Porting a Crowd-Sourced German Lexical Semantics Resource to Ontolex-Lemon

Thierry Declerck[1,2], Melanie Siegel[3]

[1] German Research Center for Artificial Intelligence, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany

[2] Austrian Centre for Digital Humanities, Sonnenfelsgasse 19, 1010 Vienna, Austria

[3] Darmstadt University of Applied Science, Max-Planck-Str. 2, 64807 Dieburg, Germany

E-mail: declerck@dfki.de, melanie.siegel@h-da.de

Abstract

In this paper we present our work consisting of mapping the recently created open source German lexical semantics resource "Open-de-WordNet" (OdeNet) into the OntoLex-Lemon format. OdeNet was originally created in order to be integrated in the Open Multilingual Wordnet initiative. One motivation for porting OdeNet to OntoLex-Lemon is to publish in the Linguistic Linked Open Data cloud this new WordNet-compliant resource for German. At the same time we can with the help of OntoLex-Lemon link the lemmas of OdeNet to full lexical descriptions and so extend the linguistic coverage of this new WordNet resource, as we did for French, Italian and Spanish wordnets included in the Open Multilingual Wordnet collection. As a side effect, the porting of OdeNet to OntoLex-Lemon helped in discovering some issues in the original data.

Keywords: Open Multilingual Wordnet; OntoLex-Lemon; OdeNet; Lexical Semantics

1. Introduction

Wordnets are well-established lexical resources with a wide range of applications in various Natural Language Processing (NLP) fields, like Machine Translation, Information Retrieval, Query Expansion, Document Classification, etc. (Morato et al., 2004). 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 other resources and enriched with lexical semantics information, building the so-called Open Multilingual Wordnet (OMW) (Bond & Paik, 2012), which is merging more than 35 open wordnets that are linked through the Collaborative Interlingual Index (CILI) (Bond & Foster, 2013; Bond et al., 2016). The resources in OMW are of different coverage and do not always contain the same amount of information, as for example many resources are lacking definitions (or "glosses"), contrary to the PWN resource, or example sentences.

Recently we made some experiments to enrich OMW resources with morphological resources. The resources we were dealing with are "WOLF (Wordnet Libre du Français)" for French, "ItalWordNet" for Italian and "Multilingual Central Repository" for

Spanish (this resource also contains wordnets for the Catalan, Basque and Galician languages).[1] In order to link those OWM resources to full lexical and morphological descriptions we first map them onto the OntoLex-Lemon model (Cimiano et al., 2016), which is a de facto standard for the representation of lexical data in the Web (McCrae et al., 2017), especially in the Linguistic Linked Open Data cloud.[2]

Up until very recently no German resources were included in the OMW collection, which requires the data to be equipped with an open and free licence. This condition is probably the reason why GermaNet is not included in OMW. GermaNet is a manually well-designed WordNet resource for German (Hamp & Feldweg, 1997).[3] But GermaNet is not equipped with the type of license required by OMW.

In this context, a new German lexical semantics resource with the name "Open German WordNet" (OdeNet)[4] has been developed with the aim to be included as the first open German WordNet into the Open Multilingual Wordnet.[5]

This paper is organised as follows. In Section 2 we present the Ontolex-Lemon model. In Section 3 we give some more details on the OMW resources we mapped to OntoLex-Lemon in order to link them to corresponding morphological resources. The result of this mapping is shown in Section 4. The OdeNet resource is described in some detail in Section 5. We describe in Section 6 the current state of the representation of OdeNet data in OntoLex-Lemon, and the issues in the original data we discovered through this mapping exercise.

2. 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 ontological elements are equipped with an extensive linguistic description.[6]

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

[1] See Sagot and Fišer (2008), Pianta et al. (2002), Toral et al. (2010) and Gonzalez-Agirre et al. (2012), respectively.

[2] See http://linguistic-lod.org/ and also Chiarcos et al. (2012).

[3] See also http://www.sfs.uni-tuebingen.de/GermaNet/ for more details.

[4] See https://github.com/hdaSprachtechnologie/odenet for more details.

[5] See http://compling.hss.ntu.edu.sg/omw20/omw_wns for more details.

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

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.

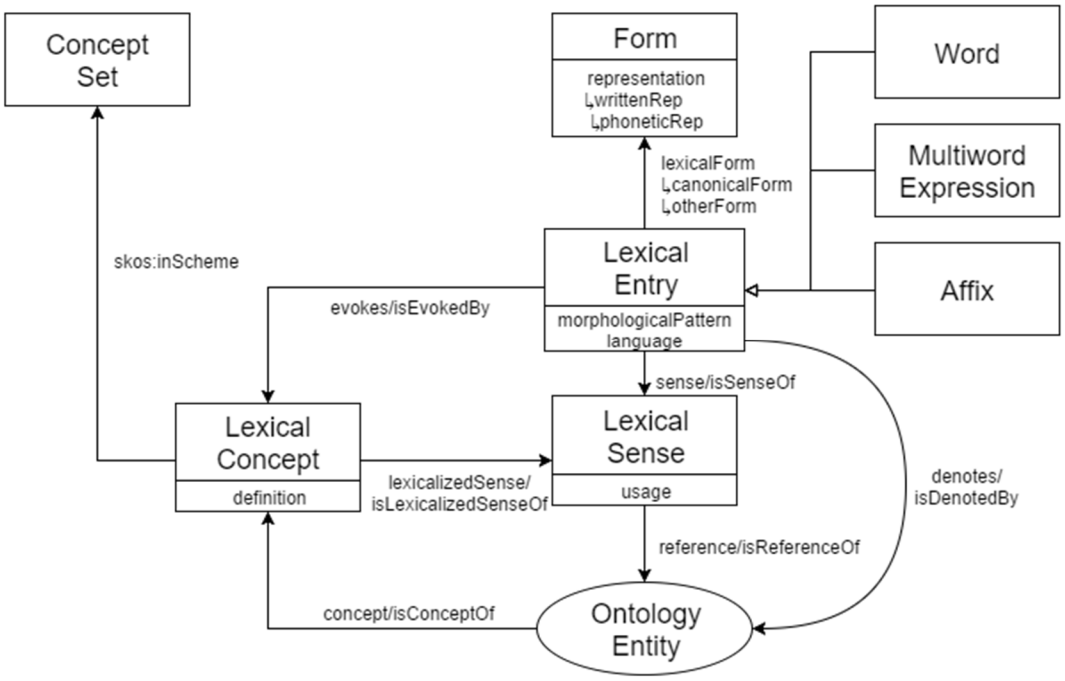

Figure 1: The core module of OntoLex-Lemon: Ontology Lexicon Interface. Graphic taken from https://www.w3.org/2016/05/ontolex/.

OntoLex-Lemon builds 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.[7] 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.[8] The "sets of

[7] 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/).

[8] 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."

cognitive synonyms (synsets)"[9], that Princeton WordNet (PWN) describes, seems to be best modelled by the ontolex:LexicalConcept class, while the ontolex:LexicalSense class is meant to represent the bridge between lexical entries and ontological entities (which do not necessarily have semantic relations between them).

3. Open Multilingual WordNet

The three Open Multilingual Wordnet resources (for French, Italian and Spanish) we were dealing with are available at the Open Multilingual Wordnet (OMW) page.[10] OMW is an initiative that brings together wordnets in different languages, which are linked through the Collaborative Interlingual Index (CILI). As stated on the web page of OMW, those wordnets are of different quality, and some of those were in fact extracted from different types of language resources. OMW provided for some corrections and for an harmonization of such resources, and published them in a uniform tabular format, which is displayed below, exemplified here by entries from the Italian OMW 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 il tizio comincio' a rotolarsi
        per terra in preda a dolori lancinanti
```

In the two examples displayed above, the uniform tabular format of OMW delivers information on the synset IDs (08388207-n and 14842992-n), which include the part-of-speech ("n") of the associated lemma(s). The nominal lemmas associated with the synset-ID 08388207-n are "nobiltà" (nobility), "aristocrazia" (nobility, aristocracy) and "patriziato" (aristocracy). The nominal lemmas associated with the synset-ID

[9] Quoted from https://wordnet.princeton.edu/.

[10] See http://compling.hss.ntu.edu.sg/omw/. For more details see also Bond and Paik (2012).

14842992-n are “terra” (earth, land, soil), “terreno” (ground, terrain, soil) and “suolo” (land, earth, ground). If available, definitions (“glosses") are provided (marked with the feature “ita:def”), as well as examples (marked with the feature “ita:exe”).[11]

This tabular format is used for all the OMW data sets. This makes it easier to map OMW data to a formal representation that supports the interoperability and interlinking of language resources. The next section shows the result of the mapping of OMW resources to OntoLex-Lemon.

4. Mapping the OMW Resources to OntoLex-Lemon

As mentioned earlier, the format generated by the OMW imitative is very convenient with regard to mapping onto more complex representation frameworks. A Python script was implemented for porting the OMW data sets to OntoLex-Lemon.

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 we expect to have the lemmas present in already existing lexicons, we will just link the synsets to those lemmas, which are encoded as instances of the OntoLex-Lemon LexicalEntry class. This way we achieve a higher level of modularity. Since the synsets are now encoded as instances of the LexicalConcept class, each synset-ID gets a Unique Resource Identifier (URI), and does not have to be repeated for each lemma it is associated with, but can just link to those via the OntoLex-Lemon property isEvokedBy, as seen in Figure 1. This way we have also a more compact (graph-based) representation as in the original representation of the OMW data.

We have now 38,512 such instances of LexicalConcept for Spanish, 15,553 for Italian, and 59,091 for French.[12]

In Listing 1.1 we show examples of the OntoLex-Lemon encoding of two synsets for Spanish. The lemmas associated with these synsets are “cura”. In Section 2, we explain how in OntoLex-Lemon the synsets are linked to the lemmas, which are differentiated in the OntoLex-Lemon representation,[13] which we add here, but not in the original OMW file, as in OMW the lemmas are just literals and not real lexical entries, associated with more complex linguistic information, additionally to PoS.

---

[11] We observe that using this type of text format for representing the data, one has to repeat the relevant information (for example the synset-ID) for each line introducing a lemma associated with the synset.

[12] The lower number for the Italian resource is due to the fact that we consider only the subset of ItalWordNet that has been curated by OMW.

[13] Depending on the view on the word “cura” (meaning *cure* or *priest*, if the gender of the word is feminine or masculine) we can have either one lexical entry or two. Taking into consideration the distinct genders and etymologies for “cure”, we decided to have two entries.

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

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

: lex_cura -13491616-n a ontolex : LexicalEntry ;
   lexinfo : gender lexinfo : masc ;
   lexinfo : partOfSpeech lexinfo : noun ;
   ontolex : evokes : synset_spawn-13491616-n ;
   ontolex : canonicalForm : form_cura ;
   ontolex : otherForm : form_cura_plural .

: lex_cura -10470779-n a ontolex : LexicalEntry ;
   lexinfo : gender lexinfo : fem ;
   lexinfo : partOfSpeech lexinfo : noun ;
   ontolex : evokes : synset_spawn-10470779-n ;
   ontolex : canonicalForm : form_cura ;
   ontolex : otherForm : form_cura_plural .
```

Listing 1.1: The OntoLex-Lemon representation of two Spanish synsets with the corresponding lemmas

Current work is dedicated in enriching the three wordnets encoded in OntoLex-Lemon with further morphological semantic information. For this we already mapped the French, Italian and Spanish morphological resources included in the MMmorph data sets (Petitpierre & Russell, 1995) into OntoLex-Lemon,[14] and we are bridging the two types of data sources.

5. The Open-de-WordNet (OdeNet)

The "Open-de-WordNet" (OdeNet)[15] initiative is intended as a contribution to the Open Multilingual Wordnet Initiative. It is a WordNet for the German language under an

[14] This mapping is described in Declerck and Racioppa (2019).

[15] https://github.com/hdaSprachtechnologie/odenet.

open license (CC BY-SA 4.0). The main source for the synset entries is the OpenThesaurus German synonym lexicon.[16] OpenThesaurus compiled approximately 120,000 entries in a crowd sourcing procedure. OdeNet transferred those data to synsets in the Global WordNet format.[17] Subsequently, the resulting synsets were enriched with part-of-speech (PoS) information, semantic identifiers from OMW were identified and hierarchy relations were added.

As mentioned above, PoS information is associated with the synsets. We observe that only four PoS categories are used: Adjectives, Nouns, Verbs and "p", which seems to be attributed to all synset/lemma combinations not being one of the three other categories. This strategy is not satisfying, and we are working on mapping all the "p" tagged lemmas to existing entries in a German lexicon in order to further specify their PoS. We also observe that phrasal multi-word units are also equipped with one of those PoS tags. In most cases this is sensible and could be accepted, as with "in Rechnung stellen" (*to bill*) or "Abschied nehmen" (*say goodbye*),[18] but led to errors with idioms, as with "das geht auf keine Kuhhaut" (*this is impossible*), which cannot be marked as a verb (or as a verb phrase).

A difficulty related with the presence of such multi-word units (MWUs) for the lemmas associated with the synsets is the fact that very few morphological and lexical data sets have such MWUs as their lemmas or headwords, so that it can be hard to automatically map a lemma of OdeNet to a German lexical or morphological resources and therefore some manual work will be needed to encode such multi-word units in the OntoLex-Lemon representation. A segmentation algorithm can be helpful in this case, relating the basic components of a MWU to existing headwords in a lexicon.

Another issue with the OdeNet data is the fact that a high number of definitions associated with the synsets are only in English, as they have been first imported from the Princeton WordNet. Those definitions still need to be translated or adapted to German, preferably by a human expert.

The lemmas are also translated into English and so mapped to PWN via the semantic multilingual identifier (ili). For example "Flügel;Tragfläche;Flugzeugflügel" is translated with "wing", which is annotated in PWN with the multilingual semantic ID "i61201". This feature is important as it can ensure the cross-linking of OdeNet to other wordnets in OMW.

For the example "Flügel;Tragfläche;Flugzeugflügel" (*wing*) we have in the OdeNet

[16] https://www.openthesaurus.de/ and the Open Multilingual WordNet English[17] resource. OpenThesaurus is a large resource, generated and updated by the crowd.

[17] See http://globalwordnet.github.io/schemas/.

[18] But in fact we would prefer to categorize those expressions as being verb phrases.

format the following lexical entries and the corresponding entry for the synset:

```
<LexicalEntry id="w3226">
    <Lemma writtenForm="Flügel" partOfSpeech="n"/>
    <Sense id="w3226\_648-n" synset="odenet-648-n"/>
    <Sense id="w3226\_4974-n" synset="odenet-4974-n"/>
    <Sense id="w3226\_8657-n" synset="odenet-8657-n"/>
    <Sense id="w3226\_9783-n" synset="odenet-9783-n"/>
    <Sense id="w3226\_10207-n" synset="odenet-10207-n"/>
    <Sense id="w3226\_11256-n" synset="odenet-11256-n"/> </LexicalEntry>
<LexicalEntry id="w39183">
    <Lemma writtenForm="Tragfläche" partOfSpeech="n"/>
    <Sense id="w39183\_9783-n" synset="odenet-9783-n"/>
</LexicalEntry>

<LexicalEntry id="w39184">\\
    <Lemma writtenForm="Flugzeugflügel" partOfSpeech="n"/>\\
    <Sense id="w39184\_9783-n" synset="odenet-9783-n"/>\\
</LexicalEntry>

<Synset id="odenet-9783-n" ili="i61201" partOfSpeech="n" dc:description="one of
    the horizontal airfoils on either side of the fuselage of an airplane">
    <SynsetRelation target='odenet-3131-n' relType='holo\_ part'/>
    <SynsetRelation target='odenet-18647-n' relType='hyponym'/> </Synset>
```

From the 36,000 OdeNet synsets, about 20,000 contain links to OMW. Approximately 10,000 hyponymy relations and 2,650 antonymy relations are inserted.

In a first evaluation 7% of the PoS entries and 18% of the ili entries were not correct. There is also a need to add more relations and to correct existing ones. With the porting to OntoLex-Lemon we hope, among other things, to discover other issues for OdeNet entries that need correction.

6. Porting OdeNet to OntoLex-Lemon

In order to make OdeNet available in the Linguistic Linked Open Data cloud[19] we need to transform its encoding format (compliant to the GWA[20] WordNet XML DTD[21]) to

[19] http://linguistic-lod.org/llod-cloud, see also Chiarcos et al. (2012).

[20] "GWA" stands for Global WordNet Association. See http://globalwordnet.org/.

[21] http://globalwordnet.github.io/schemas/WN-LMF-1.0.dtd.

an RDF[22] representation. As the target representation framework we have chosen the OntoLex-Lemon model,[23] the core module of which is depicted in Figure 1.

This model is not only the de-facto standard for representing lexical data in the Linked Data framework, but it also includes a property called ontolex:lexicalConcept, which is very important for representing the relation between WordNet synsets and lexical data.[24]A key issue we had to handle with the original crowd-sourced data was that additional textual information was added to the headword, and our script for transforming the OdeNet data to OntoLex-Lemon had to clean the headword field and encode the additional information in a "comment" field. A second issue is related to the improper use of part-of-speech (PoS) information, as soon as the data was not about a noun, a verb or an adjective (the main part-of-speech information in WordNet dictionaries). We filtered out all the entries marked with PoS "p" and will link the entries to well-established German lexical data in the Linguistic Linked Data cloud in order to extract the correct PoS information. We also mapped some OdeNet codes into the LexInfo vocabulary for PoS and semantic relations.[25]

As for now, we have in the OntoLex-Lemon encoding of OdeNet 120,012 lexical entries, the same number of lexical senses and 36,192 synsets, which are encoded as *ontolex:LexicalConcepts* and included in an SKOS [26] based conceptual hierarchy, supporting also the description of lexical semantic relations between synsets, like synonymy, hyponomy, etc.

It is interesting to notice that 44,506 entries contain a blank and can therefore be considered as Multi Word Expressions (MWEs). And if we add to this figure all the 14,080 compound entries[27] we note that approximately half of the lexical entries in the OntoLexLemon representation can be considered as segmentable lemmas.

We give now some details on the OntoLex-Lemon encoding of the first entry in OdeNet, which is "Kernspaltung" (*nuclear fission*). This example is a compound word, which we need to segment in order to be able to represent its components. This representation is supported by the Decomp module of OntoLex-Lemon, which is displayed in Figure 2. First we display the original OdeNet XML representation for "Kernspaltung":

---

[22] RDF stands for "Resource Description Framework", see also https://www.w3.org/RDF/.

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

[24] See the section "Lexical Linkset" in https://www.w3.org/community/ontolex/wiki/Final_Model_Specification.

[25] See https://www.lexinfo.net/ontology/2.0/lexinfo and also Cimiano et al. (2011).

[26] See https://www.w3.org/2004/02/skos/ for more details.

[27] This figure was computed merely by comparison with the list of split nominal compounds offered by the GermaNet project on its web page: http://www.sfs.uni-tuebingen.de/GermaNet/documents/compounds/split_compounds_from_GermaNet13.0.txt, We expect to have a larger number of compounds by applying a decomposition algorithm, not only to nominal entries.

```
<LexicalEntry id="w1">
    <Lemma writtenForm="Kernspaltung"
           partOfSpeech="n"/>
    <Sense id="w1_1-n" synset="odenet-1-n"/>
</LexicalEntry>
<LexicalEntry id="w2">
    <Lemma writtenForm="Kernfission"
           partOfSpeech="n"/>
    <Sense id="w2_1-n" synset="odenet-1-n"/>
</LexicalEntry>
```

Lexical senses are grouped in synsets, i.e., groups of word senses with the same meaning. Hierarchical relations are introduced as synset relations:

```
<Synset id="odenet-1-n" ili="i107577"
  partOfSpeech="n" dc:description="a
nuclear reaction in which a massive
  nucleus splits into smaller nuclei with
  the simultaneous release of energy">
<SynsetRelation target='odenet-5437-
  n' relType='hypernym'/>
</Synset>
```

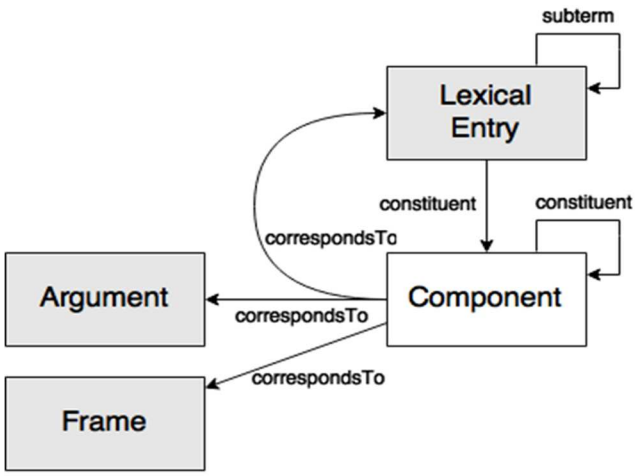

Figure 2: The Decomposition module of OntoLex-Lemon. Graphic taken from https://www.w3.org/2016/05/ontolex/.

In the following Listings we show the Ontolex-Lemon representation of “Kernspaltung”.

```
: entry_w1 rdf : type ontolex : LexicalEntry ;
   decomp : constituent :Kern_comp ;
   rdf :_1 :Kern_comp ;
   decomp : subterm : entry_w3542 ;
   decomp : constituent : spaltung_comp ;
   rdf :_2 : spaltung_comp ;
decomp : subterm: entry_w23527 ;
   lexinfo : hypernym : synset_odenet -5437-n ;
   lexinfo : partOfSpeech lexinfo : noun ;
   ontolex : canonicalForm :form_w1 ;
   ontolex : sense : sense_w1_1-n ;
   ontolex : evokes : synset_odenet-1-n ;
.
```

Listing 1.2: The lexical entry for *Kernspaltung*

In Listing 1.2 we display the full OntoLex-Lemon entry. One aspect that can be immediately noted is the possibility to represent the components of the compound word. This demonstrates one of the benefits of linking synsets to the (complex) representation of lexical entries, as we can state (see below) the semantic relations between synsets associated with the components of a compound word and its own synset.

Listing 1.3 below shows the form information associated to the w1 entry in Listing 1.2.

```
:form_w1 rdf : type ontolex :Form ;
   ontolex : writtenRep " Kernspaltung "@de ;
.
```

Listing 1.3: The ontolex:Form *Kernspaltung*

Listing 1.4 shows the conversion of the original OdeNet sense information into an instance of the ontolex:LexicalSense class.

```
: sense_w1_1-n rdf : type ontolex : LexicalSense ;
   ontolex : isLexicalizedSenseOf
      : synset_odenet-1-n ;
   ontolex : isSenseOf : entry_w1 ;
   ontolex : reference
      https ://www. wikidata .org/wiki/Q11429 ;
.
```

Listing 1.4: The LexicalSense associated to the entry for *Kernspaltung*

In this code we see how the property ontolex:isLexicalizedSenseOf is linking a sense to

a synset, while the entry itself can be linked to the synset via the property ontolex:evokes, as shown in Listing 1.1. The property (ontolex:reference) also links the sense to an ontological entity, here in the form of a Wikidata entry.

Listing 1.5 shows the representation of the synset associated with both the w1 lexical entry and the w1_1-n sense. There we can also see that this lexical concept (synset) is also "evoked" by other entries/senses. For example by the entries for "Kernfission" or "Atomspaltung", which are synonyms of "Kernspaltung". The lexinfo:hypernym property provides information on the semantic relation this synset has to another synset.

```
: synset_odenet-1-n
    rdf : type ontolex : LexicalConcept ;
    skos : inScheme :ODEnet ;
    skos : definition "a nuclear reaction
    in which a massive nucleus splits
    into smaller nuclei with the
    simultaneous release of energy " ;
   wn: i l ii l i : i107577;
    ontolex : isEvokedBy : entry_w1 ;
    ontolex : isEvokedBy : entry_w2 ;
    ontolex : isEvokedBy : entry_w3 ;
    ontolex : isEvokedBy : entry_w4 ;
    ontolex : lexicalizedSense : sense_w1_1-n ;
    ontolex : lexicalizedSense : sense_w2_1-n ;
    ontolex : lexicalizedSense : sense_w3_1-n ;
    ontolex : lexicalizedSense : sense_w4_1-n ;
    lexinfo : hypernym : synset_odenet -5437-n ;
.
```

Listing 1.5: The LexicalConcept (synset) associated with the entry for *Kernspaltung*

Finally, in Listing 1.6 we show the "entries" for the components of the compound word "Kernspaltung". Those components are pointing to the lexical entries they are related to. The entry :entry_w23527 is, for example, the one corresponding to the noun "Spaltung" (*split*, *fission*, *separation*, *cleavage*, etc.), which has again its own senses and associated synsets. We can here disambiguate the meaning of "Spaltung" as used in the compound, as being the one of "fission". And the whole compound can then be considered as an hyponym of the synset for "fission".

```
:Kern_comp
   rdf : type decomp : Component ;
   decomp : correspondsTo      : entry_w3542 ;
.
: spaltung_comp
   rdf : type decomp : Component ;
   decomp : correspondsTo      : entry_w23527 ;
.
```

Listing 1.6: The two components of the entry *Kernspaltung*

In Listing 1.2 above, we can see the information on the ordering those components have in this entry, marked with the "rdf:_1" and "rdf:_2" constructs. For sure, those component "entries" can be re-used separately for other compounds, such as "Atomspaltung". So that we can collect all the corresponding meanings of a word, even when they are used in compounds, as well as depending on their position in the compounds. Details on the decomposition module of OntoLex-Lemon are shown in Figure 2.

The porting of OdeNet to OntoLex made evident that the introduced senses in OdeNet are not really playing a role. We will in the near future replace the OdeNet senses with lexical senses established in other resources. We will also link the synsets to ontological resources, whereas the BabelNet resource from Navigli and Ponzetto (2012) can be very helpful here. We also see that there is no need to associate a PoS with a synset, as this information is present with the associated lemmas. This way we are reaching a higher level of modularity with the OntoLex-Lemon representation.

7. Current Work

We are currently linking the newly created data in the OntoLex-Lemon representation with the already existing UBY-OmegaWiki lemon-based encoding for German[28], which at the time of its creation (2014) could not make use of the *ontolex:LexicalConcepts* property. This work will result in the merging of two large lexical semantics German resources in OntoLex-Lemon, and make this resource accessible in the Linguistic Linked Data cloud.

8. Acknowledgements

The work presented in this paper has been partially supported by the H2020 project "Prêt-àLLOD" with Grant Agreement number 825182. Contributions by Thierry Declerck have been additionally supported in part by the H2020 project "ELEXIS"

[28] See https://lemon-model.net/lexica/ubyow_deu/.

with Grant Agreement number 731015. We would like to thank the anonymous reviewers for their very valuable comments, which we tried to adequately address in the final version of the paper.

9. References

Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, pp. 1352–1362. http://aclweb.org/anthology/P13-1133.

Bond, F. & Paik, K. (2012). A survey of wordnets and their licenses. *Small*, 8(4), p. 5.

Bond, F., Vossen, P., McCrae, J. P. & Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference*, volume 2016.

Chiarcos, C., Nordhoff, S. & Hellmann, S. (eds.) (2012). *Linked Data in Linguistics Representing and Connecting Language Data and Language Metadata*. Springer. https://doi.org/10.1007/978-3-642-28249-2.

Cimiano, P., Buitelaar, P., McCrae, J. P. & Sintek, M. (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), pp. 29–51.

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

Declerck, T. & Racioppa, S. (2019). Porting Multilingual Morphological Resources to OntoLex-Lemon. In R. Mitkov & G. Angelova (eds.) *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019*. INCOMA Ltd.

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

Gonzalez-Agirre, A., Laparra, E. & Rigau, G. (2012). Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., Garcia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation*, 46(4), pp. 701–719.

McCrae, J. P., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (eds.) *Proceedings of eLex 2017*. Brno: Lexical Computing CZ s.r.o., pp. 587–597.

Morato, J., Marzal, M., Lloréns, J. & Moreiro, J. (2004). WordNet Applications. pp. 270–278. http://www.fi.muni.cz/gwc2004/proc/105.pdf.

Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artif. Intell.*, 193, pp. 217–250. http://dx.doi.org/10.1016/j.artint.2012.07.001.

Petitpierre, D. & Russell, G. (1995). MMORPH: The Multext Morphology Program. Multext deliverable 2.3.1, ISSCO, University of Geneva. URL http://www.issco.unige.ch/downloads/multext/mmorph.doc.ps.tar.gz.

Pianta, E., Bentivogli, L. & Girardi, C. (2002). MultiWordNet: Developing an Aligned Multilingual Database. In *In Proceedings of the First International Conference on Global WordNet*. Mysore, India, pp. 293–302.

Sagot, B. & Fišer, D. (2008). Building a free French wordnet from multilingualresources. In E.L.R.A. (ELRA) (ed.) *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Toral, A., Bracale, S., Monachini, M. & Soria, C. (2010). Rejuvenating the Italian WordNet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*. Mumbai.