Open Dutch WordNet

Marten Postma
VU Amsterdam
Amsterdam, The Netherlands
m.c.postma@vu.nl

Emiel van Miltenburg
VU Amsterdam
Amsterdam, The Netherlands
emiel.van.miltenburg@vu.nl

Roxane Segers
VU Amsterdam
Amsterdam, The Netherlands
roxane.segers@gmail.com

Anneleen Schoen
VU Amsterdam
Amsterdam, The Netherlands
a.m.schoen@vu.nl

Piek Vossen
VU Amsterdam
Amsterdam, The Netherlands
piek.vossen@vu.nl

Abstract

We describe Open Dutch WordNet, which has been derived from the Cornetto database, the Princeton WordNet and open source resources. We exploited existing equivalence relations between Cornetto synsets and WordNet synsets in order to move the open source content from Cornetto into WordNet synsets. Currently, Open Dutch Wordnet contains 117,914 synsets, of which 51,588 synsets contain at least one Dutch synonym, which leaves 66,326 synsets still to obtain a Dutch synonym. The average polysemy is 1.5. The resource is currently delivered in XML under the CC BY-SA 4.0 license[1] and it has been linked to the Global Wordnet Grid. In order to use the resource, we refer to: https://github.com/MartenPostma/OpenDutchWordnet.

1 Introduction

The main goal of this project is to convert the Dutch lexical semantic database Cornetto version 2.0 (Vossen et al., 2013) into an open source version. Cornetto is currently not distributed as open source, because a large portion of the database originates from the commercial publisher Van Dale.[2] The main task of this project is hence to replace the proprietary content of the database with open source content. In order to create Open Dutch WordNet, we used all the synsets and relations from WordNet 3.0 (Fellbaum, 1998) as our basis. We then exploited existing equivalence relations between Cornetto synsets and WordNet synsets in order to replace WordNet synonyms by Dutch synonyms. We further added new concepts that were not matched through hyperonym relations to the WordNet hierarchy. Any new and manually-created semantic relation from Cornetto was added to the database as well. We limited the synonyms, concepts and relations to those on which there are no copy-right claims. In addition, the inter-language links in various external resources were used to add synonyms to the resource. The result is an open source wordnet that combines the merge and expand method described in (Vossen, 1999).

The resource is currently delivered in XML under the CC BY-SA 4.0 license.[3] In order to inspect and improve the resource, a Python module has been created. This module can be found at : https://github.com/MartenPostma/OpenDutchWordnet.

The outline of this paper is as follows. We start with the motivation to create Open Dutch WordNet in section 2, followed by the methodology to create the resource in section 3. An overview of the main components will be provided in section 4. Finally, we discuss the process of making the resource and plans to improve the resource in section 5.

2 Background and motivation

The first version of the Dutch WordNet was developed within the EuroWordNet project starting from a database developed by Van Dale publisher. This database already contained synset-like structures and lexical semantic relations that could be used to efficiently derive a wordnet structure. Licenses were agreed for commercial and research usage. The Dutch WordNet and the Referentie Bestand Nederlands (RBN) (Van der Vliet, 2007) were combined in the Cornetto project (Vossen et al., 2013). RBN has detailed information on

[1] https://creativecommons.org/licenses/by-sa/4.0/

[2] http://www.vandale.nl/

[3] https://creativecommons.org/licenses/by-sa/4.0/

morpho-syntactic, semantic and pragmatic properties of lexical units, with a focus on the combinatorics. The Cornetto database thus provides the semantic organization of a wordnet and the details on each synonym in a synset as can be found in lexical unit based lexicons. An important characteristic of Cornetto is that it has been developed independently from Princeton WordNet (PWN). The synsets in Cornetto were then mapped to synsets in PWN following a merge approach (Vossen, 1999). First, all possible equivalence relations were created between synonyms in synsets using bilingual dictionaries, after which the mappings were ranked on the basis of shared properties, e.g. hyperonyms and hyponyms already linked manually, similar domain labels, and synset membership of multiple translations (Vossen et al., 2008). The Van Dale publisher however decided to stop all collaborations with the research community. This motivated us to develop Open Dutch WordNet, for which we wanted to keep as much as possible the concepts and word meanings that are defined independently of PWN. This implies that we cannot simply follow an expand approach to translate English synonyms in PWN to Dutch words but we need to also match PWN synsets to RBN lexical units.

Figure 1 introduces the main components of the Dutch lexical semantic database Cornetto.

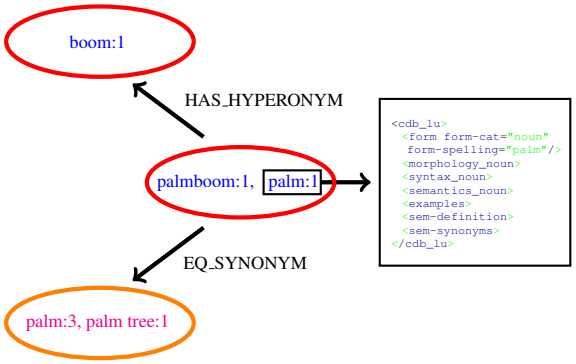

Figure 1: The most important components of Cornetto are visualized. The ellipses in red are examples of Cornetto synsets, which contain Lexical Units (LU). Each LU can contain rich information about its morphology, syntax and semantics. Cornetto synsets can have Internal Semantic Relations (ISRs) to other Cornetto synsets (e.g. HAS_HYPERONYM), but also Equivalence Semantic Relations (ESRs) to PWN synsets (e.g. EQ_SYNONYM).

Figure 1 visualizes the most important components of Cornetto. Cornetto synsets, or Cornetto sets of synonyms, are shown in red. The synonyms inside the Cornetto synsets are called Lexical Units (LU), because they can contain rich information about its morphology, syntax and semantics, especially if these LU's originate from RBN. Synonyms that originate from the Van Dale database only have part-of-speech information. Cornetto synsets can have Internal Semantic Relations (ISRs) to other Cornetto synsets (e.g. HAS_HYPERONYM), but also Equivalence Semantic Relations (ESRs) to PWN synsets (e.g. EQ_SYNONYM). ESRs are mainly used to define synonymy or near synonymy between Cornetto synsets and PWN synsets. Most ISR relations originate from the Van Dale database. A small set of relations were added manually in the various projects. All synonyms and relations have provenance tags which enables us to trace data from Van Dale and data that can transferred to the Open Dutch WordNet.

Table 1 presents the provenance statistics for the most important components of the database:

Component	Van Dale	RBN	Cornetto
LU	60	57	1.5
S	70	1	0
ISR	77	0	33
ESR	0	0	82

Table 1: The provenance information for Lexical Units (LU), Synsets (S), Internal Semantic Relations (ISR), and Equivalence Semantic Relations (ESR) is shown for each of the three sources: Van Dale, RBN, and Cornetto (if the source is Cornetto, this means that the data was created manually in the Cornetto project and does not originate from Van Dale).

Table 1 clearly shows that a large part of the LU's, synsets, and ISRs originate from Van Dale. The removal of this licensed content creates large gaps in the resource. The main goal is hence to use open source resources to replace the licensed content with open source content as much as possible. One of the most promising components to transfer information from Cornetto into Open Dutch WordNet are the ESRs that were created semi-

automatically during the EuroWordNet and Cornetto project and are 100% open source.

3 Methodology

We used the following procedure to create Open Dutch WordNet.

We use English WordNet3.0 (PWN) (Miller, 1995; Fellbaum, 1998) as our basis for the concept structure. This means that we copied the PWN synsets and relations to ODWN and ignored all synsets and relations from Van Dale. The next step is to transfer the LU's from RBN to the PWN-based synsets.

Before copying these LU's we improved the quality of the ESRs. We defined a set of ESRs that are either likely to be more difficult or that play an important role in the transfer. This subset was checked manually and was also used as training to filter the remaining ESRs using a decision tree algorithm. This process is described in subsection 3.1.

Subsequently, we make use of the ESRs between Cornetto synsets and WordNet synsets to copy the LU's that do not originate from Van Dale from a Cornetto synset into a WordNet synset, which is described in subsection 3.2.

The transfer still leaves us with many synsets from PWN without a Dutch LU. We therefore use open source resources to translate the WordNet synonyms into Dutch, which is described in subsections 3.3 and 3.4, respectively. This results on the one hand in more synsets to have Dutch synonyms but also in further evidence for transferred synonyms to be correct because of evidence through other sources.

Finally, we manually checked 8,257 Dutch synonyms, which is described in subsection 3.5.

3.1 Revision of equivalence relations

Firstly, we manually filtered the ESRs, from which we focused on the synonymy relations. Each ESR links a Cornetto synset to a WordNet synset with a certain relation type. The mapping of an ESR is one of many to many. We considered three main aspects of Cornetto synsets in deciding whether to manually check an ESR: the synset depth, the number of children, and the number of ESRs. We decided to manually check the deepest and shallowest synsets because these relations got little attention in previous projects. In addition, we checked the synsets with most children because they play an important role in a wordnet. Finally, the Cornetto synsets with most ESRs were checked because we suspect that the equivalence relation is complex and likely to contain many wrong mappings. Four students manually checked 12,966 of the total 82,285 ESRs, of which 6,575 were removed.

The manually revised relations were used to train a pruned C4.5 decision tree algorithm (Quinlan, 1993; Hall et al., 2009) that was used to filter the remaining ESRs. An ESR consists of an equivalence relation between a Cornetto synset and a WordNet synset. We used properties of the Cornetto synset and the WordNet synset as well as of the synset relation itself as features.

1. the number of equivalence relations in which a Cornetto synset and a Wordnet synset are present.

2. the depth of the Cornetto synset and the Wordnet synsets. The difference of the depth is also used.

3. Because a Cornetto synset can be present in multiple ESRs to WordNet synsets and vice versa, we average the semantic similarity scores (using the Leacock & Chodorow similarity measure (Leacock and Chodorow, 1998)) of of all combinations of these ESRs.

Interestingly enough, the features in which Cornetto properties were used yielded the best results. This might be caused by the fact that the relations were also generated using Cornetto. The filtering of the ESRs using the decision tree algorithm resulted in an additional removal of 32,258 ESRs.

3.2 Cornetto synonyms

When there exists an ESR between a Cornetto synset and a WordNet synset and the relation type is either EQ_SYNONYM or EQ_NEAR_SYNONYM, all LU's that do not originate from Van Dale are inserted into the WordNet synset. Using figure 1 as an example, the LU's *palmboom:1* and *palm:1* would replace *palm tree:1* and *palm:3*. If the ESR was checked manually, the provenance tag is **cdb2.2_Manual**. If the ESR was checked using the decision tree algorithm, the provenance tag is **cdb2.2_Auto**. The provenance tag **cdb2.2_None** is given to all other strategies that were used to add LU's to

Open Dutch WordNet. One of the most dominant strategies of this class is when a LU in a Cornetto synset does not have a direct ESR (no ESR or one of EQ_HAS_HYPERONYM) to a WordNet Synset but the parent of the Cornetto synset does have an ESR to a WordNet synset. In that case a new synset (not represented in WordNet) is created as a hyponym of the target of the ESR of the hyperonym. Finally, the ESRs are used to insert Cornetto synset relations into Open Dutch WordNet that do not originate from Van Dale but were created manually in one of the projects.

3.3 External resources

Using various external open source resources such as Wiktionary (Foundation, 2014b), Omegawiki [4], and Google (Google, 2014), Oliver (2014) translated both monosemous and polysemous lemmas into Dutch for the part of speeches noun, verb, and adjective. For the monosemous lemmas, the English lemmas are simply translated into Dutch. For the polysemous lemmas, the gloss overlap between examples in an external resource and the possible WordNet synsets for a lemma are used to determine the correct synset for a lemma. We used a similar procedure to add synonyms from Wikipedia (Wikipedia, 2014; Foundation, 2014a).

3.4 Adjectives extended

We created a mapping for two kinds of adjectives: monosemous adjectives, that have only one sense in WordNet, and 'slightly polysemous adjectives' that have exactly one adjectival sense and one nominal sense. Adjectives of the latter kind are typically nationalities (*Cameroonian*), religious denominations (*Buddhist*), and words like *purebred*. To create the mapping, we translated the English word forms using Google Translate and Bing Translate. We also use the word alignments from the OPUS project (Tiedemann, 2012). These resources provide us with Dutch candidate word forms that should correspond to the original WordNet synonyms in synsets. We then checked for each word form how many senses are associated with them in RBN. If there is only one (and the word is indeed an adjective), we conclude that this Dutch sense corresponds with the original WordNet synset.

One problem with the translation-based approach is that Dutch adjectives are sometimes inflected with the suffix *-e*. For example, the English *ontological* is automatically translated by Google to *ontologische*. In RBN, all word forms are stored without the inflectional ending, which means that the translation does not match the lemma. To solve this issue, in the cases where we could not find a direct match, we applied an automatic stemming rule to remove the suffix and tried to find a match using the stem.

[4] http://www.omegawiki.org/

3.5 Manual editing

Finally, we checked the resulting Dutch wordnet manually. We focused on two main editing tasks. Firstly, we inspected all synsets that had 10 or more synonyms since excessive synsets may contain false synonyms. In addition, because one Cornetto synset could have multiple ESRs, it occurred that the same sense was copied into multiple WordNet synsets. This may lead to excessive polysemy. The second task therefore consisted of indicating which WordNet synset was the correct synset for a sense that occurred in more than one WordNet synset. In total, 8,257 LU's were checked in this phase.

4 Overview and statistics

In this section, we provide an overview of Open Dutch Wordnet in terms of general statistics, the format it is delivered in, evaluation, and a Python module which allows to interact with the resource.

Open Dutch Wordnet contains 117,914 synsets, of which the majority are noun synsets: 98,049. There are 18,782 verb synsets and 1,083 adjectival synsets. 51,588 synsets contain at least one Dutch synonym, which leaves 66,326 synsets still to obtain a synonym. The resource contains 92,295 synonyms, of which 75,173 are nouns, 15,979 are verbs, and 1,143 are adjectives. The average polysemy is 1.5. 19,996 relations were added to the WordNet hierarchy.

4.1 Format

Open Dutch WordNet is stored in a type of XML called Global WordNet Grid LMF (`https://github.com/globalwordnet/schemas`), which is an adaptation of WordnetLMF (Vossen et al., 2012). The XML contains two main elements: LexicalEntry and Synset. LexicalEntry elements contain information about a specific synonym, whereas Synset elements contain information about synsets. A simplified example

of a LexicalEntry element can be found in figure 2:

```
<LexicalEntry id="ondernemer-n-1"
                partOfSpeech="noun">
      <Lemma writtenForm="ondernemer"/>
      <Sense
       id="r_n-25922"
       senseId="1"
       definition="iemand met eigen bedrijf"
       synset="eng-30-10060352-n"
       provenance="cdb2.2_Auto+wiktionary+google"
       annotator="">
</LexicalEntry>
```

Figure 2: A simplified example of a LexicalEntry element is shown.

In figure 2, an example of a LexicalEntry element is shown. The attributes **id** and **partOfSpeech** of the LexicalEntry element indicate the identifier and the part of speech, respectively. In this example, the identifier is *ondernemer-n-1*, which refers to the first noun sense of the Dutch translation of *entrepreneur* in the sense of "someone who organizes a business venture and assumes the risk for it". The attribute **writtenForm** of the element Lemma indicates the lemma. Following the structure of Cornetto, the LexicalEntry structure represents a lexical unit and not a form unit. The motivation for this is that form properties can differ from one meaning to another for a lemma. The same form can thus appear in multiple LexicalEntry elements.

Finally, the Sense element contains five attributes:

1. **senseId** refers to the synonym sense number.
2. **id** stores the synonym sense identifier. If the identifier starts with *r*, the synonym originates from RBN. In this case, more information about the synonym can be found in RBN. In all other cases, this is not available.
3. **definition** presents the definition for the sense.
4. **synset** points to the synset to which this synonym belongs.
5. Concatenated by '+', the attribute **provenance** shows which resources proposed this particular synonym for this particular synset.
6. the attribute **annotator** shows the name of an annotator and marks that the synonym has been checked manually. The default value is an empty string. Currently, 6,370 LexicalEntry elements have been checked manually.

The LexicalEntry used in Figure 2 belonged to the synset "eng-30-10060352-n". Figure 3 presents a simplified example of that Synset element.

```
<Synset id="eng-30-10060352-n"
                  ili="i89775">
      <Definitions>
         <Definition
         gloss="iemand met eigen bedrijf"
         language="nl"
         provenance="odwn"/>

         <Definition
         gloss="someone who organizes
         a business venture and
         assumes the risk for it"
         language="en"
         provenance="pwn"/>

      <SynsetRelations>
         <SynsetRelation
         provenance="pwn"
         relType="has_hyperonym"
         target="eng-30-09882716-n"/>

         <SynsetRelation
         provenance="odwn"
         relType="role_agent"
         target="eng-30-01651293-v"/>
         ....
         </SynsetRelations>
</Synset>
```

Figure 3: A simplified example of a Synset element is shown.

In figure 3, a simplified example is shown of a Synset element. The Synset attributes **id** and **ili** provide information about the synset identifier and the interlingual index identifier, respectively: `http://data.lider-project.eu/ili`.

The elements Definitions/Definition provide information about the **gloss**, **language**, and **provenance** of the definitions. Finally, the element SynsetRelations/SynsetRelation stores the information about the relations between synsets. Again the **provenance** attribute is used to mark whether the relation originates from PWN or from Cornetto.

4.2 Analysis Lexical Entries

Open Dutch WordNet contains 92,295 synonyms, originating from various resources. Table 2 presents information about the number of synonyms from each resource:

Table 2 presents the number of synonyms proposed by each resource. Note that the same synonym can be proposed by multiple resources, which is why the sum of all numbers is higher than

Provenance	instances	% of all LE
cdb2.2_Auto	32806	35.5
cdb2.2_None	19073	20.7
wiktionary	17968	19.5
cdb2.2_Manual	13075	14.2
omegawiki	12589	13.6
google	8374	9.1
opus	612	0.7
bing	506	0.5
wikipedia	375	0.4

Table 2: The number of synonyms from each resource is shown. In addition, the second column indicates what percentage this number is relative to all synonyms in Open Dutch Wordnet.

the total number of synonyms. The vast majority of synonyms originate from the ESRs (prefixed by *cdb2.2*) between Cornetto synsets and WordNet synsets.

In order to evaluate the quality of each resource for the creation of Open Dutch Wordnet, we randomly evaluated 50 monosemous and polysemous instances. The results can be found in table 3:

Provenance	m	p
Google	0.84	NA
Wiktionary	0.86	0.68
Wikipedia	0.88	0.62
Omegawiki	0.90	0.86
Cdb2.2_Manual	0.88	0.74
Cdb2.2_Auto	0.80	0.80
Cdb2.2_None	0.96	0.78

Table 3: The evaluation results of randomly selected 50 monosemous (m) and polysemous (p) instances per resources is shown.

Table 3 shows that the overall precision of the resource is high as far as the quality of a synonym that bears a certain provenance is concerned. What it does not show, is a fair comparison of the quality of each resource, because not exactly the same strategy was used to extract information from each resource. For example, only monosemous words were used from the output from Google. Overall, we observe that 87% of the proposed monosemous synonyms were correct in the evaluation, whereas this was 76% for the polysemous synonyms. The most valuable external resource for Open Dutch WordNet seems to be Omegawiki, which is not only present in 13.6% of the LexicalEntry elements, but also performed well in the evaluation. For comparison, Sevens (Sevens et al., 2014) performed an independent evaluation of the equivalence relations in Cornetto and reported precision of 52.18% for a sample based on all synsets and 88.94% for a subset that was likely to have manually created links. Although it is difficult to compare both samples for evaluation, the precision for Open Dutch Wordnet is thus very much in line with the precision of Cornetto as reported by them.

4.3 Depth Distribution

66,326 synsets in Open Dutch Wordnet still lack a synonym. We were interested in knowing in which part of the hierarchy these synsets were located. Breadth-first search was used to calculate synset depth. Figure 4 presents the distribution of synsets with and without synonyms per depth layer.

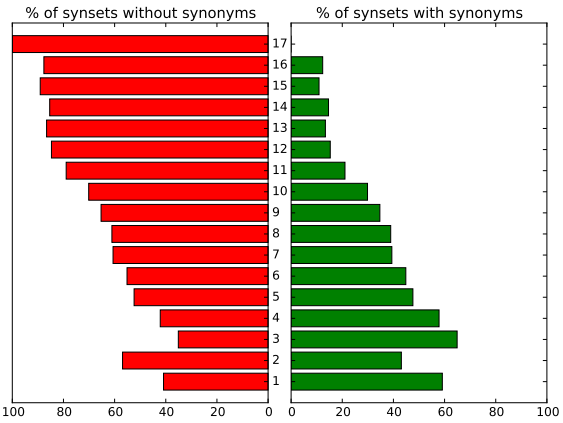

Figure 4: For each depth layer in Open Dutch WordNet, which ranges from the top level 1 to the most deepest layer 17, the percentage of synsets in that layer with and without synonyms is shown.

Figure 4 presents the distribution of synsets with and without synonyms per depth layer. In general, we observe that the top layers have relatively few synsets without synonyms, whereas the opposite is true for the deeper layers. It is likely that these lower level synsets can be filled easily if bilingual resources extend their coverage. These words usually have a single meaning and only one translation.

Also the opposite situation occurs that we added new synsets to the hierarchy that are not in WordNet. These synsets appear to be spread

over all levels of the hierarchy. It is more difficult to resolve these cases since searching for possible matches in WordNet that could have been missed can only partially be supported through e.g. gloss-comparison but in the end needs to be verified manually. To support this process, we visualized these concepts in the hierarchy. An example can be found in Figure 5.

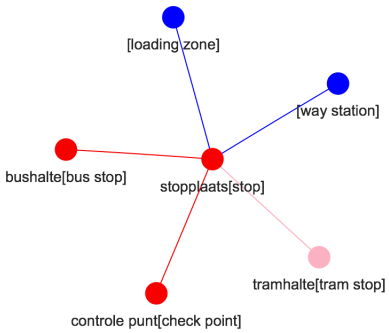

Figure 5: In this visualisation, pink nodes are new concepts, red nodes are WordNet synsets with Dutch synonyms and blue nodes are WordNet synsets without Dutch synonyms.

Figure 5 presents an example of a new concept that has been added to the hierarchy. We added the concept of *tramhalte* (tram stop) as a hyponym of the concept 'stop'. In general, we observed that we mostly added concepts that are represented in Dutch by compounds, such as *polderlandschap* (flat, barren landscape).

4.4 Python module

A Python module has been created to use Open Dutch WordNet. The module can be found at `https://github.com/MartenPostma/OpenDutchWordnet`. It is designed in Python 3.4. The module allows the user to inspect the LexicalEntry and Synset elements and to gather general statistics about the resource. Finally, it is possible to edit the resource using this module.

5 Discussion and future work

In this section, we discuss the process of creating Open Dutch WordNet as well as future work to further improve the resource.

A part of Open Dutch WordNet consists of synonyms that originate from the inter-language links in external resources such as Omegawiki, Wiktionary, and Wikipedia. It is interesting to observe that we obtained mostly noun synonyms

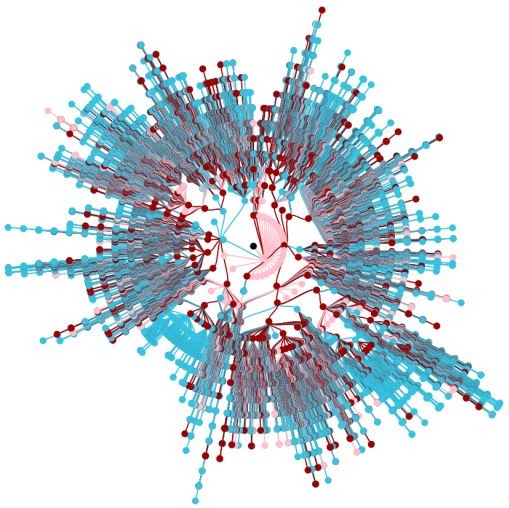

Figure 6: This figure visualizes the noun hyperonym hierarchie in ODWN. The black center node represents the top noun node ('entitiy'). In this visualisation, pink nodes are new concepts, red nodes are WordNet synsets with Dutch synonyms and blue nodes are WordNet synsets without Dutch synonyms.

from these resources. There are two main reasons why this is the case. Firstly, nouns simply have more entries in these resources. In addition, it is obviously more difficult to disambiguate verbs than nouns. In order to get a better understanding of where we added Dutch noun synonyms, we visualized the noun hyperonym hierarchy, which can be found in Figure 6.

In Figure 6, the noun hyperonym hierarchy is visualized, focusing on which synsets contain a Dutch synonym. The lower left side shows a large blue spot, which means that no Dutch synonyms are located in that part of the hierarchy. We identified the synset *genus* ('taxonomic group containing one or more species)' as the main hyperonym of this part. In addition, we observe pink nodes around the top node, which we identified as religious terms such as *Heer* (Lord), and *Jaweh* (Jaweh).

In order to improve the resource, we strive to both improve the quality and quantity of the resource. The quality will be improved by manually inspecting the synsets ranging from 5 to 10 synonyms. The quantity will be improved by adding synonyms in the deeper parts of the resource. This can be done by using more or improved public bilingual resources, both English-Dutch but also

by combining more languages, or by using parallel corpora. In addition, we plan to assess the most important parts of the hierarchy. This involves the top nodes of the hierarchies and the base concepts. Errors in these synsets are likely to propagate to other synsets in lower parts of the hierarchy. Finally, the relations imported from Cornetto are now added to the PWN relations. As a result, we obtained 115,077 hyperonym relations from PWN and 19,996 hyperonym relations from Cornetto. Additional hyperonym relations result in tangled hierarchies with more complex semantics. Whereas PWN has 559 top nodes for verbs, ODWN has 154 tops. The reduction of the tops is due to the additional relations that were created in Cornetto to provide more structure to the verb hierarchy. In Cornetto, there are only two top nodes for the verb hierarchy.

Open Dutch WordNet currently contains a limited amount of monosemous adjectives. We hope to be able to map the polysemous adjective synsets to PWN synsets by translating the Dutch glosses and by making use of the synset relations in Cornetto and Princeton WordNet. Because Dutch is very close to German, another possibility is to map the Cornetto synsets to GermaNet (Hamp and Feldweg, 1997) and make use of the rich set of synset relations that it provides.

Finally, the current format of the resource is XML. We would also like to make the resource available in RDF (Klyne and Carroll, 2006).

6 Conclusion

We described Open Dutch WordNet, which is derived from the Cornetto database, Princeton WordNet and various external resources. We exploited existing equivalence relations between Cornetto synsets and WordNet synsets in order to replace WordNet synonyms by Dutch synonyms. In addition, the inter-language links in various external resources such as Wiktionary and Omegawiki were used to add synonyms to the resource. In addition, we manually evaluated each resource and manually edited the most problematic synsets. The Princeton-based hierarchy was also extended with manually created relations came from Cornetto.

Open Dutch Wordnet contains 92,295 synonyms, which are located in 51,588 synsets. There are 75,173 nouns, 15,979 verbs, and 1,143 adjectives. In total, the resource consists of 117,914 synsets, which leave 66,326 synsets still to obtain a synonym. The average polysemy is 1.5.

The resource is currently delivered in XML under the CC BY-SA 4.0 license.[5] In order to use and improve the resource, a Python module has been created. This module can be found at: `https://github.com/MartenPostma/OpenDutchWordnet`.

Acknowledgments

This project has been co-funded by the Nederlandse Taalunie (`http://taalunie.org/`). In addition, we thank Anne Broekhuis, Anja Stoop, Marjolein Klaassen, and Amber Witsenburg for their work on evaluating the ESRs manually. Moreover, we thank Isa Maks (`https://www.linkedin.com/pub/isa-maks/24/b47/`) and Hennie van der Vliet (`https://www.linkedin.com/pub/hennie-van-der-vliet/0/869/512`) for their valuable input. Finally, we would like to thank Adam Rambousek (`http://www.muni.cz/fi/people/60380`) for his help in creating and updating the DebVisDic editor.

References

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

Wikimedia Foundation. 2014a. Wikipedia. `http://en.wikipedia.org/`.

Wikimedia Foundation. 2014b. Wiktionary. `http://en.wiktionary.org/`.

Google. 2014. Google translate. `https://translate.google.nl/`.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Birgit Hamp and Helmut Feldweg. 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*, pages 9–15.

Graham Klyne and Jeremy J Carroll. 2006. Resource description framework (rdf): Concepts and abstract syntax.

[5] `https://creativecommons.org/licenses/by-sa/4.0/`

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

George A. Miller. 1995. Wordnet: a Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Antoni Oliver. 2014. Wn-toolkit: Automatic generation of wordnets following the expand model. *Proceedings of the 7th Global WordNetConference, Tartu, Estonia*.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Leen Sevens, Vincent Vandeghinste, and Frank Van Eynde. 2014. Improving the precision of synset links between cornetto and princeton wordnet. *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing,Coling 2014, Dublin, Ireland*, pages 120–126.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, pages 2214–2218.

Hennie Van der Vliet. 2007. The Referentiebestand Nederlands as a multi-purpose lexical database. *International Journal of Lexicography*, 20(3):239–257.

P Vossen, I Maks, R Segers, and H Vliet. 2008. van der, zutphen, h. van,(2008). the cornetto database: the architecture and alignment issues. In *Proceedings of the Fourth International GlobalWordNet Conference-GWC 2008*, pages 22–25.

Piek Vossen, Claudia Soria, and Monica Monachini. 2012. Wordnet-lmf: a standard representation for multilingual wordnets. In G. Francopoulo, editor, *LMF: Lexical Markup Framework, theory and practice*, pages 51–66. Hermes, Lavoisier, ISTE.

Piek Vossen, Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke. 2013. Cornetto: a Combinatorial Lexical Semantic Database for Dutch. In Jan Odijk Peter Spyns, editor, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 165–184. Springer.

Piek Vossen. 1999. Eurowordnet: General document. version 3 final. *University of Amsterdam. EuroWordNet LE2-4003, LE4-8328*.

Wikipedia. 2014. Plagiarism — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350`.