

Alignment of the Dutch databases in Cornetto

Version: 1
Date: January 22th, 2007



Project: STE05039

<http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm>

Funded by the Stevin framework (www.taalunieversum.org/stevin)

Author: Hetty van Zutphen

Table of Contents:

1	Introduction.....	3
2	Alignment of lexical resources	3
2.1	Lexical resources.....	3
2.2	Alignment strategies	3
2.3	Alignment results	4
3	The Cornetto link database.	5
4	Cornetto link database statistics	6

1 Introduction

The goal of the Cornetto project (STE05039), funded by the Nederlandse Taalunie in the STEVIN framework (www.taalunieversum.org/stevin), is to create a lexical semantic database for Dutch, covering approx. 40K entries, including the most generic and central part of the language.

Cornetto combines and aligns two semantic resources for Dutch: The Dutch Wordnet (vossen 1998) and the Referentie Bestand Nederlands (Martin et al 1999). This document describes the method and the results of the alignment. Link statistics are shown, as well as the structure of the resulting database.

In the next section, the alignment sources, strategies and results are discussed. The following section goes into the structure of the Cornetto database. The last section shows the global size of the results.

2 Alignment of lexical resources

2.1 Lexical resources

Dutch Wordnet is originally based on the conceptual structure of the Van Dale VLIS database. It contains 55K word forms and 70K word meanings. The Referentie Bestand Nederlands (RBN) contains 52K word forms and 66K word meanings. The VLIS database contains 87K word forms and 107K word meanings.

In the Cornetto project, these databases are aligned. This means that word meanings which are equivalent, will be stored as links in a database.

2.2 Alignment strategies

For the alignment, a number of strategies were used. For all strategies, the basic requirement is that the wordform and the part of speech in the databases match.

The strategies used to determine which wordforms can be aligned are:

1. The words have one meaning and no synonyms in both databases
2. The words have one meaning in both databases
3. The words have one meaning in RBN and more than one meaning in VLIS/DWN
4. The word has one meaning in VLIS/DWN and more in RBN
5. If the broader term (BT) of a set of words is linked, all words which are under that BT in the semantic hierarchy and which have the same form are linked

6. If some narrow term in the semantic hierarchy is related, siblings of that NT that have the same form are also linked.
7. Words that have a linked domain, are linked
8. Words with definitions in which one in every three words is the same (there must be more than one match) are linked.

2.3 Alignment results

The number of links found per strategy is as follows

Strategy	1	2	3	4	5	6	7	8
Number of links	9.936	25.366	22.892	1.357	7.305	21.691	11.008	22.664

Evaluation of the strategies

Of the results of each strategy, a sample was made of 100 records. Each sample was checked by a number of persons. For each record, the wordform, part of speech, id, sequence number and the definition was shown for both RBN and VLIS. The testers had to determine whether the definitions described the same meaning of the word, or not.

The results of the test were combined and the result was a list of percentages of items which were considered good links. The averages per strategy were:

Strategy	1	2	3	4	5	6	7	8
Average percentage correct scores	97.1	88.5	53.9	68.2	85.3	74.6	70.2	91.6

On the basis of these results, the strategies were ranked: some were considered very good (3), some were considered average (2), and some were considered relatively poor (1).

The ranking factors per strategy are:

Strategies 1, 2 and 8: 3

Strategies 5, 6 and 7: 2

Strategies 3 and 4: 1

The ranking factor is used to determine the score of a link (see below). The score of the link is determined by the number of strategies which apply and the ranking factor of the strategies

The result of the alignment is stored in the Cornetto Database. This contains the links between RBN and VLIS/DWN items, as well as entries from both databases that could not be linked.

3 The Cornetto link database.

The cornetto link database is stored in XML format. The encoding is UTF-8. The database contains three types of records:

- Records which represent a link between RBN and VLIS/DWN
- RBN records that were not linked
- DWN/VLIS records that were not linked

The XML tag for each record is “cid” Each tag has 15 attributes, as you can see in the following example:

```
<cid
cid_id="4651"
c_lu_id="r_n-6013"
c_pos="NOUN"
c_form="bas"
c_seq_nr="1"
r_lu_id="n-6013"
r_seq_nr="1"
d_lu_id="d_21940"
d_seq_nr="2"
d_sy_id="d_10187"
score="64.6"
date=""
name=""
status=""
selected="true"
/>
```

The attributes have the following values:

cid_id	a new id, ranging from 1 to n in the cornetto database. Each record has a unique cid_id
c_lu_id	this is identical to the r_lu_id with ‘r_’ prefixed to it if the word is either linked, or if it is an unlinked RBN item; otherwise, it is equivalent to the d_lu_id, which starts with ‘d_’
c_pos	the part of speech (NOUN, VERB, ADJECTIVE or REST)
c_form	the wordform
c_seqnr	equivalent to the r_seq_nr if the word is either linked, or if it is an unlinked RBN item; otherwise, it is increased for each (different) d_lu_id for the same wordform;
r_lu_id	RBN id: contains a part of speech (POS) prefix plus the id. POS-prefixes are ‘n’ for nouns, ‘v’ for verbs, ‘a’ for adjectives and ‘r’ for others
r_seq_nr	RBN sequence number
d_lu_id	VLIS lexical entry id

d_seq_nr	VLIS sequence number
d_sy_id	DWN synset id
score	the score of the link:determined by the number of strategies which apply and the ranking factor of the strategies
date=""	the date at which the link was evaluated
name=""	the name of the person or program that evaluated the link
status=""	the status in the manual process
selected	"true" for the highest scoring link for the same RBN form, and "false" for any lower scoring link

For link records, both the RBN attributes (r_lu_id, r_seq_nr) and the VLIS/DWN attributes (d_lu_id, d_seq_nr, d_sy_id) contain a value. For RBN records that are not linked, the VLIS/DWN attributes as well as the score remain empty; for VLIS/DWN records that were not linked, the RBN attributes and the score remain empty.

4 Cornetto link database statistics

In total, 136K entries are stored in the Cornetto link database. There are 58K links, representing 41K word forms. In total 47K different RBN word meanings were linked, and 48K different VLIS/DWN word meanings. 19K word meanings from RBN were not linked, as well as 59K word meanings from VLIS/DWN. The size of the entire Cornetto link database is 28,5Mb.

References

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

Maks, I., W. Martin, and H. de Meerseman (1999) *RBN Manual*, Vrije Universiteit Amsterdam.

Vossen, P (1998, ed.) *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.