolex/ for more details.

[4] See https://globalwordnet.github.io/schemas/##rdf.

[5] RDF stands for "Resource Description Framework". See

mapped onto the former *lemon* model (McCrae et al., 2014). Our present goal is thus to integrate conceptual descriptions, lemmas and morphological descriptions in the extended ontological framework specified by the OntoLex-Lemon model.[6]

In the next sections, we give some background information on OMW and MMorph. We continue with a section on OntoLex-Lemon, followed by sections that describe how OntoLex-Lemon supports the linking of lemmas in the OMW resources to full morphological descriptions. Doing so, morphological descriptions can be associated with the conceptual entries of WordNet.

2 Open Multilingual WordNet

OMW is an initiative that brings together Wordnets in different languages, linking them to the original Princeton WordNet (PWN). As stated on the web page of OMW, those Wordnets were of different quality, and some of those were in fact extracted from different types of language resources. We are dealing with three OMW WordNet resources.[7] OMW provided for an harmonization of such resources, and published them in a uniform format, which is displayed just below, showing here a few examples from the Italian resource:

```
08388207-n ita:lemma nobiltà
08388207-n ita:lemma aristocrazia
08388207-n ita:lemma patriziato
08388207-n ita:def_0
    l'insieme degli aristocratici
08388207-n ita:def_1
    l'insieme dei nobili
...

14842992-n ita:lemma terra
14842992-n ita:lemma terreno
14842992-n ita:lemma suolo
14842992-n ita:def_0 parte
    superficiale della crosta
    terrestre sulla quale si
    sta o si cammina
14842992-n ita:exe_0 si piegò
    con fatica per raccogliere da
    terra i sacchetti, pronta a
    salire sull'autobus
14842992-n ita:exe_1 l'uomo
    cominciò a rotolarsi per terra
    in preda a dolori lancinanti
```

As the reader can see in the two examples above, OMW resources deliver information on the synset number, together with the part-of-speech of the associated lemma. In some cases, definitions (marked with `ita:def`) are provided, as well as examples (marked with `ita:exe`).

This format is used for all languages of the OMW corpus. This eases its mapping to a formal representation supporting the interoperability and interlinking of language resources, such as the OntoLex-Lemon model (see Section 4).

3 MMorph

MMorph was originally developed by ISSCO at the University of Geneva in the past MULTEXT project[8]. For our purposes, we used the extended MMorph version developed at DFKI LT Lab (*MMorph3*). This version includes huge lexical resources for English, French, German, Italian and Spanish. Very generally, the tool relates a word to a morphosyntactic description (MSD) containing free-definable attribute and values. The MMorph lexicon which is used to realize such MSD consists of a set of lexical entries and structural rules.[9] For example, the following rule creates in Italian a noun plural concatenating the noun stem and the gender-specific suffixes:

Listing 1: Rule for noun plural generation in Italian. Note how the rule ensures that the gender doesn't change in the plural form.

```
N.ms: "o" NSuffix[num=sing gen=masc
    type=oa]
N.mp: "i" NSuffix[num=plur gen=masc
    type=oa]
N.fs: "a" NSuffix[num=sing gen=fem
    type=oa]
N.fp: "e" NSuffix[num=plur gen=fem
    type=oa]

FlexN: Noun[gen=$1 num=$2 form=surf]
    <- Noun[gen=$1 num=sing
            form=stem type=$T]
       N_ASfix[gen=$1 num=$2
               type=$T]
```

This rule will apply only to the lexical entries (feminine and/or masculine nouns) matching the defined features, e.g.

```
Noun[gen=masc num=sing form=stem
     type=oa]
"patriziat" = "patriziato"
"suol" = "suolo"
```

https://www.w3.org/RDF/ for more details.

[6] OntoLex-Lemon is indeed representing an ontology of lexical elements.

[7] French, Spanish and Italian, with a focus on the latter. See http://compling.hss.ntu.edu.sg/omw/ for downloading the resources. For more details see also (Bond and Paik, 2012).

[8] See https://www.issco.unige.ch/en/research/projects/MULTEXT.html for more details on the resulting MMorph2.3.4 version.

[9] See (Petitpierre and Russell, 1995)

The morphology is completed by a set of spelling rules to catch the orthographic peculiarities of a specific language (e.g. `fung + i = funghi` in Italian).

The MMorph lexica can be dumped to full form lists for the usage in further programs, as can be seen in the following examples:

```
"nobiltà" = "nobiltà"
    Noun[ gen=fem num=sing|plur ]
"suoli" = "suolo"
    Noun[ gen=masc num=plur ]
"suolo" = "suolo"
    Noun[ gen=masc num=sing ]
```

The entries above are completed by labelled features for gender and number, but the user can freely define further features, if needed (e.g. *clitics* for verbal entries or *rection* of prepositions). Multiple values of a feature are expressed by "|".

Because of their well-structured form, the dumped Mmorph lexica are ideally suited for the mapping into the OntoLex-Lemon format.

4 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.[10] This rich linguistic grounding includes the representation of morphological and syntactical properties of a lexical entry 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 represented in Figure 1, which displays the core module of the model.

OntoLex-Lemon is based on and extends the *lemon* model (McCrae et al., 2012). A major difference is that OntoLex-Lemon includes an explicit way to encode conceptual hierarchies, using the SKOS standard.[11] As shown

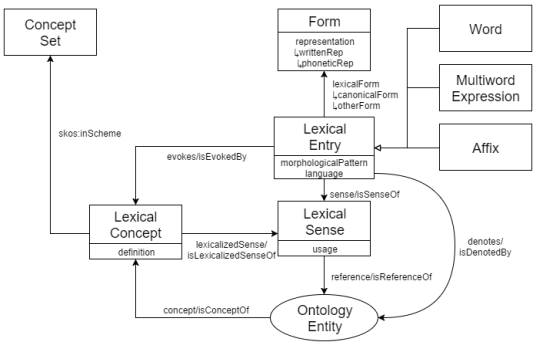

Figure 1: The core module of OntoLex-Lemon: Ontology Lexicon Interface. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

in Figure 1, lexical entries can be linked, via the `ontolex:evokes` property, to such SKOS concepts, which can represent WordNet synsets. This structure parallels 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.[12] The `ontolex:LexicalConcept` class seems to be most appropriate to model the "sets of cognitive synonyms (synsets)"[13] described by Princeton WordNet (PWN), while the `ontolex:LexicalSense` class is meant to represent the bridge between lexical and ontological entities.

5 Mapping the OMW Resources to OntoLex-Lemon

As mentioned above, the format generated by the OMW initiative is very convenient to map dif-

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

[11] 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/`)

[12] Quoting from Section 3.6 "Lexical Concept" `https://www.w3.org/2016/05/ontolex/`: "We [...] capture the fact that a certain lexical entry can be used to denote a certain ontological predicate. We capture this by saying that the lexical entry denotes the class or ontology element in question. However, sometimes we would like to express the fact that a certain lexical entry evokes a certain mental concept rather than that it refers to a class with a formal interpretation in some model. Thus, in lemon we introduce the class Lexical Concept that represents a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses. A lexical concept is thus a subclass of skos:Concept."

[13] Quoted from `https://wordnet.princeton.edu/`.

ferent information onto more complex representation frameworks. To transform the OWN data onto the OntoLex-Lemon representation, a Python script was used. A design decision was to extract only the synset information and to encode the synsets as instances of the `LexicalConcept` class of OntoLex-Lemon. As some OWM lemmas are present in the MMorph resources, we just link the synsets to those lemmas, which are encoded as instances of the OntoLex-Lemon `LexicalEntry` class (see next section). We will need to create new instances of the OntoLex-Lemon `LexicalEntry` class for the OWM lemmas not present in the MMorph resources.

We have now 15553 such `LexicalConcept` instances for Italian. This is due to the fact that we consider only the subset of ItalWordNet that has been curated by OMW. We also noted that we have less instances of the `LexicalConcept` as lines for each synset in the original files, as the synset indices are represented by a unique URI in OntoLex-Lemon.

In Listing 2 we show examples of the OntoLex-Lemon encoding of two synsets for Spanish.[14] The lemmas associated with these synsets are "cura". In Section 7, we explain how the synsets are linked to the lemmas, which are differentiated in the OntoLex-Lemon representation, but not in the original OMW file.

Listing 2: The OntoLex-Lemon representation of two Spanish synsets

```
:synset_spawn-13491616-n
  rdf:type ontolex:LexicalConcept ;
  skos:inScheme :spawnet .

:synset_spawn-10470779-n
  rdf:type ontolex:LexicalConcept ;
  skos:inScheme :spawnet .
```

6 Mapping MMorph to Ontolex-Lemon

To transform the MMorph data into OntoLex-Lemon we used a Python script including the `rdflib` module[15], which supports the generation of RDF-graphs in `rdf : xml`, `turtle`, or other relevant formats. In Listing 3, we show examples of the resulting data for the lemma "viola" in Italian.

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

```
:lex_viola_fem a ontolex:LexicalEntry ;
    lexinfo:partOfSpeech lexinfo:noun ;
    ontolex:canonicalForm :form_viola_f ;
    ontolex:otherForm :form_viola_f_pl .

:lex_viola_masc a ontolex:LexicalEntry ;
    lexinfo:partOfSpeech lexinfo:noun ;
    ontolex:canonicalForm :form_viola_m ;

:form_viola_f a ontolex:Form ;
    lexinfo:gender lexinfo:feminine ;
    lexinfo:number lexinfo:singular ;
    ontolex:writtenRep "viola"@it .

:form_viola_f_pl a ontolex:Form ;
    lexinfo:gender lexinfo:feminine ;
    lexinfo:number lexinfo:plural ;
    ontolex:writtenRep "viole"@it .

:form_viola_m a ontolex:Form ;
    lexinfo:gender lexinfo:masculine ;
    lexinfo:number lexinfo:plural ,
        lexinfo:singular ;
    ontolex:writtenRep "viola"@it .
```

As the reader can observe, we have two lexical entries for the entry "viola", as this is requested by the OntoLex-Lemon guidelines, following which a word with different grammatical genders should have one lexical entry per gender. "Viola" in feminine is the music instrument, while in masculine it means "violet". This is in fact an important feature for linking synsets to lemmas having distinct genders, as we will exemplify in Section 7.

The transformation of nominal entries from MMorph to the OntoLex-Lemon format resulted in 21085 instances of the class `LexicalEntry` for Italian. We still need to consider the lemmas of the OMW resources that are not in MMorph. This is concerning mostly multiword entries in OMW.

We will also investigate the use of other lexical resources, but the current use of the MMorph was motivated by the fact that we could have access to the different languages available in one and the same format, facilitating thus the uniform mapping into OntoLex-Lemon.

7 Linking the OMW Resources to the MMorph Resources

We see the use of OntoLex-Lemon for representing WordNets not only as a chance to port information from one format to another (including the possibility to publish WordNets in the Linguistic Linked Opend Data cloud[16]), but also as an opportunity to extend the coverage of WordNet descrip-

[14] For the representation of OntoLex-Lemon data, we chose the `turtle` syntax serialization. More on the turtle syntax: `https://www.w3.org/TR/turtle/`.

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

[16] See `http://linguistic-lod.org/llod-cloud` and (Chiarcos et al., 2012)

tions to more complex lexical phenomena, beyond lemma and PoS considerations. One case that has been studied in the recent past concerns the meaning that can be specifically associated to English plurals listed in PWN (Gromann and Declerck, 2019). We are interested in applying a similar approach to grammatical gender: we could link a Wordnet synset to a specific gender, as this information is normally not included in the Wordnets, which consider only the part-of-speech of the associated lemmas.

OntoLex-Lemon supports this linking in a straightforward manner. As can be seen in Figure 1, there is a property putting a `LexicalConcept` in relation to a `LexicalEntry`, i.e. the property `evokes` and its reverse `isEvokedBy`. Therefore we just need to add this property to both the OntoLex-Lemon representations of a synset and its corresponding entry. In Listing 4 we show such a case, taking again the word "cura" as an example.

Listing 4: Interlinking a synset and an entry for *cura*

```
:synset_spawn-13491616-n
  rdf:type ontolex:LexicalConcept ;
  skos:inScheme :spawnet ;
  ontolex:evokes :lex_cura_1 .

:lex_cura_1 a ontolex:LexicalEntry ;
  lexinfo:gender lexinfo:fem ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :form_cura ;
  ontolex:otherForm :form_cura_plural ;
  ontolex:isEvokatedBy
          :synset_spawn-1349161-n .

:synset_spawn-10470779-n
  rdf:type ontolex:LexicalConcept ;
  skos:inScheme :spawnet ;
  ontolex:evokes :lex_cura_2 .

:lex_cura_2 a ontolex:LexicalEntry ;
  lexinfo:gender lexinfo:mas;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :form_cura ;
  ontolex:otherForm :form_cura_plural ;
  ontolex:isEvokatedBy
          :synset_spawn-10470779-n .
```

Just adding the properties `evokes` and its reverse `isEvokedBy` to the corresponding elements in the generated OntoLex-Lemons data sets is providing for this morphological enrichment of the original Wordnets. Once the original (different types of) resources have been mapped onto the OntoLex-Lemon model, it is very easy to interlink or even to merge them into a richer representation. An extension of this work consists in describing restrictions on the usage of certain Wordnet concepts, as for example in the Italian case of the noun "bene" versus its plural form "beni", or English "silk" versus the plural form "silks", which are associated with different and sometimes not shareable meanings.[17] We are making use for this of a strategy described in an extension to the core module of Ontolex-Lemon, called "Lexicog",[18] which foresees the description of instances of a class named `FormRestriction`, so that it is possible to state that a meaning is available only with the use of a specific form, like singular or plural.

8 Conclusion

We described our work on porting Open Multilingual Wordnet resources into the OntoLex-Lemon model, in order to establish an interlinking with corresponding morphological resources, such as the MMorph resource set. For this purpose, the morphological resources were also ported onto OntoLex-Lemon. As a result we noticed that this model can be easily used for bridging the WordNet type of lexical resources to a full description of lexical entries, which coult possibly lead to an extension of the coverage of WordNets beyond the consideration of lemmas and PoS information.

We documented our interlinking work with the example of the full morphological representation of Italian words, putting them in relation with the corresponding OMW data sets. We also started to investigate the description of usage restrictions, which allows us to state formally that certain Wordnet concepts should be used only in the singular or in the plural form.

As a final goal of our work, we see the interlinked or merged resources in the Linguistic Linked Open Data (LLOD) cloud. We will investigate how our work can be combined with resources present in the LLOD, especially with the BabelNet framework, which is already integrating a huge number of lexical resources, including Princeton WordNet, and encyclopedic data sets (Ehrmann et al., 2014).

[17] The reader can see the different meanings associated to those plural words while querying for those in the user interface of PWN: `http://wordnetweb.princeton.edu/perl/webwn`.

[18] The current state of this "Lexicography" module is available at `https://www.w3.org/community/ontolex/wiki/Lexicography`.

Acknowledgments

The presented work has been supported in part by the H2020 project "Prêt-à-LLOD" with Grant Agreement number 825182. Contributions by Thierry Declerck have been supported additionally in part and by the H2020 project "ELEXIS" with Grant Agreement number 731015.

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362, Sofia.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. *Small*, 8(4):5.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*, volume 2016.

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

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

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano, and Roberto Navigli. 2014. Representing multilingual data as linked data: the case of BabelNet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 401–408, Reykjavik, Iceland, May. European Languages Resources Association (ELRA).

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

Dagmar Gromann and Thierry Declerck. 2019. Towards the detection and formal representation of semantic shifts in inflectional morphology. In Maria Eskevich, Gerard de Melo, Christian Fth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski, editors, *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, 5.

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. 2012. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–709.

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*.

John P. McCrae, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In Iztok Kosem, Jelena Kallas, Carole Tiberius, Simon Krek, Miloš Jakubíček, and Vít Baisa, editors, *Proceedings of eLex 2017*, pages 587–597. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., 9.

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

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *In Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India.

Antonio Toral, Stefania Bracale, Monica Monachini, and Claudia Soria. 2010. Rejuvenating the italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*, Mumbai.