

Cornetto Evaluation Report

Cornetto Deliverable D-15

Version 1

11 May 2009

Author: Tanja Gaustad van Zaanen, Polderland Language and Speech Technology

Project STE05039

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

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

Table of Contents

1	Introduction.....	3
2	Document Validation.....	3
2.1	General Documentation	3
2.2	Technical Information	3
2.3	Content Information	4
3	Formal Validation.....	5
3.1	Validation of files	5
3.2	Validation of tools	5
4	Content Validation	5
4.1	Validation of Linguistic Correctness	5
4.1.1	<i>Samples</i>	<i>5</i>
4.1.2	<i>Manual Checks Performed</i>	<i>6</i>
4.1.3	<i>Findings of Manual Validation</i>	<i>6</i>
5	Conclusion	9
6	Appendix.....	10
6.1	List of lexical units included in the evaluation from the core part of Cornetto ...	10
6.1.1	<i>Nouns.....</i>	<i>10</i>
6.1.2	<i>Verbs</i>	<i>11</i>
6.1.3	<i>Adjectives.....</i>	<i>12</i>
6.2	List of lexical units included in the evaluation from the extended part of Cornetto with manually verified equivalence relations, SUMO labels and domain labels	12
6.2.1	<i>Nouns.....</i>	<i>12</i>
6.2.2	<i>Verbs</i>	<i>12</i>
6.3	List of lexical units included in the evaluation from the extended part of Cornetto	12
6.3.1	<i>Nouns.....</i>	<i>12</i>
6.3.2	<i>Verbs</i>	<i>13</i>
6.3.3	<i>Adjectives.....</i>	<i>13</i>
6.3.4	<i>Adverbs.....</i>	<i>13</i>

1 Introduction

This document reports on the evaluation of Cornetto, a lexical semantic database for Dutch. The results of the Cornetto project consist of three major components, namely the Cornetto XML database, the general acquisition tool, and the domain acquisition tool. Of these components only the Cornetto XML database is evaluated here.

The evaluation checks the completeness and content of the documentation of Cornetto. It also provides a formal validation of the data files. Furthermore, the content validation of the lexical semantic database includes 100 lexical units from the core Cornetto database and 35 lexical units from the extended Cornetto database.

The document is structured as follows: In section 2, the documents included in the Cornetto project as deliverables are validated. We include remarks on the technical information contained in the documentation as well as the actual content information. The formal validation in section 3 checks the files and tools included in Cornetto. The main part of the evaluation can be found in section 4 with a detailed description of the content validation. We conclude with a summary of our findings in section 5. The appendix in section 6 contains the lists of manually validated lexical units.

2 Document Validation

2.1 General Documentation

The general documentation for the Cornetto project is written in English. It is comprehensible and in some instances very detailed. All the information on how to acquire and use Cornetto is provided either in the written documentation or on the website (<http://www2.let.vu.nl/oz/cornetto/index.html>). Also, all contact information is provided and support in case of questions or problems is given within a very short delay.

2.2 Technical Information

All technical information on the set-up and content of the database tables can be found in deliverable D-1 "The Design of the Cornetto Database". It explains in detail and with useful figures what the different database tables contain and how they are linked. Deliverables D-02 through D-07 supply technical (and linguistic) information on various aspects of the Cornetto database. The general overview document (deliverable D-16) provides a comprehensive summary and introduction to Cornetto. All of these deliverables are technically sound and give a good insight into the decisions taken during the development of the database.

Deliverables D-08 through D-11 report on the acquisition toolkits and related issues and have not been evaluated here. The same applies for the task-based evaluation reports D-13 and D-14. We trust that they are of the same high quality as the rest of the documentation.

2.3 Content Information

The Cornetto lexicon is a monolingual resource for Dutch. The entries in the database are linked to the English WordNet in order to allow for the import of WordNet domains and ontologies. The various entry types are specified in the documentation and comprise Lexical Units (LU), Synsets as well as ontology terms and axioms (derived from SUMO and MILO). The Cornetto database also contains various semantic relations (vertical and horizontal), combinatorial relations, ontologies (top-level and domain), and morpho-syntactic information. All the relevant information on the sort of linguistic knowledge included in an entry is either given in the Cornetto documentation or a reference is cited.¹

The documentation clearly specifies the domain of the lexicon (general language) and the aim at the highest possible degree of coverage. In order to maximize coverage, the development team has chosen to use all data from both Dutch WordNet (DWN) and Referentie Bestand Nederlands (RBN) even though that meant including data that could not be manually verified. Good coverage is certainly essential for NLP applications provided that the quality of the data does not suffer.

Cornetto contains nouns, verbs and adjectives. The data was taken over from the sources used (DWN and RBN) and no new lemmas were added, but the combined and enhanced data in Cornetto now covers more general language. We are also given ideas for the potential applications of Cornetto in NLP. As it is a general purpose tool, the given list is not exhaustive and is not expected to be.

Great care has been taken to critically evaluate and if necessary correct the automatic alignments. One such means is the assignment of an alignment quality label (AQL) which indicates the strategy employed paired with a confidence indication (see pp. 23-26 in D-16). In total, 46% of all alignments have an AQL which roughly corresponds to the aimed at 50% of inspected alignments (see "Evaluation of Cornetto", p. 4). The majority of the remaining 54% of data are nouns solely occurring in DWN as a word. Of the 46% with an AQL, approximately 10% has been verified manually. Given these statistics and our manual validation, we think that the overall quality of the resulting lexical resource is warranted.

Also, the AQLs allow potential users to only choose a sub-part of the data if specific quality criteria have to be met. This added flexibility is a useful feature of the Cornetto database. Limiting the choice of data to entries with specific AQLs also presents a drawback, however: not all entries for a given lemma necessarily have the same AQL (see e.g. the entries for *sterretje*, *zullen*, *zeer*). In practice, this can mean that only one of three entries for a given lemma is taken into account which can lead to a very different data view.

¹ The main reference concerns more detailed information on the "Referentie Bestand Nederlands" which was used as one of the data sources.

3 Formal Validation

3.1 Validation of files

The XML files included in Cornetto are complete according to the specifications and do not present any problems.

3.2 Validation of tools

The main tool included in Cornetto is the DebVisDic server, a dictionary editor and browser. The Cornetto client has been realized as plug-in for web browsers (currently supporting Mozilla Firefox 3.x). There were some problems with the installation of the plug-in at first, but the support was excellent. The step-by-step instructions on the Cornetto website are easy to follow and using the tool once it has been installed is quite intuitive. No major problems with the Cornetto client were encountered during this evaluation.

4 Content Validation

4.1 Validation of Linguistic Correctness

Linguistic correctness was evaluated through data samples and manual inspection. The samples were checked manually using the DebVisDic server as an interface to search for lemmas and visualize the associated information. In certain cases, the XML files were searched directly with a script to identify candidate lemmas with specific properties.

4.1.1 Samples

In a first step, 100 lexical units from the core Cornetto were selected from the lists provided as an appendix to the general Cornetto documentation (deliverable D-16). These 100 lexical units were spread over the different parts-of-speech (POS) contained in Cornetto. Of the total of 119,000 lexical units that have been delivered in Cornetto, 72% are nouns, 14% verbs, 13% adjectives, and 1% adverbs. The 8500 lexical units that form the core and have been manually corrected show a similar distribution between POS: 77% nouns, 13% verbs, and 10% adjectives.² From these lists, 77 nouns, 13 verbs and 10 adjectives have been randomly selected for the current evaluation and then manually checked (see the complete lists in the appendix).

Additionally to the core data, 509 nouns and 618 verbs have been manually checked for equivalence relations, SUMO labels and domain labels during the Cornetto project. In turn, from these lists, 10 nouns and 10 verbs have been chosen and manually checked (see the complete lists in the appendix).

In a last step, 15 entries from the extended database have been randomly selected and evaluated by hand. Table 1 provides an overview of the number of manually verified entries per data source and POS.

² We did not find any mention of adverbs. They seem to not have undergone any manual validation or correction.

	<i>From Cornetto Core</i>	<i>From Cornetto with manually verified equival. relations</i>	<i>From Cornetto extended</i>	<i>Total</i>
<i>Nouns</i>	77	10	6	93
<i>Verbs</i>	13	10	4	27
<i>Adjectives</i>	10	0	2	12
<i>Adverbs</i>	0	0	3	3
Total	100	20	15	135

Table 1: Overview of number of selected lexical units per POS and per data source.

Both the core database (all manually verified) and the extended database have been included in the evaluation. Also, all different POS have been taken into account and all possible aspects of the work done during Cornetto (see list of checks performed below).

4.1.2 Manual Checks Performed

The following manual checks have been performed during the evaluation:

- Alignment, including AQL scores
- SUMO labels
- Domain labels
- Translation relations

With regard to the AQL scores, we examined whether they agree with our own evaluation as they are supposed to indicate the overall confidence in the alignment. A minor comment on AQL scores in general relates to the abbreviations used. In deliverable D-05 (p.12) it is stated that “the suffix of the AQL indicates the confidence of the mapping [...] The AQL D-75 thus means it is correct with a confidence of 75%”. It is a shame that the same principle has not been applied when naming the AQL D-58 and D-55. Their respective confidence in the alignment is 97% and 95%.

4.1.3 Findings of Manual Validation

In a first step, the qualitative aspect of the manual validation is reported on. To this end, the findings are listed per check performed. Also, suitable examples are given where appropriate to illustrate the problem found. At the end of the section, we summarize the overall evaluation for each check and add some general comments. Then, table 2 presents the quantitative results of the manual validation.

4.1.3.1 Alignment and AQL scores

- Status: For some lexical units that have been manually aligned the status does not reflect that fact: it is either empty or has another AQL assigned to it (see for example *set*, *tegenvoeter* or *dubben*).

- Extra lexical units: A number of manually checked words have unnecessary lexical units which could have been deleted during manual editing.³
- Wrong alignments: For some of the validated samples, wrong alignments have been found.
 - Looking at *curie*, we noticed that the link between the lexical unit entry for *curie:1*, *pauselijke regering* and the synset entry for *curie:1* is not correct. In fact, the synset entries for *curie:1*, *pauselijke regering* and *curie:3*, *eenheid van radioactieve straling* are equivalent, but should be different.
 - For *mastodont*, the alignment between lexical units and synsets has been reversed: lexical unit *mastodont:1*, *uitgestorven olifantensoort* should be aligned with synset *mastodont:2* and vice versa.
 - In the case of *γ*, the two synset entries (and therefore the links to the English WordNet) are equivalent whereas they should contain different information.

4.1.3.2 SUMO labels

- Wrong label: The attribution of a certain SUMO label leads to mistakes in the attribution of English equivalents. The SUMO label "+Bag" for *blaas:3*, *blaar* (in combination with the attributed WordNet domain labels) leads to wrong links with English (see below).

4.1.3.3 Domain labels

- Wrong domain: The addition of a wrong WordNet domain leads to incorrect equivalence relations.
 - The combination of SUMO label ("Bag") and WordNet domains (factotum, number) links *blaas:3*, *blaar* with the English equivalents *bag:1*, *bag:4*, *handbag:4* and *bag:5*, *bagful:1* whereas the Dutch example is "blazen op de huid".
 - For *crisis:1*, *kritieke situatie* it seems that the WordNet domain "law" is inappropriate as it results in the equivalence relation *martial law:1*.
 - In the example of *kit*, adding the WordNet domain "Factotum" for *kit:2*, *vulmateriaal* overgenerates and produces *hod:1* as an incorrect near synonym in English.

4.1.3.4 Translation (and other) relations

- Wrong equivalence relation: There are instances of manually verified words where the attributed equivalence relations are wrong.
 - The equivalence relation *blister:1* should be cited for *blaas:3*, *blaar* rather than for *blaas:2*, *holte*.
 - Most of the equivalence relations in the synset for *salon:4*, *samenkomst* (the only lexical unit connected to a synset) are actually more appropriate for *salon:1*, *kamer voor het ontvangen van gasten*, with the exception of

³ We do not refer to the dummy lexical units created during the alignment of synsets and lexical units from RBN, but to empty lexical units that are not necessary in any way.

the English *suite:4* which should be linked to *salon:2*, *bankstel*. This particular meaning seems to reflect the fact that RBN also takes into account Belgian usages, but DWN does not (as much).

- Wrong relation: The verb *dubben:3* has the SUMO labels “doubts” and “IntentionalPsychologicalProcess” as well as the English equivalents of doubting, but the only relation in Dutch is a hyperonym relation with *tapen*, *opnemen*.

4.1.3.5 General comments and summary of findings

We have encountered quite a number of typos and/or mistakes in the examples included with the lexical units (e.g. “geonden” instead of “gebonden”, “met dit gangetje komen we het* nooit”). It would obviously be a huge undertaking to correct them all, but it seems that for manually corrected or added material overall correctness should be warranted. Also, for spelling variants (think of *wc* and *WC*), it would have been nice to synchronize the numbering of entries.

In deliverable D-16 it is stated that “[d]ummy lexical units are created for those DWN word meanings that could not be mapped to RBN lexical units automatically” (p.26). For one of the manually evaluated words this has been forgotten: *bits* has two synset entries but only one lexical unit. In our understanding of the documentation, a dummy lexical unit should have been created automatically for the second synset.

There are also missing relations in synsets. For the word *sterretje*, there is no hyponym connection for *sterretje:klein hemellichaam* to *ster:1, gasvormig hemellichaam* even though the documentation on diminutives in deliverable D-01 (p.40) states that there should be.

The findings of the manual validation show that overall Cornetto is a useful database and presents the outcome of a lot of effort, even though there are also some remaining issues. The main point of critique is that no wrong alignments for edited entries in Cornetto (especially for manually aligned words) should have been found. Correcting these misalignments would increase the value of Cornetto. Most translation relations between Cornetto and the English WordNet are accurate and complete. There are, however, some instances where the equivalence relations are slightly off. Furthermore, adding AQL labels is a laudable initiative to reflect the confidence in the alignment. In practice, though, these labels will still have to prove their worth.

On a positive note, the word *economie* is a very nice example of the added value of the different kinds of information contained in Cornetto (SUMO labels, WordNet domains, etc.) and the way their interplay enhances the actual information. This example shows that Cornetto can be a very useful resource for NLP applications for Dutch.

Table 2 presents the quantitative view on the findings from the manual validation. The numbers are split according to the source of the data. We also include the total number of entries from Cornetto to better relate the findings on a small portion of the data to the database as a whole.

	<i>Total in Cornetto</i>	<i>Total checked</i>	<i>Problems found</i>
<i>Cornetto Core</i>	8500	100 (1%)	13 (13%)
<i>Cornetto with manually verified equiv. relations</i>	1127	20 (2%)	6 (30%)
<i>Cornetto extended</i>	109373	15 (0.01%)	0 (0%)
Total	119000	135 (0.1%)	19 (14%)

Table 2: Quantitative results of manual validation of Cornetto

As we see from table 2, only a fraction of the total data contained in Cornetto has been evaluated (0.1%). This glimpse, however, provides useful feedback and insights into remaining problems and the quality of Cornetto as a whole. The extended part of Cornetto does not present any problems and the automatic treatment of these entries is justified. The core Cornetto contains more polysemous data and is therefore also more error prone (during automatic processing as well as during manual editing). This is reflected in the number of problems found during the evaluation. Whether the problems still persisting in the data will have repercussions and how big those will be, depends for a major part on the application Cornetto is used for.

Overall Cornetto is a valuable resource for Dutch NLP and presents a wealth of information otherwise inaccessible. If extra funding can be found to perfect this resource, it could become even more valuable for the commercial and non-commercial NLP community in the Netherlands.

5 Conclusion

This evaluation shows that Cornetto is a useful lexical semantic database for Dutch of good quality and with useful features for many NLP applications. Regarding the documentation and the formal validation, there are no points of critique. With respect to the content, a few problems surfaced during the evaluation. Overall, however, Cornetto can definitely be seen as a successful project which resulted in a unique lexical semantic database for Dutch. Hopefully, Cornetto will prove its worth through the inclusion in numerous NLP applications for Dutch in the (not too distant) future.

6 Appendix

6.1 List of lexical units included in the evaluation from the core part of Cornetto

6.1.1 Nouns

a
aanleg
aarde
alt
assimilatie
bami
band
beharing
betaling
blaas
boog
cachot
college
crisis
curie
dek
dol
draak
dreutel
economie
eis
eten
film
flard
frisbee
gang
garde
geleding
gezicht
ham
hol
inhoud
juffertje
keet
kit
knoedel
knots
kunst
lexicon
licht
maal
mee
moer

naakt
net
omber
ontzetting
optie
paprika
piek
plas
ree
riet
ronde
salon
scheut
set
slinger
snede
staat
sterretje
strik
taak
TAP
toog
uitmonstering
uitzending
vaandel
vat
verzet
vreemdeling
WC
wereld
y
zee
zoon
zwik

6.1.2 Verbs

aarzelen
behouden
boeien
eten
gelden
groeien
hechten
maken
noemen
sluiten
trouwen
verliezen
zullen

6.1.3 Adjectives

aardig
bits
clean
druilerig
ernstig
fijn
goedkoop
los
vlug
zeer

6.2 List of lexical units included in the evaluation from the extended part of Cornetto with manually verified equivalence relations, SUMO labels and domain labels

6.2.1 Nouns

barenswee
credo
gom
kaffer
mastodont
onschuld
reserve
tegenvoeter
toevalligheid
zwetskous

6.2.2 Verbs

boekstaven
dubben
kleuren
poekelen
slaan
tanen
teloorgaan
verschieten
vlaggen
wiebelen

6.3 List of lexical units included in the evaluation from the extended part of Cornetto

6.3.1 Nouns

baby
interieur
leven
loon
vriend
weekend

6.3.2 Verbs

besnijden
doen
poetsen
tutten

6.3.3 Adjectives

tof
super

6.3.4 Adverbs

niet
nooit
toch