Proceedings of the Globalex Workshop on Linked Lexicography, pages 33–35
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Towards an Extension of the Linking of the Open Dutch WordNet with Dutch Lexicographic Resources

Thierry Declerck[1,2]
[1] DFKI GmbH, Multilinguality and Language Technology Lab
[2] Austrian Centre for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences
[1] Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
[2] Sonnenfelsgasse 19, A-1010 Vienna, Austria
declerck@dfki.de

Abstract

This extended abstract presents on-going work consisting in interlinking and merging the Open Dutch WordNet and generic lexicographic resources for Dutch, focusing for now on the Dutch and English versions of Wiktionary and using the Algemeen Nederlands Woordenboek as a quality checking instance. As the Open Dutch WordNet is already equipped with a relevant number of complex lexical units, we are aiming at expanding it and proposing a new representational framework for the encoding of the interlinked and integrated data. The longer term goal of the work is to investigate if and on how senses can be restricted to particular morphological variations of Dutch lexical entries, and how to represent this information in a Linguistic Linked Open Data compliant format.

Keywords: Dutch WordNet, Lexicography, Linking, OntoLex-Lemon

1. Introduction

Work on interlinking or merging language data for Italian, Spanish and French included in wordnets on the one side and morphological data sets on the other side is documented in (Racioppa and Declerck, 2019). The authors accessed for this experiment Wordnet data that are available at the Open Multilingual Wordnet (OMW, (Bond and Paik, 2012; Bond and Foster, 2013)) portal.[1] OMW brings together wordnets in different languages, harmonizing them in a uniform tabular format that lists synsets IDs and the associated lemmas. OMW is linking those Wordnets to the original Princeton WordNet (PWN, (Miller, 1995; Fellbaum, 1998)). Additionally, XML versions of LMF and *lemon* representations[2] of the data are provided.

The morphological data used in those experiments were taken from updated versions of the MMorph data sets.[3] (Declerck et al., 2019) describe a similar experiment conducted for combining the German data from MMorph with an emerging lexical semantics resource for German.

In all those experiments, the OntoLex-Lemon model (Cimiano et al., 2016)[4] was used for representing the linking and merging of the language data originating from both the Wordnet and the MMorph frameworks.

In our current work we expand this kind of experiments beyond the use of morphologies and consider also full lexical resources.

It has been shown that the access and use of Wiktionary can be helpful in a series of applications. (Kirov et al., 2016), for example, describe work to extract and standardize the data in Wiktionary and to make it available for a range of NLP applications, while the authors focus on extracting and normalizing a huge number of inflectional paradigms across a large selection of languages. This effort contributed to the creation of the UniMorph data (`http://unimorph.org/`).

BabelNet[5] is integrating Witkionary data[6] with a focus on sense information, in order to support, among others, word sense disambiguation and tasks dealing with word similarity and sense clustering (Camacho-Collados et al., 2016).

(McCrae et al., 2012b) is directly paving the way for our work, whereas we are upgrading the described approach by the use of OntoLex-Lemon and focusing on establishing relations between senses and morphological forms, and not only between senses and lexical entries.

In our current work, which is dealing with the Dutch language, we consider for the Wordnet side the Open Dutch WordNet (ODWN) and for the lexicographic side the XML dump of the Dutch edition of Wiktionary. We also access the XML dump of the English edition of Wiktionary, in order to extract the descriptions of Dutch nouns included in this edition and to compare them with those proposed in the Dutch edition. When discovering discrepancies between the two, we check manually if a corresponding entry is included in the "Algemeen Nederlands Woordenboek",[7] as a referential point for taking a decision on which data source is to be selected.

[1] See `http://compling.hss.ntu.edu.sg/omw/` for downloading the resources.

[2] LMF stands for"Lexical Markup Framework", an ISO standard. See (Francopoulo et al., 2006) and `http://www.lexicalmarkupframework.org/` for more details. *lemon* stands for "LExicon MOdel for oNtologies". See (McCrae et al., 2012a) and `https://lemon-model.net/` for more details.

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

[4] OntoLex-Lemon is a further development of the *lemon* model. See also `https://www.w3.org/2016/05/ontolex/` for more details on the model.

[5] See (Navigli and Ponzetto, 2010) and `https://babelnet.org/`.

[6] As far as we are aware of, BabelNet integrates only the English edition of Wiktionary, but includes all the languages covered by this edition.

[7] See `http://anw.inl.nl/` and (Tanneke Schoonheim, 2010).

2. Open Dutch WordNet

(Postma et al., 2016) describe how the Open Dutch WordNet (ODWN) combines lexical semantics information and lexical units. This is partially done, as the authors of ODWN had to remove from the predecessor resource, called “Cornetto” (Vossen et al., 2008), a large part of the lexical units, which were owned by a publishing house not willing to publish them as open source. So that “only” around 50,000 full lexical units are associated to the 117,914 ODWN synsets. Those lexical units are originating from the “Referentie Bestand Nederlands” (RBN).[8] In order to replace the removed lexical units, data from public sources, including Wiktionary, were accessed, but this was limited to the “lemmas” that could be associated to a (Dutch) synset to be aligned to a PWN synset. Our aim is thus to add to those lemmas a full lexical description.

ODWN also converted its data to *lemon*, and in our current work we are aiming at upgrading this formal representation to OntoLex-Lemon, the successor of *lemon*, as this new model is designed to also accommodate conceptual lexical data such as those one can find in a wordnets.

3. Wiktionary

Our work consists in accessing lexical data from the XML dump of the Dutch edition of Wiktionary,[9] with a focus for now on Dutch nouns. When we say “XML dump” of Wiktionary, we have to precise that most of the data within the XML encoded general entries are in fact encoded using the MediaWiki markup language, which is more intended for generating a human readable web page. Some of the data is included in such a way that tools are called for generating the information to be displayed in HTML tables, like the (possibly complex) display of inflection of entries.

As mentioned above, we are also accessing the English Wiktionary for Dutch nouns, as there all metadata and definitions etc. are in English, easing thus the comparison between entries of different languages. There are about 52,000 entries for Dutch words in the English Wiktionary.[10]

Consulting the Dutch Wiktionary, we see that from the total of 754,631 entries (also called “pages”), 388,786 are about Dutch words (and 11,330 about English words).

A first comparison of both sources for Dutch words shows that there is in general a certain level of congruence of information between them, while the Dutch Wiktionary is more expansive on semantically related words. It might happen that one source is displayed more definitions (“senses”) than the other, and this constitutes a challenge for the automatic merging of sense-related information.[11] Also the ways of encoding the lexical information are distinct. So, for the Dutch word “route” (*road*, *way*), the English Wiktionary encodes the information on Part-of-Speech, gender, plural form(s) and diminutive(s) this way:

```
{{nl-noun|f|-s|pl2=-en|routetje}}
```

while in the Dutch edition the more or less corresponding data is displayed this way:

```
{{-nlnoun-|{{pn}}|[[{{pn}}n]]
[[{{pn}}s]]|bezield=nietgeanimeerd|
        meta=abstract|telbaar=ja}}
{{-noun-|nld}}
'''{{pn}}''' {{m}}
```

where we can notice that the information on diminutive form(s) is missing, whereas there is some semantic information added.[12]

But it seems that the information on the gender is contradictory, as the English Wiktionary is indicating for the entry the feminine gender, and the Dutch version the masculine gender. Using here the Algemeene Nederlandse Woordenboek (ANW)[13] as a “referee”, we see that the word is in fact “ mannelijk of vrouwelijk” (*masculine or feminine*), which corresponds to the distribution of genders in Dutch, following which nouns are either of grammatical gender “common” (“feminine or masculine”) and “neuter”.

So that even within Wiktionary there is a need to harmonize data representation across distinct language-based editions. For this we are currently porting the Wiktionary data into OntoLex-Lemon. This way we can compare, link and merge with lexical data from the ANW[14] and associate those lexical unit with the OntoLex-Lemon encoding of the ODWN synsets.

4. Conclusion

In this extended abstract, we presented current work aiming at adding further lexical data to the Open Dutch WordNet. This goal requires that we first harmonize all the data sources we are considering, using for this purpose the OntoLex-Lemon model. The longer term goal of our work is to be able to represent the association of senses to morphological variants of lexical entries.

5. Acknowledgements

The presented work is supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreeement no. 731015). It is also partially supported by the COST Action CA18209 - NexusLinguarum “European network for Web-centred linguistic data science”.

We also thank the anonymous reviewers for their helpful comments.

[8] The Referentiebestand Nederlands - RBN (Version 2.0.1) (2014) is available at the Dutch Language Institute: `http://hdl.handle.net/10032/tm-a2-n2`. See also (van der Vliet, 2007).

[9] The dumps of Wiktionary can be downloaded at `https://dumps.wikimedia.org/backup-index-bydb.html`. The human readable Dutch version of Wiktionary is accessible at `https://nl.wiktionary.org/wiki/Hoofdpagina`.

[10] Data is taken from `https://en.wiktionary.org/wiki/Wiktionary:Statistics`.

[11] This topic is at the core of a challenge on “Monolingual Word Sense alignment (MWSA)” organized in the context of the ELEXIS project (`https://elex.is/`). See for more details on this challenge: `https://sinaahmadi.github.io/resources/mwsa.html`.

[12] Both human readable entries can be accessed at `https://en.wiktionary.org/wiki/route#Dutch` and `https://nl.wiktionary.org/wiki/route` respectively.

[13] `http://anw.inl.nl/article/route`.

[14] A description of ANW lexical data encoded in OntoLex-Lemon is given in (Tiberius and Declerck, 2017).

6. Bibliographical References

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August. Association for Computational Linguistics.

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

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

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

Declerck, T., Siegel, M., and Gromann, D. (2019). Ontolex-lemon as a possible bridge between wordnets and full lexical descriptions. In Christiane Fellbaum, et al., editors, *Proceedings of the Tenth Global Wordnet Conference*, pages 264–271, wyb. Stanisława Wyspiańskiego 27 50-370 Wrocław Poland, 7. Oficyna Wydawnicza Politechniki Wrocławskiej, Oficyna Wydawnicza Politechniki Wrocławskiej.

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

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical markup framework (lmf). In *International Conference on Language Resources and Evaluation-LREC 2006*.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of wiktionary morphological paradigms. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

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

McCrae, J., MontielPonsoda, E., and Cimiano, P., (2012b). *Integrating WordNet and Wiktionary with lemon*, pages 25–34. Springer Berlin Heidelberg, Berlin, Heidelberg.

Miller, G. A. (1995). Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, July. Association for Computational Linguistics.

Petitpierre, D. and Russell, G. (1995). MMORPH: The Multext morphology program. Multext deliverable 2.3.1, ISSCO, University of Geneva.

Postma, M., van Miltenburg, E., Segers, R., Schoen, A., and Vossen, P. (2016). Open dutch wordnet. In *Proceedings of the Eight Global Wordnet Conference*, Bucharest, Romania.

Racioppa, S. and Declerck, T. (2019). Enriching open multilingual wordnets with morphological features. In Raffaella Bernardi, et al., editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics*. CEUR, 10.

Tanneke Schoonheim, R. T. (2010). Dutch lexicography in progress: the algemeen nederlands woordenboek (anw). In Anne Dykstra et al., editors, *Proceedings of the 14th EURALEX International Congress*, pages 718–725, Leeuwarden/Ljouwert, The Netherlands, jul. Fryske Akademy.

Tiberius, C. and Declerck, T. (2017). A lemon model for the anw dictionary. In Iztok Kosem, et al., editors, *Proceedings of the eLex 2017 conference*, pages 237–251. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., 9.

van der Vliet, H. (2007). The referentiebestand nederlands as a multi-purpose lexical database. *International Journal of Lexicography*, 20(3):239–257.

Vossen, P., Maks, E., Segers, R., and van der Vliet, H. (2008). Integrating lexical units, synsets and ontology in the cornetto database. In European Language Resources Association (ELRA), editor, *Proceedings of LREC 2008, Marrakech*.