Porting a crowd-sourced German lexical semantic Resource to Ontolex-Lemon

Thierry Declerck[1] **and Melanie Siegel**[2]

[1] German Research Center for Artificial Intelligence
[2] Hochschule Darmstadt – University of Applied Sciences

Abstract

In this submission we present our work consisting in mapping the recently created "Open-de-WordNet" resource into the OntoLex-Lemon format, in order to make this new WordNet-compliant resource for German available in the Linguistic Linked Data cloud.

1. The Open-de-WordNet

The Open-de-WordNet (Odenet) initiative[1] is aiming at including a crowd-sourced German resource in a multilingual WordNet context, where the concepts (the synsets) of the languages are linked, and where the resources are under an open-source license, being eventually included in the NLTK language processing package[2], so that the resource can be further developed by researchers while using the resource for their Natural Language Processing projects. Odenet is combining two resources: The OpenThesaurus German synonym lexicon[3] and the Open Multilingual WordNet English[4] resource, with the aim to link the resulting WordNet-compliant German resource to the Princeton WordNet of English (PWN)[5]. Odenet was automatically created, but manual corrections were made, German definitions have been introduced, relations between synsets have been supplemented and "ili" links (links to WordNet multilingual concepts) were added. Odenet is coming with an open license (CC BY-SA 4.0).

2. Porting Odenet to OntoLex-Lemon

In order to make Odenet available in the Linguistic Linked Open Data cloud[6] we need to transform its encoding format (compliant to the GWA WordNet XML DTD[7]) to an RDF[8] representation. As the target representation framework we have chosen the OntoLex-Lemon model[9], 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[10]. A

[1] `https://github.com/hdaSprachtechnologie/odenet`.
[2] `https://www.nltk.org/`, see also Bird, Klein, and Loper, 2009.
[3] https://www.openthesaurus.de/. OpenThesaurus is a large resource, generated and updated by the crowd.
[4] http://compling.hss.ntu.edu.sg/omw/
[5] `https://wordnet.princeton.edu/`, see also Fellbaum, 1998.
[6] `http://linguistic-lod.org/llod-cloud`, see also Chiarcos, Nordhoff, and Hellmann, 2012.
[7] `http://globalwordnet.github.io/schemas/WN-LMF-1.0.dtd`.
[8] RDf stands for "Resource Description Framework", see also `https://www.w3.org/RDF/`.
[9] See Cimiano, McCrae, and Buitelaar, 2016 and `https://www.w3.org/2016/05/ontolex/`.
[10] See the section "Lexical Linkset" in `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`

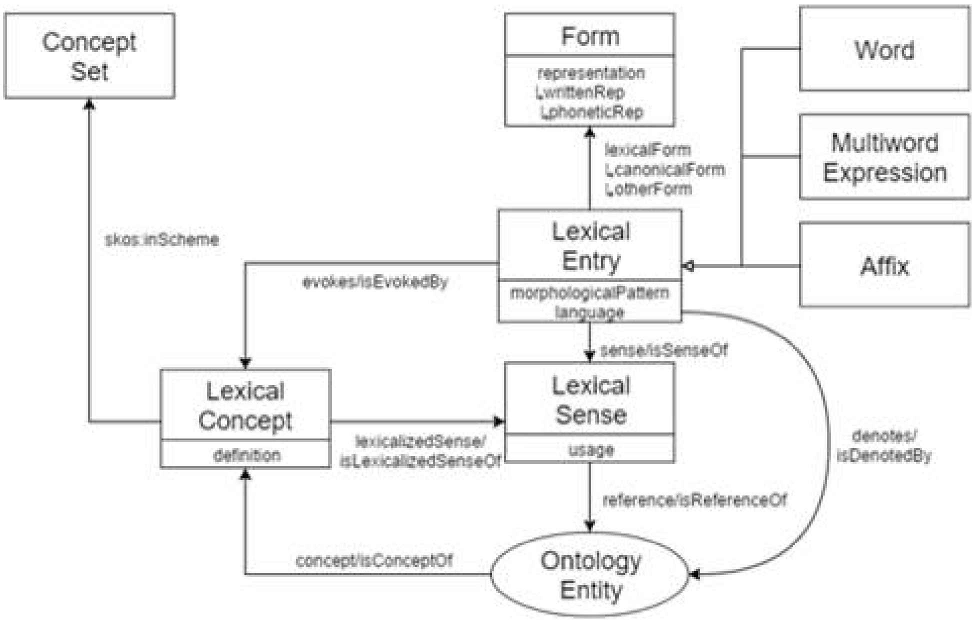

Fig. 1: The core module of OntoLex-Lemon, taken from `https://www.w3.org/2016/05/ontolex/`

main issue we had to deal with the original crow-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. As for now, we have in the OntoLex-Lemon encoding of OdeNet 120012 lexical entries, the same number of lexical senses and 36192 synsets, which are encoded as *ontolex:LexicalConcepts* and included in a *skos*[11] based conceptual hierarchy, supporting also the description of lexical semantic relations between synsets, like synonymy, hyponomy etc.

3. Current Work

We are currently linking the newly created data in the OntoLex-Lemon representation with the already existing UBY-OmegaWiki lemon-based encoding of lexical semantic resource for German[12], 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 semantic German resources in OntoLex-Lemon and make this resource accessible in the Linguistic Linked Data cloud.

[11] See `https://www.w3.org/2004/02/skos/` for more details.
[12] See`https://lemon-model.net/lexica/ubyow_deu/`.

References

Books

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python*. 1st. O'Reilly Media, Inc. ISBN: 0596516495, 9780596516499.

Chiarcos, Christian, Sebastian Nordhoff, and Sebastian Hellmann, eds. (2012). *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata*. Springer. ISBN: 978-3-642-28248-5. DOI: `10.1007/978-3-642-28249-2`. URL: `https://doi.org/10.1007/978-3-642-28249-2`.

Fellbaum, Christiane, ed. (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA ; London: The MIT Press. ISBN: 978-0-262-06197-1. URL: `http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=8106`.

Technical Reports

Cimiano, Philipp, John P. McCrae, and Paul Buitelaar (2016). *Lexicon Model for Ontologies: Community Report*. W3C Community Group Final Report.