

Extended Cornetto

Cornetto Deliverable D-05

Version 1

September 2008



Project STE05039

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

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

Authors

Isa Maks and Piek Vossen, Vrije Universiteit Amsterdam



Table of contents

1	INTRODUCTION	3
2	EXTENDED CORNETTO	4
2.1	WORDNET EQUIVALENCE RELATIONS, DOMAINS AND SUMO TERMS.....	6
2.2	ALIGNMENT QUALITY LABEL	9



1 Introduction

The Cornetto project (STE05039) is funded by the Nederlandse Taalunie in the STEVIN framework (www.taalunieversum.org/stevin). The goal was to build a lexical semantic database for Dutch, covering 40K entries, including the most generic and central part of the language. Cornetto combines the structures of both the Princeton Wordnet and FrameNet for English, by combining and aligning two existing semantic resources for Dutch: the Dutch wordnet (Vossen 1998) and the Referentie Bestand Nederlands (Martin et al 1999). To create the initial Cornetto database, the word meanings in the Referentie Bestand Nederlands (RBN) and the Dutch WordNet (DWN) have been automatically aligned. The final Cornetto Database can be divided into two subsets:

- Core Cornetto - the part that has been manually checked, edited and confirmed after an initial set of automatic alignment procedures (appr. 8,500 lexical units and 3,000 words). For more details please refer to Cornetto Deliverable D-04.
- Extended Cornetto- the part of Cornetto that is not manually checked and covers approximately 89,000 words and 110,000 lexical units.

Table 1: Core and extended Cornetto

	Lexical units	Words
Core	8,500	3,000
Extended	110,000	89,000

The core Cornetto database is the part of the database that is manually revised and checked. The work involved:

1. Manually aligning the lexical units from RBN and DWN:
 - a. linking lexical units
 - b. splitting lexical units
 - c. merging lexical units
 - d. deleting lexical units



- e. adding lexical units as synonyms to existing synsets
2. Adding essential information to the lexical units involved:
 - a. morpho-syntax
 - b. semantics, e.g. resumes and case frames
 - c. examples
3. Adding essential relations to the synsets:
 - a. hyperonym relations
 - b. other significant relations
4. Manually verifying or creating the mapping to English Wordnet2.0
5. Manually verifying or creating the mapping to the SUMO ontology
6. Manually verifying or creating the mapping to WordNet Domains

In this deliverable, we describe the work and decisions made for the work on Extended Cornetto.

2 Extended Cornetto

The extension consisted of including the maximal amount of data in the resulting resource. This means that in principle all entries from DWN and RBN are included in the resource. Likewise, we are going to deliver more data than originally promised (over 91K entries instead of 40K entries). The main motivation for this is that coverage of a lexical resource is an important aspect for NLP applications. Furthermore, we expect to further improve the data in future releases and now provide the basic administrative units for making these improvements and enrichments. In many cases, the additional data are very specific, occurring either only in RBN or in DWN. In that case, alignment is not possible but the basic information provided in either resource is still valuable. The price for delivering more data than we promised is that we have not been able to manually verify and correct all the data. Furthermore, certain relations and other important data are not given for these words. They may lack a hyperonym relation, morpho-syntactic details for the Lexical Units, definitions or an equivalence relation to English WordNet.

The original DWN database was derived from the Van Dale Lexicographic Information System (VLIS). This is a much larger repository of which a portion was published in DWN. However, since the RBN entries may match VLIS entries that are not part of the DWN selection, we have extended the coverage of DWN with remaining VLIS entries. To give an



impression of the data, the next table (Table 2) gives a global overview of the total number of form units (FUs) that are on both original database and their potential overlap. As can be seen, Cornetto can at most have 37,956 entries if we would restrict ourselves to words that occur in both resources, assuming the ideal case that all meanings can also be mapped one-to-one. By taking the union of the words, we get a much richer database (more than 96K entries), part of which is very rich, manually confirmed and corrected and aligned, part of which has partial and not-confirmed data.

Table 2: Form units (FUs) in the original source databases

Both in RBN and DWN	37.956
Total number FUs in Cornetto	96,659

This gives us a larger volume of lexical data to start with, with the drawback that a proportion of the data is not validated. The large data set has been further processed in the following way. We first assumed that the most critical cases have been covered in the work for the core Cornetto. There are two major risks for the remaining data:

1. The alignment between RBN and DWN is not correct;
2. The mapping with the English wordnet is not correct and through this mapping also the labeling with SUMO and Wordnet Domains.

We decided to tackle these problems independently and in different ways. The alignment of RBN and DWN was revised on the basis of a subset of heuristics that can be applied easily to large amounts of words with low polysemy. Furthermore, we derived a so-called Alignment Quality Label for all the aligned cases so that users of the database have an idea about the reliability of the data.

The matching with English wordnet and via that to SUMO and Wordnet Domain was handled by manually revising all nouns and verbs with too many equivalence relations, by ignoring remaining vague mappings, and by independently assigning WordNet domain labels.

Both approaches are explained in the next subsections in more detail.



2.1 Wordnet Equivalence relations, Domains and SUMO Terms

Cornetto synsets have translation equivalence relations to Princeton Wordnet2.0 (PWN2.0). Through these relations current and future PWN extensions can be imported into Cornetto¹. Obviously, the quality of these imported extensions of the Cornetto Wordnet heavily depends on the quality of the equivalent mappings. Therefore, one of the goals of the Cornetto project was to correct and revise the equivalence relations to PWN.

Pos	Synsets manually revised	Number of synsets
nouns	Nr of equivalence relations > 7	509
verbs	Nr of equivalence relations > 4	618

Table 3: Revisions of equivalence relations

For these words we selected or created the correct equivalence relation(s) to PWN2.0, we revised the SUMO mapping and the domain relations. In many cases also the other meanings of these words have been checked. We did not consider the RBN-DWN alignment since these are checked through other selections (see next section).

If synsets still got 4 or more equivalence relations after this manual revision, we have removed the imported SUMO labels and domain labels. In case of many mappings, these are too unreliable and they can be derived from the hyperonyms of the synsets.

For the Wordnet domain labels, we created an alternative method. Instead of importing the domains through the equivalence relation, we aligned the domain labels in the VLIS database directly with the domain labels in Wordnet domain. The alignment is shown below. If a synset had no Wordnet Domain labels, we thus derived them through the mapping table by inserting a Wordnet Domain for every matching VLIS domain that is associated with a synset.

¹ Like for example , Wordnet Domains, SUMO, Wordnet Affect, Sentiwordnet



AARDR	geography;anthropology	GELUIDSL	music;acoustics
ADEL	heraldry	GENOTM	gastronomy;pharmacy
AGR	agriculture	GEOL	geology
ANAT	anatomy	GEREED	building_industry
ANTR	anthropology	GESCH	history
ARB	economy	GOLF	golf;play
ARCH	building_industry	GRAF	publishing
ARCHEOL	archaeology	HAND	commerce
ASTROL	astrology	HERALD	architecture;philology;heraldry
ASTRON	astronomy	HORECA	commerce;building_industry
ATLET	athletics;sport	HOUTIND	botany;building_industry
AUDIO	electricity;radio	HUIS	furniture
AUTOSP	sport	IND	industry
BANKW	banking	JACHT	hunting
BEELDH	sculpture	JUR	law
BELAST	tax	KARAK	psychology;person;quality
BESTUUR	enterprise	KLED	fashion
BEURS	exchange	KUNST	art;dance;drawing;painting;philately;music;photo
BIBL	philology		graphy;plastic_arts;jewellery;numismatics;sculpture;theatre
BIJB	religion	KUNSTST	chemistry
BIJENT	zoology;zootechnics	KV-ADEL	heraldry
BILJART	play;sport	KV-BIJBB	religion
BIOCHEM	chemistry;biology	KV-BIJBF	religion
BIOL	biology;zoology;botany;physiology;anatomy	KV-DAGEN	time_period
BK	art	KV-EENHE	metrology;physics
BOEKH	administration;book_keeping	KV-ELEME	chemistry
BOHAND	publishing	KV-EVANG	religion
BOSB	agriculture;botany	KV-GEOGR	geography
BOUW	building_industry	KV-INWON	administration;geography
BUR	administration	KV-KAART	card
COMP	computer_science	KV-KLEUR	color;quality
COSM	body_care	KV-NAAMV	grammar
CUL	gastronomy	KV-PREDI	heraldry
DIERGEN	medicine;physiology;zoology	KV-ROMCI	number
DIERK	zoology	KV-SCHOO	school
DRNAAM	zoology	KV-STERB	astrology;astronomy
DRUK	publishing	KV-TALEN	linguistics
EC	economy;money	KV-TIJD	metrology;time_period
ELEKTR	electricity	KV-WINDR	geology
FARM	pharmacy	KV-ZODIA	astrology;astronomy
FIL	philosophy	LANDB	agriculture
FILM	cinema;theatre	LEDERIND	anatomy;zoology
FIN	economy;money;exchange	LETTERK	literature
FOTO	photography	LOG	philosophy;mathematics
FRUITT	botany	LUCHTV	transport;aeronautic
FYSIOL	physiology;anatomy	MAT	chemistry;industry;botany;zoology
GEBOUW	building_industry		

D05 Extended Cornetto



MED	medicine;anatomy;dentistry;pharmacy;psychiatry ;radiology;surgery	SCHIETSP	archery;sport
MED-B	medicine	SEKS	sexuality
MED-D	medicine	SOC	anthropology;sociology
MED-I	metrology;medicine	SOVERZ	economy
MED-T	medicine	SPEELG	play
MEDIA	publishing;telecommunication	SPEL	play
MET	chemistry;industry	SPOORW	transport
METEO	meteorology	SPORT	sport;badminton;baseball;basketball;cricket;footb all;golf;rugby;soccer;table_tennis;tennis;volleyball;cycling;s kating;skiing;hockey;mountaineering;rowing;swimming;sub ;diving;racing;athletics;wrestling;boxing;fencing;archery;fis hing;hunting;bowling
METR	mathematics;metrology;time_period;number	SPORTVIS	fishing
MIJNB	industry;chemistry	STAT	mathematics,statistics
MIL	military	TAALK	linguistics
MILIEU	town_planning;ecology	TECHN	mechanics;metrology
MOLEN	building_industry;industry	TELECOMM	telecommunication;telegraphy;telephony;radio
MUNT	money	TENNIS	table_tennis;tennis
MUZ	music	TEXT	industry
MYTH	mythology	THEAT	theatre;art
NAT	physics;chemistry	THEOL	religion,theology
NIJVERH	artisanship	TOER	tourism;town_planning
OB	administration;politics	TOPOGR	geography
OCC	religion;occultism	TUINB	botany;agriculture;zootechnics
OLIE	chemistry	UURW	metrology
ONDERW	pedagogy;university;school	VECHTSP	boxing;sport
OPTICA	metrology;quality;physics	VEET	zoology;zootechnics
OPTINS	optics	VERF	color;painting;drawing
PAARDESP	racing;sport	VERK	transport;auto;cycling;tourism
PAPIER	building_industry;industry	VERZ	insurance
PARK	geography	VIS	fishing;merchant_navy
PATHOL	medicine;physiology	VOERT	transport;auto
PEDAG	pedagogy;school	WAPEN	military;atomic_physic
PLANTK	botany	WATERSP	diving;swimming;sport;tourism
PLNAAM	botany	WET	pure_science;applied_science;and;computer_sci ence;pure_science;university
POL	politics	WIELersp	cycling;sport;transport
POLIT	law	WIJNB	agriculture;gastronomy;industry
PROT	religion	WINTERSP	skiing;skating;sport;tourism;transport
PSYCH	psychology	WISK	mathematics;geometry;number
PTT	post;telecommunication	WON	furniture;building_industry;fashion
R-K	religion;roman_catholic	WWB	transport;building_industry
RADIO/TV	tv;telecommunication	ZEILV	merchant_navy;tourism;transport;sport
REL	religion		
RO	town_planning;transport		
RUIMTEV	astronautics		
SCHEEPV	merchant_navy		
SCHEIK	chemistry		
SCHERM	fencing		



2.2 Alignment Quality Label

The alignment of the RBN Lexical Units with the DWN synsets of the Extended Cornetto has been performed in two steps:

- Automatic alignment

Automatic Alignment of all words has been done following a number of mapping strategies which use the overlapping features between the two resources². The result is a mapping table (cid) representing the alignment, where each word meaning in RBN is mapped to one of more word meanings of the same word (with the same part-of-speech) in DWN. The mappings are scored; these scores can be found in the cid table (cid_score).

- Automatic (re-)alignment

Finally some extra alignments strategies have been carried out with regard to:

- (1) monosemous and bisemous words and
- (2) words with a considerable overlap between the RBN and DWN definition

The different alignment mappings are assessed by determining accuracy at a sample of 100 alignments per type. This resulted in an alignment quality label (AQL) which is stored in the in the field *status* of the cid table.³ The users of the database can use these labels to get an indication of the core-Cornetto database that has both rich lexical unit and synset data, and the extended Cornetto database that is partially represented. For different purposes, different selection of the data can be made.

² For more information see CornettoDeliverableD02

(<http://www.let.vu.nl/onderzoek/projectsites/cornetto/deliver.html>)

³ For more information about the design of the database, please refer to Cornetto Deliverables D-02 and D-16

**Table 4: Alignment quality labels for Core and Extended Cornetto**

AQL	Accuracy	Pos	Number of alignments	Number of words
Core Cornetto Manual	100%		8499	3050
		A	1071	
		V	1612	
		N	5816	
Extended Cornetto M-97	97%		25488	25488
		A	4065	
		V	2885	
		N	18510	
Extended Cornetto B-95	95%		4986	2493
		A	2714	
		V	1672	
		N	600	
Extended Cornetto BM-90	90%		4479	n.a.
		A	752	
		V	938	
		N	2782	
Extended Cornetto Resume-75	75%		1052	n.a.
		A	121	
		V	851	
		N	80	
Extended Cornetto D-75	75%	N	2092	n.a.
Extended Cornetto D-58	97%	V	777	n.a.
Extended D-55	95%	A	173	n.a.
Extended Cornetto NO-STATUS	n.a.		56122	n.a.
		A	6350	
		V	5358	
		N	43818	
		Adverb	484	



	Pronoun	13	
	Interjection	80	
	Conjunction	6	
	Numeral	6	
	Preposition	7	

- **Manual** includes:
- frequent verbs (CGN -freq > 250) ⁴
- frequent adjectives (CGN-freq > 300)
- highly polysemous nouns (number of senses -> 3 in Cornetto after the alignment)
- highly polysemous adjectives (number of senses > 3 in RBN or DWN)
- highly polysemous verbs

M-97 includes:

- monosemous words, i.e. words that have in both DWN and RBN one sense only

B-95 includes:

- bisemous words, i.e. words that have in both DWN and RBN two senses. The first sense and the second sense of RBN are aligned with the first sense and the second sense of DWN, respectively,

BM-90 includes:

- words that have one sense in RBN and two senses DWN or viceversa. The first sense of the bisemous word is aligned with the only sense of the monosemous word. The second sense remains unaligned.

Resume -75 includes:

- Lexical Unit – Synset alignments based upon a substantial overlap between the RBN and DWN definitions and synonyms.

D-75 includes:

- Noun Lexical Unit – Synset alignments that have a automatic alignment score > 30%.

D-58 includes:

- Verb Lexical Unit – Synset alignments that have a automatic alignment score > 30%.

D-55 includes:

- Adjective Lexical Unit – Synset alignments that have a automatic alignment score > 30%.

No-status

- Lexical Unit- Synset alignments between a DWN synset and a dummy Lexical Unit.

⁴ CGN = Corpus Gesproken Nederlands



Dummy lexical units are created for those DWN word meanings that could not be mapped to RBN lexical units automatically. These LUs have only minimal morpho-syntactic information and no semantics, pragmatics and examples. As explained above, the suffix of the AQL indicates the confidence of the mapping based on samples of 100 records differentiated per part-of-speech. The AQL D-75 thus means it is correct with a confidence of 75%.