Open Source Dutch WordNet

Marten Postma	m.c.postma@vu.nl
Piek Vossen	p.t.j.m.vossen@vu.nl

Release	1.0
Date	December 1, 2014
License	CC BY-SA 4.0
Website	http://wordpress.let.vupr.nl/odwn/

This project has been funded with the help of the Nederlandse Taalunie (http://taalunie.org/).

Contents

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

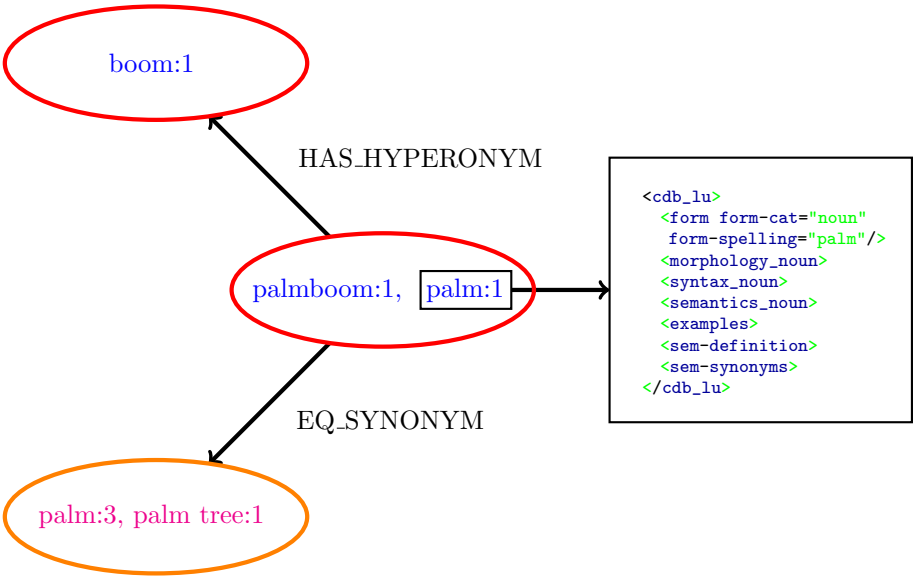


Figure 1: The most important components of Cornetto are visualized. The ellipses in red are examples of Cornetto synsets, which contain Lexical Units (LU). Each LU can contain rich information about its morphology, syntax and semantics. Cornetto synsets can have Internal Semantic Relations (ISRs) to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

[1] This documentation has been written with the user of Open Source Dutch WordNet in mind. It is not to be considered a full technical report of each step of the creation process.

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

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 (Van der Vliet, 2007). Cornetto synsets can have Internal Semantic Relations (ISRs) to other Cornetto synsets (e.g. HAS_HYPERONYM), but also Equivalence Semantic Relations (ESRs) to Wordnet synsets (e.g. EQ_SYNONYM). ESRs are mainly used to define synonymy or near synonymy between Cornetto synsets and Wordnet synsets.

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

Source	**LU**	**%**	**S**	**%**	**ISR**	**%**	**ESR**	**%**
Van Dale	59391	50.3	69562	98.7	76630	70.0	0	0
RBN	56991	48.3	0	0.0	0	0	0	0
Cornetto	1586	1.3	937	1.3	33057	30.0	82285	100

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.

Hence, in order to create an open source version, our goal was to try to replace 50.3% of the LU's, 98.7% of the synsets, and 70% of the ISRs with open source content. Because all the ESRs are open source, we opted for the following procedure to accomplish this goal:

1. We use English WordNet 3.0 (Miller, 1995; Fellbaum, 1998) as our basis, which we converted into Cornetto2.0 XML structure. This means that we replace the Van Dale synsets and ISRs by Wordnet synsets and ISRs.

2. The next step is to replace the English synonyms in the Wordnet synsets by LU's. This is done in two ways:

   (a) When there exists an ESR between a Cornetto synset and a WordNet synset, 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*. However, the automatically-generated ESRs were first filtered before this technique was applied. Four students manually checked 12,966 ESRs, of which 6,575 were removed. Afterwards, the unchecked ESRs were filtered using a decision tree algorithm that used the manual inspection as training. This resulted in a removal of 32,258 ESRs.

(b) Using open source resources (Wikipedia (Wikipedia, 2014;Foundation, 2014a), Wiktionary (Foundation, 2014b), Google Translate (Google, 2014)), the English synonyms in English WordNet are translated into Dutch.

The remainder of this documentation is outlined as follows. Section 2 introduces the XML structure of Open Source Dutch WordNet. This is followed by the main statistics about the database in section 3. Finally, the formats in which the database is available are detailled in section 4, followed by the acknowledgements in section 6.

2 XML structure

An example of a synset in Open Source Dutch WordNet can be found below. The original WordNet 3.0 synset with offset 06722453 now contains Dutch synonyms.

```
<cdb_synset c_sy_id="eng-30-06722453-n" posSpecific="NOUN" comment="EQ_SYNONYM">
       <synonyms>
              <synonym c_lu_id="r_n-39115" c_lu_id-previewtext="uitspraak:2" status="cdb2.2_Manual"/>
              <synonym c_lu_id="o_n-102764766" c_lu_id-previewtext="dictum:1" status="cdb2.2_Manual"/>
              <synonym c_lu_id="t_n-679214835" c_lu_id-previewtext="verklaring:4" status="google_api+wiktionary"/>
       </synonyms>
       <definition>a message that is stated or declared;
       a communication (oral or written) setting forth particulars or facts etc</definition>
       <wn_internal_relations>
                     <relation relation_name="HAS_HYPERONYM" target="eng-06598915-n" source="pwn"/>
                     <relation relation_name="HAS_HYPERONYM" target="eng-06284225-n" source="odwn"/>
       </wn_internal_relations>
</cdb_synset>
```

A formal description of all child elements with all possible attributes will be described in the next subsections.

2.1 Cdb_synset Element

Each Cdb_synset element contains three attributes: c_sy_id, posSpecific, and comment. The possible values of these attributes are explained.

I c_sy_id
This attribute indicates the provenance of the synset. If the synset is prefixed by *eng* (95,356 synsets), the synset origin is English WordNet. All other synsets originate from Open Source Dutch Wordnet and start with **odwn** (21,636 synsets).

II posSpecific
The synset part of speech is denoted by this attribute. It can either be NOUN (98,107 synsets) or VERB (18,885 synsets).

III comment
This attribute explains how the LU's from the Cornetto synsets are inserted into Open Source Dutch Wordnet. As an example, we refer to figure 1. The LU's *palmboom:1* and *palm:1* are both located in a Cornetto synset, which

has an EQ_SYNONYM ESR with the Wordnet synset that contains the English synonyms *palm:3* and *palm tree:1*. In Open Source Dutch WordNet, The LU's *palmboom:1* and *palm:1* will be inserted into the synset that originally contained the English synonyms *palm:3* and *palm tree:1*, and the synset will receive the attribute value EQ_SYNONYM for the attribute comment. All possible values are explained in below:

EQ_SYNONYM
The LU('s) was/were located in a Cornetto synset that had an EQ_SYNONYM ESR with the WordNet synset it/they is/were copied into.

EQ_NEAR_SYNONYM
The LU('s) was/were located in a Cornetto synset that had an EQ_NEAR_SYNONYM ESR with the WordNet synset it/they is/were copied into.

dummy
This is the default value, meaning that this synset only contains one or more English synonyms, and no Dutch synonyms.

eq-parent-match
The LU('s) was/were located in a Cornetto synset that did not have a direct ESR (no ESR or one of EQ_HAS_HYPERONYM) to a WordNet Synset. However the parent of the Cornetto synset did have an ESR to a WordNet synset. A new synset is created as a hyponym of the target of the ESR of the hyperonym of the Cornetto synset.

eq-parent-match;EQ_HAS_HYPERONYM
The only difference between **eq-parent-match;EQ_HAS_HYPERONYM** and **eq-parent-match** is that in this case, the target of the ESR of the synset itself and its parent are the same. The same procedure is used as in **eq-parent-match**.

eq-synonym-preempt
It can occur that multiple Cornetto synsets have an ESR to a WordNet Synset. If one these ESRs is one of EQ_SYNONYM, only the LU's in the synset with that relation will be copied into the WordNet synset. For the LU's in Cornetto synsets with ESRs of EQ_NEAR_SYNONYM, new synsets will be created, which will be hyponyms to the original and therefore automatically becoming co-hyponyms to each other

2.1.1 Synonyms Element

The synonyms elements contains zero or more synonyms elements.

2.1.1.1 Synonym Element

Each synonym element contains four attributes: c_lu_id,c_lu_id-previewtext, and status.

c_lu_id
The prefix value of this attribute indicates the provenance of the LU. There are five distinct prefixes:

prefix	origin
r_	RBN
c_	Cornetto
o_	Open Source Dutch WordNet
t_	Open Source resources
eng_:	original wordnet synonyms

c_lu_id-previewtext
The value of this attribute denotes a synonym. It can be prefix by *eng*, which means that it is an original English synonym, else the synonym is Dutch.

status
The provenance information of which source proposed the LU for the particular synset is explained by this attribute. The possible values are explained below:

Value attribute	origin
cdb2.2_Manual	manually checked ESR
cdb2.2_Auto	automatically checked ESR
babelnet	Babelnet
google_api	Google translate
wiktionary	interlanguage links of wiktionary
wikipedia	interlanguage links of wikipedia

2.2 Definition Element

The value of the definition element contains the original English WordNet definitions or a Dutch definition.

2.3 Wn_internal_relations

The wn_internal_relations element contains zero or more Relation elements.

2.3.1 Relation

The relation element contains three attributes: relation_name, target, and source.

I relation_name The value of this attribute indicates the name of an internal semantic relation inside Open Source Dutch Wordnet. In the current version, there exist 70 different internal semantic relations. For more information, we refer to Vossen (1999).

II target The value of this attribute denotes the target of the internal semantic relation.

III source The value of this attribute indicates the source of the internal semantic relation. It can either be from Princeton Wordnet (eng), or Open Source Dutch Wordnet (odwn).

3 Statistics

This section provides statistics about the resource. A general overview is given in subsection 3.1, after which the open source resources are evaluated in subsection 3.2. Finally, the synsets without Dutch synonyms are analysed in 3.3.

3.1 Overview Open Source Dutch WordNet

In total, Open Source Dutch WordNet contains 116,992 synsets, of which 95,356 originate from WordNet 3.0 and 21,636 synsets are new synsets. The number of English synsets without dutch synonyms is 60743, which means that 34,613 WordNet 3.0 synsets have been filled with at least one Dutch synonym.

3.2 Evaluating Open Source resources

In order to evaluate Open Source Dutch WordNet, we make use of the 12,966 ESRs that have been checked manually. Table 3.2 presents our findings.

R	FreqR	FreqM	R & M	(R & M) / R	(R & M) / M
wiktionary	20707	9478	2586	0.125	0.273
google_api	22708	9478	2769	0.122	0.292
wikipedia	422	9478	29	0.069	0.003
babelnet	20440	9478	923	0.045	0.097

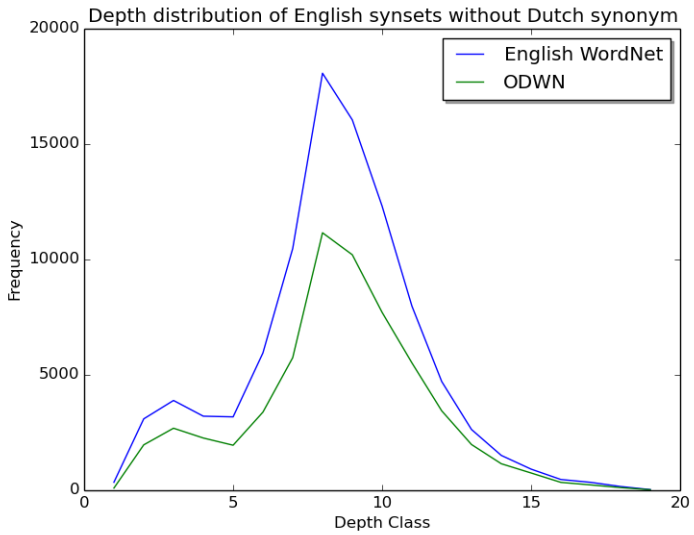
Table 2: Table 3.2 presents an evaluation of the open resources used to translate Wordnet 3.0 synonyms. The first column indicates an open source resource. The second column presents the number of synonyms in Open Source Dutch Wordnet that have this resource as origin. The third column indicates the number of synonyms that have a manually checked ESR as an origin. The fourth column presents the overlap between R and M. The final two columns present the overlap as percentage of of R and as percentage of M.

Table 3.2 presents an evaluation of the open resources used to translate Wordnet 3.0 synonyms. Note that we only add a Dutch synonym if *babelnet* or

a combination of a least two other resources propose the same synonym for the same synset. Firstly, the number of synonyms proposed by *wikipedia* is relatively low, which can be explained by the fact that *wikipedia* mainly consists of proper names and not of nouns, verbs and adjectives. Secondly, it's interesting to see that the other three resources are all present with about 20,000 synonyms. Finally, it's interesting to see that about 30% of the synonyms proposed by *wikipedia* and *google_api* overlap with the synonyms from the manually checked ESRs, whereas this is much lower for *babelnet*.

3.3 Inspection Synsets without Dutch synonyms

There are 60,743 synsets still without a Dutch synonym. We were interested in knowing the depth distribution of these synsets. The following graph presents this distribution.



The blue line indicates the depth distribution of WordNet 3.0. The green line indicates the depth distribution of synsets without Dutch synonyms in Open Source Dutch WordNet. As can be seen, the majority of synsets still to be filled with Dutch synonyms have a depth of 7 on average.

4 Formats

Open Source Dutch Wordnet is currently available in two different formats.

I The first format is Cornetto2.0 XML structure, which has been discussed in section 2. In Cornetto 2.0 XML structure, the information about the LU's

is coming from a different resource, which is RBN (Van der Vliet, 2007). Alongside creating Open Source Dutch WordNet, we also created an open source version of RBN, which we will call ORBN.

II The second format is the Lexical Markup Framework (LMF) (Francopoulo et al., 2007), which has been adapted to wordnets in general (Soria et al., 2009) and to Cornetto (Maks et al., 2013).

5 Future work

We will distribute Open Source Dutch WordNet in the Resource description framework (RDF) (Lassila and Swick, 1999). using the conversion scripts from `https://github.com/jrvosse/cornetto2rdf-conversion`.

6 Acknowledgements

This project has been co-funded by the Nederlandse Taalunie (`http://taalunie.org/`). In addition, special thanks go to Anne Broekhuis, Anja Stoop, Marjolein Klaassen, and Amber Witsenburg for their valuable work on evaluating the ESRs manually. Moreover, thanks go to Isa Maks (`https://www.linkedin.com/pub/isa-maks/24/b47/`) for the help in the LMF conversion, and Jacco van Ossenbruggen (`https://www.linkedin.com/in/jrvosse`) for the help in the RDF conversion. Finally, we would like to thank Adam Rambousek (`http://www.muni.cz/fi/people/60380`) for his help in creating and updating the editor.

References

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

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

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

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. Lexical markup framework: Iso standard for semantic information in nlp lexicons. In *Proceedings of the Workshop of the GLDV Working Group on Lexicography at the Biennial Spring Conference of the GLDV*, 2007.

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

Ora Lassila and Ralph R Swick. Resource description framework (rdf) model and syntax specification. 1999.

Isa Maks, Hennie Van der Vliet, Attila Görög, and Piek Vossen. Cornetto lmf lexical resource for dutch. Internal Report Deliverable D9 (CLARIN-NL-11-020), VU University Amsterdam, Amsterdam, 2013.

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

Claudia Soria, Monica Monachini, and Piek Vossen. Wordnet-lmf: fleshing out a standardized format for wordnet interoperability. In *Proceedings of the 2009 international workshop on Intercultural collaboration*, pages 139–146. ACM, 2009.

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

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

Piek Vossen, Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke. 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, 2013.

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