



Manual for Semantic Annotation in the DutchSemCor project

Guidelines for creating a Dutch corpus annotated with word senses from the
Cornetto database

Prof. Dr. Piek Th.J.M. Vossen and Attila Görög MA

Computational Lexicology & Terminology Lab
Vrije Universiteit Amsterdam

a.gorog@let.vu.nl

p.vossen@let.vu.nl

Content

1. Introduction

2. Methodology: different cycles of the annotation process

3. Tools used in the DutchSemCor project

3.1. Semantic Annotation Tool (SAT)

3.2. Cornetto Database

3.3. Snippet Tool

3.4. Log file

4. Conclusion

5. Appendix 1, 2, 3

1. Introduction

In natural language processing, word sense disambiguation (WSD) is the problem of determining which "sense" (meaning) of a word is activated by the use of the word in a particular context, a process which appears to be largely unconscious in people. WSD is a natural classification problem: given a word and its possible senses, as defined by a dictionary, classify an occurrence of the word in context into one or more of its sense classes. The features of the context (such as neighbouring words) provide the evidence for classification.

A famous example is to determine the sense of the word "pen" in the following passage (Bar-Hillel 1960):

Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

WordNet lists five senses for the word 'pen':

1. pen — a writing implement with a point from which ink flows.
2. pen — an enclosure for confining livestock.
3. playpen, pen — a portable enclosure in which babies may be left to play.
4. penitentiary, pen — a correctional institution for those convicted of major crimes.
5. pen — female swan.

Humans don't have any difficulty when reading this sentence and trying to decipher the meaning of the word 'pen'. A computer program, however, struggles when a choice needs to be made among the five distinct senses. For this reason, automatic WSD is one of the major challenges of AI.

In order to reach acceptable results in automatic WSD, computer applications need to be trained by humans with the help of large sense-tagged viz semantically annotated corpora. The goal of the DutchSemCor project is to create such a corpus. Our aim is to deliver a one-million word Dutch corpus that is fully sense-tagged with senses and domain tags from the Cornetto lexical database (Vossen 2006 and Vossen et al. 2007, 2008) by autumn 2012.

250K words of this corpus will be manually tagged. The remainder will be automatically tagged using three different word-sense-disambiguation systems (WSD), and will be validated by human annotators. The corpus data will be based on existing corpus material collected in the projects CGN, D-CoI and SoNaR. These corpora have already been parsed and tagged in previous projects and will be extended where necessary by free examples from the Internet or snippets from the Open Taal Project¹ in order to find sufficient examples for the word senses that are less frequent and do not appear in the above corpora.

The resulting corpus for which we aim to offer the same balance in types of text as the corpora mentioned above, will be extremely rich in terms of lexical semantic information. Its availability will generate many new lines of research and technology developments in the Dutch NLP community. The resulting data will be, for instance, applicable in the fields of cognitive science, (psycho-) linguistics, language learning and language teaching, semantic web applications, information retrieval, machine translation, text mining, and document interpretation. We foresee that the corpus will create new directions of research and technology development in Dutch computational and corpus linguistics.

The project DutchSemCor is funded by the NWO and carried out by a consortium of three universities: 1. *Language and Communication*, Faculteit der Letteren, **Vrije Universiteit**

¹ <http://www.opentaal.org/>

Amsterdam; 2. *Tilburg centre for Creative Computing*, Faculteit Geesteswetenschappen, **Universiteit van Tilburg**; 3. *ISLA*, Faculteit der Natuurwetenschappen, Wiskunde en Informatica, **Universiteit van Amsterdam**.

The aim of this document is to provide a methodology for the semantic annotation process in the DutchSemCor project. It is also a guide for the project leaders, the editors and the student assistants being involved in this project. Besides certain basic linguistic rules concerning semantic annotation and the usage of the Cornetto database, we would also like to give some information on the project, on relevant websites and literature as well as a couple of tips and advices on the usage of available software and corpus data.

In **section 2**, our general methodology is explained: how do we reach an Inter-annotator Agreement (IA) of minimum 80%. In which phases will we proceed in annotating the corpus examples. How do we make sure that the annotated data will be diverse enough both semantically and syntactically. In **section 3**, we will provide more details about the used software, the Semantic Annotation Tool (**section 3.1**), the Cornetto lexical database (**section 3.2**), the Snippet Tool (**section 3.3**) and the Log file together with the Log Analyser (**section 3.4**) as well as some basic strategies which are indispensable in the annotation process. Since our aim is not only to provide annotated data but also to, if necessary, correct and edit Cornetto, we will give a short summary of the revision and editing methodology of the Cornetto entries in **section 4**.

2. Methodology: different cycles in the annotation process

In order to reach an IA of minimum 80%, we will implement the following working cycle divided into two phases. During the different steps a Log file will be consulted which can be downloaded through a link [url²](#) at any moment (see an example of the Log file in **3.4**). These are the different steps on which we will elaborate more thoroughly in the following sections:

- Pre-processing the Cornetto data
- Manual annotation 1
- Discussion
- Manual annotation 2
- Post-editing Cornetto data

Pre-processing the Cornetto data

Before starting the annotation, the Cornetto data needs to be inspected by the coordinator of the project and, if necessary, the entries need to be corrected. Also a word list is to be prepared. Every two weeks a new word list of approx. 200 words are processed by 8 annotators (4 couples). More information about the pre-editing and the post-editing of the Cornetto entries can be found in **section 4**.

Manual annotation 1

Two annotators (**A1 + A2**, altogether four couples) receive the same list of words to annotate. They start annotating the lemmas in reverse order using the ‘free option’ in the Semantic Annotation Tool (SAT).

² <http://ilk.uvt.nl/~mvgompel/dsclog.zip>

G10		sectie									
	A	B	C	D	E	F	G	H	I	J	
1		Jonica	Lisanne	Dieke	Marlisa	Laura	Anneleen	Gratia	Wilma		
2	1	actie	bron	groei	klap	minuut	provincie	speler	verlichting	free	
3	2	afzet	chili	grootmeester	knop	molen	regering	station	vis	free	
4	3	atmosfeer	commissaris	hof	kunst	motief	resolutie	straal	vondst	free	
5	4	autoriteit	commissie	hoofd	kwaliteit	neus	richtlijn	strip	voordeel	free	
6	5	bad	correctie	impuls	legende	nood	rijkdom	systeem	vraag	free	
7	6	bedrijf	cultuur	ingang	leger	nummer	ruimte	techniek	vrouw	free	
8	7	begrip	departement	inhoud	leiding	onderwerp	school	term	waarde	free	
9	8	behandeling	fout	instructie	lengte	oplossing	schrift	thee	water	free	
10	9	belasting	gegeven	kaak	lijst	uitwerking	sectie	tip	wereld	free	
11	10	beroep	gemeenschap	kamer	man	partner	sector	titel	werking	free	
12	11	besluit	gemeente	kant	markt	politiek	serie	toets	zee	free	
13	12	bord	geschiedenis	karakter	mens	praktijk	sneeuw	uitvoer	zicht	free	
14	13	brief	gewas	kikker	middel	project	soort	vader	zoon	free	
15	14	bron	actie	klap	groei	provincie	minuut	verlichting	speler	budy unique	
16	15	chili	afzet	knop	grootmeester	regering	molen	vis	station	budy unique	
17	16	commissaris	atmosfeer	kunst	hof	resolutie	motief	vondst	straal	budy unique	
18	17	commissie	autoriteit	kwaliteit	hoofd	richtlijn	neus	voordeel	strip	budy unique	
19	18	correctie	bad	legende	impuls	rijkdom	nood	vraag	systeem	budy unique	
20	19	cultuur	bedrijf	leger	ingang	ruimte	nummer	vrouw	techniek	budy unique	
21	20	departement	begrip	leiding	inhoud	school	onderwerp	waarde	term	budy unique	
22	21	fout	behandeling	lengte	instructie	schrift	oplossing	water	thee	budy unique	
23	22	gegeven	belasting	lijst	kaak	sectie	uitwerking	wereld	tip	budy unique	
24	23	gemeenschap	beroep	man	kamer	sector	partner	werking	titel	budy unique	
25	24	gemeente	besluit	markt	kant	serie	politiek	zee	toets	budy unique	
26	25	geschiedenis	bord	mens	karakter	sneeuw	praktijk	zicht	uitvoer	budy unique	
27	26	gewas	brief	middel	kikker	soort	project	zoon	vader	budy unique	
28											
29											

Wordlist

Note that the annotators are free to choose or ignore certain examples in the corpus when annotation lemmas from the first half of the list ('free option'). When the first part of the words in the list is annotated using the 'free option' by **A1**, he/she continues with the second part using hereby the 'Buddy function'. This way the examples are restricted to the ones tagged by **A2**. Of course, the two annotators cannot see each others' annotations i.e. the senses they linked to a certain example. The other annotator (**A2**) does the same, which results in optimal overlap in the number of double-annotated examples.

When annotators encounter mistakes in Cornetto or when they are not sure whether the sense divisions are correct, they fill in a digital form available on the Project Wiki-site to report problems. The coordinator of the project reacts to these problems by correcting the Cornetto entries or by clarifying vague sense. All the questions and reactions are documented this way in an excel sheet together with the changes made by the coordinator.

probleemwoorden

Vul de woorden in waar je problemen mee hebt. Graag alle velden invullen als dat kan.

***Vereist**

Woord *

Betekenis

Type probleem

☐ ik snap de betekenis(sen) niet
☐ er ontbreekt informatie (voorbeeld, synoniem etc.)
☐ de informatie die in Cornetto staat klopt niet
☐ ik kan geen voorbeeld vinden voor X betekenis
☐ overig

Verdere uitleg

Digital form for reporting lexical problems

probleemwoorden				
Woord	Betekenis	Type probleem	Verdere uitleg	Oplossing
inleiding		overig	Zou je even kunnen controleren of ik (Gratia) bij 'inleiding' echt alle tags heb verwijderd? Dan kan ik dit woord opnieuw taggen.	ja, allemaal verwijderd.
serie		overig	waar valt 'serie' in de betekenis van 'soort', zoals 'BMW-5 serie' of 'de Serie A' (de voetbalcompetitie) onder? Het is een soort reeks, dus vallen deze voorbeelden onder de 1e betekenis?	Bet. 2. = alleen voetbal of andere radioprogramma's Bet. 4 = heel andere voorbeelden zou ik onder
resolutie		overig	Bestaat er een scherp onderscheid tussen betekenis 1 (overheidsbesluit) en betekenis 2 (conclusie van een vergadering)? Beide zijn een besluit, alleen de één is op het gebied van de overheid en de ander op de rest (bijv. de VN). Is het nodig om dit als twee losse betekenissen te zien?	er is geen scherp onderscheid vergadering (conclusie). Eerste verder... Ik zou zeggen gebruik voorbeelden niet duidelijk zijn.
Er ontbreken evenwel nog bij mij (kij de				

Overview of problems with descriptions and solutions

Discussion

Every week a file with an overview of the Inter-annotator Agreement is uploaded to the Project Wiki. The students can consult the file and when necessary discuss the problems with each other. There is also an on-line forum on which annotators can discuss project-related issues.

discussiegroepen

Startpagina
Nieuw sinds laatste keer: 19 berichten

Welkom op de discussiegroep van DutchSemCor VU.

Een groep voor belangrijke postings, data voor vergaderingen en discussies.
Voor meer informatie ga naar de projectsite: <http://sites.google.com/site/dutchsemcor/>

[welkomstbericht bewerken]

Discussies 7 van 238 berichten [alles weergeven »](#)

[Eerste meeting](#)
Door - 22 jul - 4 auteurs - 4 antwoorden

[DutchSemCor etc.](#)
Door - 20 jul - 7 auteurs - 13 antwoorden

[FW: \[DutchSemCor:963\] snippet tool](#)
Door - 12 jul - 1 auteur - 0 antwoorden

[woordform-lemma probleem](#)
Door - 2 jul - 1 auteur - 1 antwoord

[Kort overzicht van dscvu@googlegroups.com - 1 bericht in 1 onderwerp](#)
Door Fons Laan - 16 jun - 1 auteur - 0 antwoorden

[Prijsvraag: Terminologie op een bierviltje](#)
Door - 15 jun - 1 auteur - 0 antwoorden

[er is er 1 jargi hoera hoera!!](#)
Door - 14 jun - 4 auteurs - 13 antwoorden

Startpagina

Discussies
Leden
Pagina's
Bestanden

[Info over deze groep](#)
[Mijn lidmaatschap bewerken](#)
[Instellingen voor discussiegroep](#)
[Taken voor beheerders](#)
[Leden uitnodigen](#)

Info over groep
Leden: 17
Taal: Nederlands
Categorieën van de discussiegroep: Niet ingedeeld
[een categorie toevoegen](#)
[Meer informatie over discussiegr](#)

In order to ensure that all problems and unanswered questions are solved, and for the continuity of the project, we also organize a central discussion every two weeks. During this meeting, we account for the differences and in some cases the sense division of Cornetto-entries will need to be changed. Also, students can give feedback on the tools and share their knowledge with each other.

Manual annotation 2

After the discussion, words with low IA have to be re-annotated until an IA of 80 % or more reached.

Post-editing the Cornetto data

Based on the data and feedback provided by the annotators during the different discussions and cycles, the Cornetto entries will be corrected if necessary. More information about pre-editing and post-editing the Cornetto entries can be found in **section 4**.

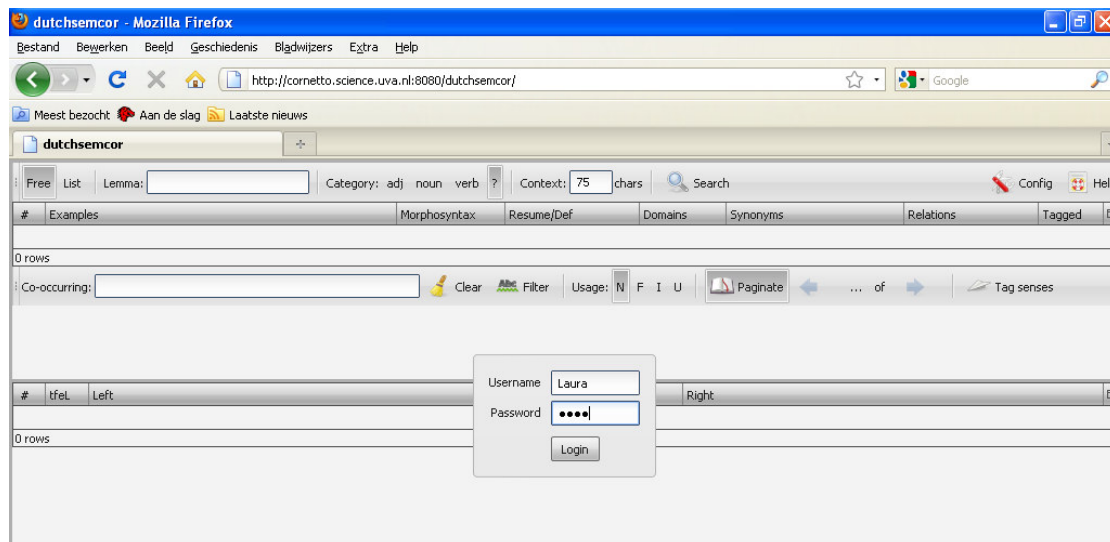
3. Tools used in the DutchSemCor project

3.1 Semantic Annotation Tool (SAT)

Our semantic annotation Tool provides human annotators with an ergonomic and easy to use web-based environment in which an optimal result can be reached for computer-aided semantic annotation. The SAT gives access to the Cornetto database (*Lexical data*) and to text fragments from the reference corpus (*Corpus data*). Cornetto is a semantically rich lexical database which contains the *Dutch WordNet*, the *RBN* (*Referentie Bestand Nederlands*, a Dutch lexicon developed by the Vrije Universiteit Amsterdam) and is also enriched with other semantic layers (*Wordnet Domains* and the *SUMO* ontology). In **section 3.2**, you can read more about the different parts of *Cornetto*.

In what follows we will describe a model annotation process illustrating the different steps with screenshots from the SAT.

1. On the login page the annotator enters the user name and the pass word received previously from the moderator.



The screenshot displays the 'dutchsemcor' application interface, divided into two main sections: 'Lexical data' (top) and 'Corpus data' (bottom).

Lexical data section: This section shows a table of lemmas. The search criteria are 'Free List', 'Lemma: beeld', 'Category: adj noun verb ?', 'Context: 75 chars'. The table has columns: #, Examples, Morphosyntax, Resume/Def, Domains, Synonyms, Relations, and Tagged. The data is as follows:

#	Examples	Morphosyntax	Resume/Def	Domains	Synonyms	Relations	Tagged
1	een marmeren/bronzen beeld	n-het-t	beeldhouwwerk	sculpture		afbeelding plaatje	0
2	een wazig beeld	n-het-t	plaatje	factotum		afbeelding plaatje	0
3	iemand een beeld van iets geven	n-het-t	voorstelling	factotum	voorstelling	weergave	0
4	een beeld van [een huisje]	n-het-nit	iets moois	factotum		iets	0
5	een treffend beeld	n-het-t	symbool	factotum psych	voorstelling plaatje	gedachte	0

Corpus data section: This section shows a table of text fragments. The search criteria are 'Free List', 'Lemma: beeld', 'Category: adj noun verb ?', 'Context: 75 chars'. The table has columns: #, tfeL, Left, Sense, Word, and Right. The data is as follows:

#	tfeL	Left	Sense	Word	Right
1	hcsillyd	Leopold III ' hangt zijn oudste dochter , prinses Esmeralda , een idyllisch		beeld	op van haar jeugd jaren in Argentuul en de familie bonheur . Wat hij in h
2	teh nek	in het bezette België , tegen het advies van zijn ministers in , zouden het		beeld	van de vorst zwaar aantasten . En na zijn huwelijk met Lillian Baels tijdens
3	seicerp	' Misschien was het zelfs al vroeger . Ik heb geen precies		beeld	van het moment waarop ik me ervan bewust geworden ben , dat mijn vader e
4	hcsirof	Het is trouwens niet aan zijn dochter om van hem een politiek of historisch		beeld	op te hangen , want dat zou niet objectief zijn . Bovendien heb ik die peri
5	teh . d	' Hij was een veelzijdige en een rijke persoonlijkheid . Het		beelden	van de getroumteerde koning ken ik niet . Ik heb hem getoond zoals hij is
6	gnittez	openbaar aanklager Carla Del Ponte liet tijdens zijn openingsuitenzetting		beeld	zien van de verschillende Joegoslavische oorlogen . Het beschikt ook over d
7	nee eic	nomen . Nu gaat ook de conjunctuurbarometer van de Nationale Bank , die ee		beelden	geeft van de verwachte economische ontwikkeling , duidelijk verder omhoog .
8	ni drew	ewerkt aan een klaardere ruimtelijke inrichting . Die transformatie werd in		beelden	gebracht door kunstenaar Candida Höfer , gespecialiseerd in het fotografer
9	ekrets	en voelen , is het samenbrengen van hyperindividueel geformuleerde , sterke		beelden	van universele emoties , denkbeelden en energieën , verbonden met hetgeen
10	nee si t	uwen uit been of geschilderd op rotsen . Het oudste bekende artefact is een		beeldje	van een mens met hondenkop dat 32.000 jaar oud zou zijn . Voorbeelden van
11	po ned	sselse parket vervolgde de veehandelaars nadat ze geïdentificeerd werden op		beelden	die de dierenrechtenorganisaties Gaia en Animaux en Péril stiekem gedraaid
12	ed . th	stiekem gedraaid hadden in november 2000 op de veemarkt van Anderlecht .		beelden	ontlokten toen een storm van protest bij de publieke opinie en zetten het p
13	ed tad	ertoe aan een onderzoek in te stellen . De rechter oordeelt vandaag dat de		beelden	een inbreuk vormen op de privacywet en daardoor een onrechtmatig bewijs zi
14	deog n	erspiegelt die een bedrijf aan zijn klanten aanrekent , geeft hij een goed		beeld	van de tarieven die een communicatieconsultant als Slangen & Partners (
15	ne , vt	rouwens leeg stonden . Helaas was de pers erbij , met fotografen en tv , en		beelden	van stuk huisraad en speelgoed en hullende peuters tussen het puin gingen d

Two parts of the SAT: the lexical data (with semantic and morfosyntactic information on lemmas) and the corpus data (with text fragments from the reference corpus)

2. After having logged in the annotator can type a lemma or automatically receives a word to align.

The screenshot shows the 'dutchsemcor' application running in a Mozilla Firefox browser window. The address bar shows the URL 'http://cornetto.science.uva.nl:8080/dutchsemcor/'. The interface is identical to the one shown in the previous screenshot, displaying the 'Lexical data' and 'Corpus data' sections. The search criteria are 'Free List', 'Lemma: beeld', 'Category: adj noun verb ?', 'Context: 75 chars'. The 'Corpus data' section shows 15 rows of text fragments, each with a 'tfeL' identifier, a 'Left' text fragment, a 'Sense' column, a 'Word' column, and a 'Right' text fragment.

3. The SAT searches for the lemma in the lexical data base and in the corpus database and shows the available data. Based on the Cornetto data (definitions, examples, grammatical and semantic information), human annotators are asked to link corpus examples to Cornetto-senses. For the purpose of targeted tagging, all occurrences of a word are displayed in a KWIC index (targeted tagging) and interfaced with the meaning specification in the Cornetto database.

dutchsemcor - Mozilla Firefox

Bestand Bewerken Beeld Geschiedenis Bladwijzers Extra Help

http://cornetto.science.uva.nl:8080/dutchsemcor/

Meest bezocht Aan de slag Laatste nieuws

dutchsemcor

Free List Lemmas: beeld Category: adj noun verb ? Context: 75 chars Search

#	Examples	Morphosyntax	Resume/Def	Domains	Synonyms	Relations	Tagged
1	een marmeren/bronzen beeld	n-het-t	beeldhouwwerk	sculpture		afbeelding plaatje	0
2	een wazig beeld	n-het-t	plaatje	factotum		afbeelding plaatje	0
3	iemand een beeld van iets geven	n-het-t	voorstelling	factotum	voorstelling	weergave	0
4	een beeld van [een huisje]	n-het-nt	iets moois	factotum			0
5	een treffend beeld	n-het-t	symbool	factotum psych	voorstelling plaatje	gedachte	0

5 rows

Co-occurring: Clear Filter Usage: N F I U Paginate 1... 50 of 1920 Tag senses

#	tfel	Left	Sense	Word	Right
1	hcsilyd	Leopold III 'hangt zijn oudste dochter, prinses Esmeralda, een idyllisch		beeld	op van haar jeugd jaren in Argentuul en de familie bonheur. Wat hij in h
2	teh nex	in het bezette België, tegen het advies van zijn ministers in, zouden het		beeld	van de vorst zwaar aantasten. En na zijn huwelijk met Lilian Baels tijdens
3	seicerp	'Misschien was het zelfs al vroeger. Ik heb geen precies		beeld	van het moment waarop ik me ervan bewust geworden ben, dat mijn vader e
4	hcsirob	Het is trouwens niet aan zijn dochter om van hem een politiek of historisch		beeld	op te hangen, want dat zou niet objectief zijn. Bovendien heb ik die peri
5	teH . d	'Hij was een veelzijdige en een rijke persoonlijkheid. Het		beeld	van de getourmenteerde koning ken ik niet. Ik heb hem getoond zoals hij is
6	gnittez	openbaar aanklager Carla Del Ponte liet tijdens zijn openingsuitlegging		beelden	zien van de verschillende Joegoslavische oorlogen. Het beschikt ook over d
7	nee eic	namen. Nu gaat ook de conjunctuurbarometer van de Nationale Bank, die ee		beeld	geeft van de verwachte economische ontwikkeling, duidelijk verder omhoog.
8	ni drew	ewerkt aan een klaarere ruimtelijke inrichting. Die transformatie werd in		beeld	gebracht door kunstenaar Candida Höfer, gespecialiseerd in het fotografer
9	ekrets	en voelen, is het samenbrengen van hyperindividueel geformuleerde, sterke		beelden	van universele emoties, denkbeelden en energieën, verbonden met hetgeen
10	nee si t	uwen uit been of geschilderd op rotsen. Het oudste bekende artefact is een		beeldje	van een mens met hondenkop dat 32.000 jaar oud zou zijn. Voorbeelden van
11	po ned	sselse parket vervolgde de veehandelaars nadat ze geïdentificeerd werden op		beelden	die de dierenrechtenorganisaties Gaia en Animaux en Pêril stiekem gedraaid
12	ed th	stiekem gedraaid hadden in november 2000 op de veemarkt van Anderlecht.		beelden	ontloken toen een storm van protest bij de publieke opinie en zetten het p
13	ed tad	ertoe aan een onderzoek in te stellen. De rechter oordeelt vandaag dat de		beelden	een inbreuk vormen op de privacywet en daardoor een onrechtmatig bewijs zi
14	deog n	erspiegelt die een bedrijf aan zijn klanten aantrekt, geeft hij een goed		beeld	van de tarieven die een communicatieconsultant als Slangen & Partners (
15	ne , vt	rouwens leeg stonden. Helaas was de pers erbij, met fotografen en tv, en		beelden	van stuk huisraad en speelgoed en hulende peuters tussen het puin gingen d

Klaar 0 MB / 23 MB 28 MB / 50 MB

start Postvak IN - Microsof... dutchsemcor - Mozilla... PROTOCOLS Screenshots SAT - M...

4. The tool is built in such way that only necessary information is presented at once in the different windows but standing on the relevant data with the cursor, more information is provided for each field (e.g. more context, more synonyms, hyponymy/ hypernymy relations, domains etc.). This way, the annotator is able to decide which extra information he/she needs in order to assign the right sense to the examples.

dutchsemcor

Free List Lemmas: beeld Category: adj noun verb ? Context: 75 chars Search

#	Examples	Morphosyntax	Resume/Def	Domains	Synonyms
1	een marmeren/bronzen beeld	n-het-t	beeldhouwwerk	sculpture	
2	een wazig beeld	n-het-t	plaatje	factotum	
3	iemand een beeld van iets geven	n-het-t	voorstelling	factotum	voorstelling
4	een beeld van [een huisje]	n-het-nt	iets moois	factotum	
5	een treffend beeld	n-het-t	symbool	factotum psych	voorstelling

of 5 rows

Co-occurring: Clear Filter Usage: N F I U Paginate

tfel Left

1 hcsilyd Leopold III

2 teh nex in het beze

3 seicerp

4 hcsirob Het is trouw

5 teH . d

6 gnittez openbaar a

7 nee eic nomen. Nu gaat ook de conjunctuurbarometer van de Nationale Bank, die ee

8 ni drew ewerkt aan een klaarere ruimtelijke inrichting. Die transformatie werd in

9 ekrets en voelen, is het samenbrengen van hyperindividueel geformuleerde, sterke

10 nee si t uwen uit been of geschilderd op rotsen. Het oudste bekende artefact is een

11 po ned sselse parket vervolgde de veehandelaars nadat ze geïdentificeerd werden op

12 ed th stiekem gedraaid hadden in november 2000 op de veemarkt van Anderlecht.

13 ed tad ertoe aan een onderzoek in te stellen. De rechter oordeelt vandaag dat de

14 deog n erspiegelt die een bedrijf aan zijn klanten aantrekt, geeft hij een goed

15 ne , vt rouwens leeg stonden. Helaas was de pers erbij, met fotografen en tv, en

Klaar 0 MB / 23 MB 28 MB / 50 MB

start Postvak IN - Microsof... dutchsemcor - Mozilla... PROTOCOLS Screenshots SAT - M...

Lemma	Category	Morphosyntax	Resume/Def	Domains	Synonyms
bronsen beeld	n-het-t	beeldhouwwerk	sculpture		
een beeld	n-het-t	plaatje	factotum		
beeld van iets geven	n-het-t	voorstelling	factotum	voorstelling	
een huisje]	n-het-nt	iets moois	factotum		
beeld	n-het-t	symbool	factotum psych	voorstelling plaatje	

#	Relation	Type
1	weergave	HAS_HYPERONYM
2	voorstellen	BE_IN_STATE
3	representeren	BE_IN_STATE
4	tatoeëren	ROLE_INSTRUMENT
5	zandstralen	ROLE_INSTRUMENT

Also the context in the corpus can be extended to 3 sentences if necessary.

dutchsemcor - Mozilla Firefox

Bestand Bewerken Beeld Geschiedenis Bladwijzers Extra Help

http://cornetto.science.uva.nl:8080/dutchsemcor/

Meest bezocht Aan de slag Laatste nieuws

dutchsemcor

Free List Lemma: beeld Category: adj noun verb ? Context: 75 chars Search

#	Examples	Morphosyntax	Resume/Def	Domains	Synonyms	Relations	Tagged
1	een marmeren/bronsen beeld	n-het-t	beeldhouwwerk	sculpture		afbeelding plaatje	24
2	een wazig beeld	n-het-t	plaatje	factotum		afbeelding plaatje	0
3	iemand een beeld van iets geven	n-het-t	voorstelling	factotum	voorstelling	weergave	2
4	een beeld van [een huisje]	n-het-nt	iets moois	factotum		iets	26
5	een treffend beeld	n-het-t	symbool	factotum psych	voorstelling plaatje	gedachte	1

1 of 5 rows

Co-occurring: Clear Filter Usage: N F I U Paginate 1 ... 50 of 1920 Tag senses

#	Left	Sense	Word	Right
1	hcsilyd Leopold III ' hangt zijn oudste dochter , prinses Esmeralda , een idyllisch	1,3	beeld	op van haar jeugdijaren in Argentuul en de familie bonheur . Wat hij in h
2	teh nec in het bezette België , tegen het advies van zijn ministers in , zouden het		beeld	van de vorst zwaar aantasten . En na zijn huwelijk met Lilian Baels tijdens
3	sekerp ' Misschien was het zelfs al vroeger . Ik heb geen precies	1,3	beeld	van het moment waarop ik me ervan bewust geworden ben , dat mijn vader e
4	hcsiro: Het is trouwens niet aan zijn dochter om van hem een politiek of historisch	5	beeld	op te hangen , want dat zou niet objectief zijn . Bovendien heb ik die peri
5	teH . d ' Hij was een veelzijdige en een rijke persoonlijkheid . Het	4,4	beeld	van de getourmenteerde koning ken ik niet . Ik heb hem getoond zoals hij is
6	gnitex openbaar aanklager Carla Del Ponte liet tijdens zijn openingsuitletzing	4,4	beelden	zien van de verschillende Joegoslavische oorlogen . Het beschikt ook over d
7	nee ek nomen . Nu gaat ook de conjunctuurbometer van de Nationale Bank , die ee	4,4	beeld	geeft van de verwachte economische ontwikkeling , duidelijk verder omhoog .
8	ni drew erwerkt aan een Klaardere ruimtelijke inrichting . Die transformatie werd in	4,4	beeld	gevoeld in het fotografer
9	ekrets en voelen , is het samenbrengen van hyperindividueel geformuleerde , sterke	4,4	beeld	verbonden met hetgeen
10	nee si tuwen uit been of geschilderd op rotsen . Het oudste bekende artefact is een	4,4	beeld	u zijn . Voorbeelden van
11	po ned sselse parket vervolgde de veehandelaars nadat ze geïdentificeerd werden op	4,4	beeld	Pérl stiekem gedraaid
12	eD , th stiekem gedraaid hadden in november 2000 op de veemarkt van Anderlecht .	4,4	beeld	pinie en zetten het p
13	ed tad ertoe aan een onderzoek in te stellen . De rechter oordeelt vandaag dat de	4,4	beeld	in onrechtmatig bewijs
14	deog n eerspiegelt die een bedrijf aan zijn klanten aanrekent , geeft hij een goed	4,4	beeld	angen & Partners (
15	ne , vt rouwens leeg stonden . Helaas was de pers erbij , met fotografen en tv , en	4,4	beeld	ussen het pulp gingen d
16	cigaM ndien niet onmogelijk te bewijzen . Intussen draaide ook Dennis Black Magic	4,4	beeld	hij treuws voor ei
17	teH . m En daar ging het om . Het	4,4	beeld	correct , maar sin

Klaar

Nog niet geladen 28 MB / 50 MB

start Postvak IN - Microsof... PROTOCOLS Screenshots SAT - Mi... dutchsemcor - Mozilla...

NL 13:57

It is also possible to group corpus examples according to the word left or right to the target lemma.

The screenshot shows the dutchsemcor web interface in a Mozilla Firefox browser. The search bar at the top contains the lemma 'beeld'. Below the search bar, there is a table of corpus examples. The table has columns: #, Examples, Morphosyntax, Resume/Def, Domains, Synonyms, Relations, and Tagged. The first five rows are visible, showing examples like 'een marmeren/bronzen beeld', 'een wazig beeld', 'iemand een beeld van iets geven', 'een beeld van [een huisje]', and 'een treffend beeld'. Below the table, there is a 'Co-occurring' window. This window has a search bar and a table of co-occurrences. The table has columns: #, tfeL, Left, Sense, Word, and Right. Two ovals highlight the 'Left' and 'Right' columns in the co-occurrence table.

In certain cases, a specific meaning of a lemma occurs frequently together with particular context words (eg.: 'beeld' + 'kunst'). In these cases, the co-occurrence feature of the SAT can be used. The user can type different word-clues or multi word expressions in the *co-occurrence* window and the corpus examples will be filtered according to these search terms showing them in bold.

The screenshot shows the 'Co-occurring' window in the dutchsemcor web interface. The search bar in the 'Co-occurring' window contains the word 'kunst'. Below the search bar, there is a table of co-occurrences. The table has columns: #, tfeL, Left, Sense, Word, and Right. The table is currently empty, indicating that no co-occurrences were found for the word 'kunst'.

dutchsemcor - Mozilla Firefox

Bestand Bewerken Beeld Geschiedenis Bladvijzers Extra Help

http://corretto.science.uva.nl:8080/dutchsemcor/

Meest bezocht Aan de slag Laatste nieuws

dutchsemcor

Free List Lemma: beeld Category: adj noun verb ? Context: 75 chars Search

#	Examples	Morphosyntax	Resume/Def	Domains	Synonyms	Relations	Tagged
1	een marmeren/bronzen beeld	n-het-t	beeldhouwwerk	sculpture		afbeelding plaatje	0
2	een wazig beeld	n-het-t	plaatje	factotum		afbeelding plaatje	0
3	iemand een beeld van iets geven	n-het-t	voorstelling	factotum	voorstelling	weergave	0
4	een beeld van [een huisje]	n-het-nt	iets moois	factotum		iets	0
5	een treffend beeld	n-het-t	symbool	factotum psych	voorstelling plaatje	gedachte	0

1 of 5 rows

Co-occurring: kunst Clear Filter Usage: N F I U Paginate 1 ... 149 of 1920 (149) Tag senses

#	tfel	Left	Sense	Word	Right
8	ni drew	ewerkt aan een klaardere ruimtelijke inrichting . Die transformatie werd in		beeld	gebracht door kunstenaar Candida Höfer , gespecialiseerd in het fotograferen
9	ekrets	en voelen , is het samenbrengen van hyperindividueel geformuleerde , sterke		beelden	van universele emoties , denkbeelden en energieën , verbonden met hetgeen
42	deog n	oneel wonen . De excursie die georganiseerd wordt door KUBES zal u een goe		beeld	geven van de ruimte en de inrichting , omdat speciaal voor deze gelegenheid
52	VT ed t	ziek clips van TMF) , Delay TV (wat de kijker de mogelijkheid geeft de TV		beelden	van Nederland 1,2 en 3 van de afgelopen 24 uur terug te zien) , Kunst &
76	ed ned	ig met het wezenlijke van de vorm . De bezoeker wordt dan ook aangeraden c		beelden	te betasten . Op de tentoonstelling "De kunst van het anders zien VIII " z
77	niz " II	betasten . Op de tentoonstelling "De kunst van het anders zien VIII " zijn		beelden	, schilderijen , keramiek en foto's van een groot aantal bij Stichting KUBE
78	teb jib	stenaars te zien . De andere expositie , getiteld " Water - Dichter bij het		beeld	" , verbindt kunstwerken rond het thema water met gedichten van blinde en
79	ed nav	aal omlijst door pianist Bert den Brink . In het boekje staan foto's van de		beelden	waarop de gedichten betrekking hebben . De catalogus kost 7,50 euro .
80	po drev	volgt nu het gedicht " Ach , laat me staren " van Tom Hugen , gebaseerd op		beeld	Zilverzwaan 1 van Jannetje Smeets - de Groot . (cd-tje met track : 2.11 min
81	teb jib	posities "De kunst van het anders zien VIII " en " Water - Dichter bij het		beeld	" zijn te zien van dinsdag 1 juni tot en met zaterdag 3 juli in de Gemeente
95	etsree	ling langs sculpturen uit Zimbabwe . Deze trotse dame was één van de eerste		beelden	die ik vorig jaar mocht bewonderen tijdens een expositie van galerie Sonda
96	nav ne	babwaanse beeldhouwkunst . Regelmatig organiseert zij tentoonstellingen v		beelden	gemaakt door Zimbabweanse beeldhouwers . Dit jaar exposeert Sonda van 1
97	ednere	de bomen , op de grasvelden en langs het water staan , veelal schitterende		beelden	opgesteld . Niet alleen buiten kunt u genieten van deze exotische pracht ma
99	nenets	utten , onder de bomen , achter een waslijntje , overal staan zwarte stenen		beelden	op palen : vogels , schildpadden en vreemde menselijke figuren staren de bezoeker
103	teb na	eelden met drie gezichten of het beeld van een zwangere vrouw . Ook aan he		beeld	van twee handen met daartussen een gezicht kan men verschillende beteken

Klaar

0 MB / 23 MB 28 MB / 50 MB

start Postvak IN - Microsof... dutchsemcor - Mozila... PROTOCOLS Screenshots SAT - Mi... NL 13:51

If the annotator is ready with analysing the sense description from the lexical database and the different corpus examples from the corpus database, the different senses can be mapped onto the corpus fragments. For this, the user needs to select one or more senses and one or more corpus examples and link them with the button 'Tag senses'.

dutchsemcor - Mozilla Firefox

Bestand Bewerken Beeld Geschiedenis Bladvijzers Extra Help

http://corretto.science.uva.nl:8080/dutchsemcor/

Meest bezocht Aan de slag Laatste nieuws

dutchsemcor

Free List Lemma: beeld Category: adj noun verb ? Context: 75 chars Search

#	Examples	Morphosyntax	Resume/Def	Domains	Synonyms	Relations	Tagged
1	een marmeren/bronzen beeld	n-het-t	beeldhouwwerk	sculpture		afbeelding plaatje	0
2	een wazig beeld	n-het-t	plaatje	factotum		afbeelding plaatje	0
3	iemand een beeld van iets geven	n-het-t	voorstelling	factotum	voorstelling	weergave	0
4	een beeld van [een huisje]	n-het-nt	iets moois	factotum		iets	0
5	een treffend beeld	n-het-t	symbool	factotum psych	voorstelling plaatje	gedachte	0

1 of 5 rows

Co-occurring: Clear Filter Usage: N F I U Paginate 1 ... 50 of 1920 Tag senses

Tag corpus with selected senses

#	tfel	Left	Sense	Word	Right
1	hcsillyd	Leopold III ' hangt zijn oudste dochter , prinses Esmeralda , een idyllisch	1,3	beeld	op van haar jeugdijaren in Argentuul en de familieale bonheur . Wat hij in h
2	teh ne	in het bezette België , tegen het advies van zijn ministers in , zouden het		beeld	van de vorst zwaar aantasten . En na zijn huwelijk met Lilian Baels tijdens
3	seicorp	" Misschien was het zelfs al vroeger . Ik heb geen precies	1,3	beeld	van het moment waarop ik me ervan bewust geworden ben , dat mijn vader e
4	hcsirof	Het is trouwens niet aan zijn dochter om van hem een politiek of historisch		beeld	op te hangen , want dat zou niet objectief zijn . Bovendien heb ik die peri
5	teh . d	" Hij was een veelzijdige en een rijke persoonlijkheid . Het		beeld	van de getroumteerde koning ken ik niet . Ik heb hem getoond zoals hij is
6	gnittez	openbaar aanklager Carla Del Ponte liet tijdens zijn openingsuiteenzetting		beelden	zien van de verschillende Joegoslavische oorlogen . Het beschikt ook over d
7	nee eic	nomen . Nu gaat ook de conjunctuurbareometer van de Nationale Bank , die ee		beeld	geeft van de verwachte economische ontwikkeling , duidelijk verder omhoog .
8	ni drew	ewerkt aan een klaardere ruimtelijke inrichting . Die transformatie werd in		beeld	gebracht door kunstenaar Candida Höfer , gespecialiseerd in het fotograferen
9	ekrets	en voelen , is het samenbrengen van hyperindividueel geformuleerde , sterke		beelden	van universele emoties , denkbeelden en energieën , verbonden met hetgeen
10	nee si t	uwen uit been of geschilderd op rotsen . Het oudste bekende artefact is een		beeldje	van een mens met hondenkop dat 32.000 jaar oud zou zijn . Voorbeelden van
11	po ned	sselse parket vervolgde de veehandelaars nadat ze geïdentificeerd werden op		beelden	die de dierenrechtenorganisaties Gaia en Animaux en Pélit stiekem gedraaid
12	eD . th	stiekem gedraaid hadden in november 2000 op de veemarkt van Anderlecht .		beelden	ontlokten toen een storm van protest bij de publieke opinie en zetten het p
13	ed tad	ertoe aan een onderzoek in te stellen . De rechter oordeelt vandaag dat de		beelden	een inbreuk vormen op de privacywet en daardoor een onrechtmatig bewijs zi
14	deog n	erspiegelt die een bedrijf aan zijn klanten aanrekent , geeft hij een goed		beeld	van de tarieven die een communicatieconsultant als Slangen & Partners (
15	ne . vt	rouwens leed stonden . Helaas was de pers erbij . met foto's af en tv . en		beelden	van stuk huisraad en soeeloed en hullede deuters tussen het duin oinden d

dutchsemcor - Mozilla Firefox

Bestand Bewerken Beeld Geschiedenis Bladvijzers Extra Help

http://cornetto.science.uva.nl:8080/dutchsemcor/

Meest bezocht Aan de slag Laatste nieuws

dutchsemcor

Free List Lemma: beeld Category: adj noun verb Context: 75 chars Search

#	Examples	Morphosyntax	Resume/Def	Domains	Synonyms	Relations	Tagged
1	een marmeren/bronzen beeld	n-het-t	beeldhouwwerk	sculpture		afbeelding plaatje	2
2	een wazig beeld	n-het-t	plaatje	factotum		afbeelding plaatje	0
3	iemand een beeld van iets geven	n-het-t	voorstelling	factotum	voorstelling	weergave	2
4	een beeld van [een huisje]	n-het-nt	iets moois	factotum		iets	0
5	een treffend beeld	n-het-t	symbool	factotum psych	voorstelling plaatje	gedachte	0

1 of 5 rows

Co-occurring: Clear Filter Usage: N F I U Paginate 1 ... 50 of 1920 Tag senses

16	Ugari...						
17	teH. m						
18	teH : s						
19	nee gic						
20	ni retsr						
21	ni nem						
22	ni ne n						
23	nee dn						
24	nee hcd						
25	gon jh						
26	teh mo						
27	teh tdr						
28	teh tod						
29	teh nav						
30	teh nav						
31	reem re						
32	teh ten						
33	kileenn						

Klaar

start Postvak IN - Microsof... PROTOCOLS Screenshots SAT - M... dutchsemcor - Mozilla... NL 13:54

Special measures are taken to detect and exclude idiomatic usages of words from the retrieved text. In case these multi-word units cannot be excluded automatically, annotators will tag them (**I** = Idiom). Furthermore, if a word in the corpus does not occur in Cornetto, it will be labelled by the human annotator as a new word meaning (**U**= Unknown) and added to the database during an editorial round (See section 4). Similarly, the SAT supports labelling figurative usage and metonymic usage (**F**= Figurative).

een huisje]	n-het-nt	iets moois	factotum		iet
eld	n-het-t	symbool	factotum psych	voorstelling plaatje	ge

Clear Filter Usage: N F I U Paginate 1 ... 50 of 1920

Figurative

met onmogelijk te bewijzen ; inussen draaide ook Dennis black magic					
En daar ging het om . Het	4				
en . Tijdens de debatten werd gediscussieerd over drie thema's : Het					
Daarop gaan we ons storten . Maar eerst is het handig een					
Windows Help te openen . Er komt dan een normaal programmavenst					
zij op de titel van dat hoofdstuk klikken . De paragraaftitels komen in	1				
heft te klikken om de tekst ervan automatisch te laten opzoeken en in	1				
alisten , en van haar huisarts . De uitgave geeft vanzelfsprekend een	1				
praktijkondersteuners in opleiding omdat het hen kan helpen zich een	1				
loen is en dat mijn vader dus blind blijft . De eerste dagen zag hij nog	1				

dutchsemcor - Mozilla Firefox

Bestand Bewerken Beeld Geschiedenis Bladwijzers Extra Help

http://cornetto.science.uva.nl:8080/dutchsemcor/

Meest bezocht Aan de slag Laatste nieuws

dutchsemcor

Free List Lemma: beeld Category: adj noun verb Context: 75 chars Search

#	Examples	Morphosyntax	Resume/Def	Domains	Synonyms	Relations	Tagged
1	een marmeren/bronzen beeld	n-het-t	beeldhouwwerk	sculpture		afbeelding plaatje	2
2	een wazig beeld	n-het-t	plaatje	factotum		afbeelding plaatje	0
3	iemand een beeld van iets geven	n-het-t	voorstelling	factotum	voorstelling	weergave	2
4	een beeld van [een huisje]	n-het-nt	iets moois	factotum		iets	0
5	een treffend beeld	n-het-t	symbool	factotum psych	voorstelling plaatje	gedachte	0

5 rows

Co-occurring: Clear Filter Usage: N F I U Paginate 1 ... 50 of 1920 Tag senses

Tag corpus with selected senses

10 tegelijk... niet met onmogelijk te bewijzen. Intussen draaide ook Dennis Black Magic...
 17 teH. m... En daar ging het om. Het
 18 teH. s... te halen. Tijdens de debatten werd gediscussieerd over drie thema's: Het
 19 nee ge... Daarop gaan we ons storten. Maar eerst is het handig een
 20 ni rets... gemene Windows Help te openen. Er komt dan een normaal programma van:
 21 ni nem... of zij op de titel van dat hoofdstuk klikken. De paragraaftitels komen in
 22 ni ne n... itel hoeft te klikken om de tekst ervan automatisch te laten opzoeken en in
 23 nee dn... specialisten, en van haar huisarts. De uitgave geeft vanzelfsprekend een
 24 nee hci... en praktijkondersteuners in opleiding omdat het hen kan helpen zich een
 25 gon jh... te doen is en dat mijn vader dus blind blijft. De eerste dagen zag hij nog
 26 teh mo... elmatig wordt er op de regionale centra gevraagd naar een hulpmiddel om het
 27 teh tdr... jkomingeving rustig. Door dichterbij de televisie te gaan zitten, wordt het
 28 teh toc... Een voorzetscherm oftewel beeldschermvergroter vergroot het
 29 teh nav... ook minder scherp. Je moet recht voor de tv zitten om vertekening van het
 30 teh nav... et voor een aantal mensen met een visuele beperking een verbetering van he
 31 reem re... e AVRO de documentaire En de wind zong uitzenden. Hierin ziet u onder meer
 32 teh ten... Iedereen maakt het wel eens mee, plotseling wordt u geconfronteerd met hel
 33 kileens... Maar het blijft niet bij die vooroordelen alleen. Het maatschappelijk

Klaar

start Postvak IN - Microsof... PROTOCOLS Screenshots SAT - M... dutchsemcor - Mozilla... NL 13:55

In the right column on the upper table under 'Tagged', the user can see how many instances have been tagged with the given sense. Since the aim of the manual annotation task is to provide at least 25 examples for each senses it is indispensable to keep an eye on this column in order to reach satisfying numbers (quantitative measurement).

dutchsemcor

Free List Lemma: beeld Category: adj noun verb Context: 75 chars Search

#	Examples	Morphosyntax	Resume/Def	Domains	Synonyms	Relations	Tagged
1	een marmeren/bronzen beeld	n-het-t	beeldhouwwerk	sculpture		afbeelding plaatje	24
2	een wazig beeld	n-het-t	plaatje	factotum		afbeelding plaatje	0
3	iemand een beeld van iets geven	n-het-t	voorstelling	factotum	voorstelling	weergave	2
4	een beeld van [een huisje]	n-het-nt	iets moois	factotum		iets	26
5	een treffend beeld	n-het-t	symbool	factotum psych	voorstelling plaatje	gedachte	1

5 rows

Co-occurring: Clear Filter Usage: N F I U Paginate 1 ... 50 of 1920 Tag senses

#	Left	Sense	Word	Right
1	hcsillyd Leopold III ' hangt zijn oudste dochter, prinses Esmeralda, een idyllisch	1,3	beeld	op van haar jeugdijaren in Argentueil en de familie bonheur. Wat hij in h
2	teh nee in het bezette België, tegen het advies van zijn ministers in, zouden het		beeld	van de vorst zwaar aantasten. En na zijn huwelijk met Lillian Baels tijdens
3	seicerp ' Misschien was het zelfs al vroeger. Ik heb geen precies	1,3	beeld	van het moment waarop ik me ervan bewust geworden ben, dat mijn vader e
4	hcsiro: Het is trouwens niet aan zijn dochter om van hem een politiek of historisch	5	beeld	op te hangen, want dat zou niet objectief zijn. Bovendien heb ik die peri
5	teh. d ' Hij was een veelzijdige en een rijke persoonlijkheid. Het	4,4	beeld	van de getourmenteerde koning ken ik niet. Ik heb hem getoond zoals hij is
6	gnittez openbaar aanklager Carla Del Ponte liet tijdens zijn openingsuitspraak	4,4	beelden	zien van de verschillende Joegoslavische oorlogen. Het beschikt ook over d
7	nee eic nomen. Nu gaat ook de conjunctuurbarometer van de Nationale Bank, die ee	4,4	beeld	geeft van de verwachte economische ontwikkeling, duidelijk verder omhoog.
8	ni drev ewerkt aan een klaardere ruimtelijke inrichting. Die transformatie werd in	4	beeld	gebracht door kunstenaar Candida Höfer, gespecialiseerd in het fotografer
9	ekrets en voelen, is het samenbrengen van hyperindividueel geformuleerde, sterke	4,4	beelden	van universele emoties, denkbeelden en energieën, verbonden met hetgeen
10	nee si t uwen uit been of geschilderd op rotsen. Het oudste bekende artefact is een	4,4	beeldje	van een mens met hondenkop dat 32.000 jaar oud zou zijn. Voorbeelden van
11	po ned sselse parket vervolgde de veehandelaars nadat ze geïdentificeerd werden op	4,4	beelden	die de dierenrechtenorganisaties Gaia en Animaux en Pélit stiekem gedraaid
12	eD. th stiekem gedraaid hadden in november 2000 op de veemarkt van Anderlecht.	4,4	beelden	ontloken toen een storm van protest bij de publieke opinie en zetten het p
13	ed tad ertoe aan een onderzoek in te stellen. De rechter oordeelt vandaag dat de	4,4	beelden	een inbreuk vormen op de privacywet en daardoor een onrechtmatig bewijs zi
14	deog n eerspiegelt die een bedrijf aan zijn klanten aanrekent, geeft hij een goed	4,4	beeld	van de tarieven die een communicatieconsultant als Slangen & Partners (
15	ne, vt rouwens leeg stonden. Helaas was de pers erbij, met fotografen en tv, en	4,4	beelden	van stuk huisraad en speelgoed en hulende peuters tussen het puin gingen d
16	cigaM k ndien niet onmogelijk te bewijzen. Intussen draaide ook Dennis Black Magic	4,4	beelden	voor zijn pornoproducties. In zijn boekjes adverteert hij trouwens voor ei
17	teH. m... En daar ging het om. Het	4,4	beeld	van de aarde zoals wij die ons voorstellen, is dus niet correct, maar sin

Summary: WSD strategies in the manual annotation process using the SAT

As mentioned in section 2 of this manual, the annotation process as applied in the DutchSemCor project is made up of different cycles and phases. Now, we will discuss the actual process of tagging, focusing on practical points necessary to solve WSD problems.

The idea is that the annotators work with a pre-selected wordlist and a limited corpus output (this is important in cases when there are too many examples available for a highly frequent word). In an ideal setting, we have a lemma with a couple of senses and for each sense approx. 100 examples from the corpus. From these instances, the annotator is to tag 25 examples per sense. Since the corpus output is limited AND filtered for diversity, the annotation task is in this ‘ideal’ setting quite straightforward and can be fulfilled at a high speed.

In most cases, however 1.) there won’t be enough examples for each sense; 2.) the SAT will not be able to provide examples which are diverse enough and finally 3.) the SAT will be unable to limit the number of examples, which will in turn cause an overload of one sense in the corpus output.

The first case i.e. the lack of examples for certain senses in the corpus can be solved by acquiring text fragments from the Internet using the Snippet Tool (See **3.3**). In the other two cases, annotators will need to limit the corpus output by grouping the examples and by using other methods (e.g. Co-occurrence strategies) to properly match Cornetto senses to the corpus instances. In what follows, we will discuss these WSD techniques in more detail.

The annotation process involves the following steps:

Step 1 - Analyzing the Cornetto data by scanning and clustering senses

Step 2 - Analyzing the corpus data by grouping corpus occurrences based on neighbouring words, co-occurrence and

Step 3 - Complementing corpus data with snippets from the internet and from Open Taal using the Snippet Tool if necessary (see **section 3.3**).

Step 4 - Mapping Cornetto senses onto corpus occurrences

Analyzing the Cornetto data

In the SAT, The upper-half of the screen represents data imported from Cornetto. The different fields in the table provide morpho-syntactic information, synonyms and other relations of the different senses, examples and information on domain.

dutchsemcor							
Free	List	Lemma: leiden	Context: 25	chars	Search		
#	Examples	Morpho	Resume/Def	Domains	Synonyms	Relations	Tag
1	een kind aan de hand <u>leiden</u>		in een richting doen gaan	factotum	voeren leiden	meevoeren meenemen	
2	een leven <u>leiden</u>		voeren naar	factotum	voeren leiden lopen	zijn vormen uitmaken	
3	iemand in verzoeking <u>leiden</u>		in een toestand brengen	factotum	veroorzaken tweewegbrengen leid	passeren afspelen gebeu	
4	een vergadering <u>leiden</u>		de leiding hebben	factotum	leiden leidend	handelen	
5	het peloton <u>leiden</u>		de voorste zijn	factotum	leiden	zijn vormen uitmaken	
6	[nergens toe/tot niets] <u>leiden</u>		resulteren in	factotum	uitdraaien uitlopen uitpakken val	veranderen kenteren ker	
rows							
Co-occurring:			Clear	Filter	Tag senses		
#	tfeL	Left	Sense	Lemma	W Id	Right	
1	et fa st	Uit de klanken valt niks af te		leiden	WR-P-E-C-0000000022.p.145.s.5.w.8	.	
2	nesteo	... ewaking van de instaptoetsen		leiden	WR-P-E-C-0000000127.p.60.s.1.w.8	.	
3	neziwif	... s tot onjuiste schrijfwijzen		leiden	WR-P-E-C-0000000321.p.3.s.2.w.9	.	
4	eiD	Die		leiden	WR-P-E-C-0000000174.p.53.s.3.w.2	voor de eerste fase tot de volgende prioritering	
5	tiu gnir	... ulp voor matthias huening uit		leiden	WR-P-E-C-0000000455.p.3.s.1.w.9	(vraag 9210.3) .	
6	eotre lj	... ns kan een slechte stijl ertoe		leiden	WR-P-E-C-0000000240.p.128.s.4.w.7	dat lezers het stuk terzij ...	
7	et fa si	Dit is af te		leiden	WR-P-E-C-0000000034.p.221.s.5.w.5	uit het colofon of het zogenaamd drukkersmerk i	
8	ne ned	... or een verklaring te vinden en		leiden	WR-P-E-C-0000000127.p.460.s.4.w.10	daaruit de noodzaak af voo ...	
9	et liuk	... ngs menig spellingvalkuil te		leiden	WR-P-E-C-0000000096.p.216.s.2.w.17	.	

Based on the Cornetto data, annotators use the following strategy to match senses onto corpus examples:

1. The annotator first checks the data in this window: the resume, the definitions, the examples, the morphological information and the different relations of the senses.
2. The annotator tries to cluster the different senses and specifies the LU internal meaning shifts.

Scanning senses

When scanning the different senses of a lemma, annotators should start with the information available in the different fields of the table. If the available information is not enough, extra data can be imported by standing on the given fields with the cursor.

Clustering senses

After acquainting themselves with the different meanings of a lemma, annotators need to look at the different meaning shifts i.e. the relations between the different senses. This is important when making meaning clusters. These clusters can be used later for the grouping of corpus examples (See the following sections).

Example

If we look at the different senses of the word “leiden” (= lead), we see 3 main clusters:

1. spatial position
 - a. dynamic (“iemand door de stad leiden”) – **sense 1**
 - b. static (“de weg leidt naar Alkmaar”) – **sense 2**
2. cause to be/happen (“iemand in verzoeking leiden”; “tot discussie leiden”) – **sense 3, 6**
3. leadership (“een vergadering leiden”; “een peloton leiden”) – **sense 4,5**

These clusters reveal us the prototypical meaning(s) of the word together with the meaning shifts which formed the different senses. This way, we can interpret the various meanings with greater accuracy.

Clustering senses has also the advantage that annotators can remember the different senses when annotating. This advantage becomes obvious when dealing with lemmas with 6 or more meanings. It is simply impossible for the annotators to familiarize themselves completely with all senses and turn to the corpus examples without going back and forth between the

Looking up the etymology of the verb “lead” in the OED, we conclude the following:

- The original meaning of the verb “leiden” is “go, travel”
- The meaning of “leiden” has been influenced by the Latin “ducere” with the meaning “lead, guide, drive”
- The prototypical meaning of “lead” is consequently the meaning **1.a spatial position, dynamic**

We can sketch now the meaning shifts of the lemma “lead” using syntactic info on complementation, argument structure, transitivity and thematic roles.

Sense	Transitivity	Argument struct.	Complementation	Meaning shift
1.a spatial pos. (dynamic) – sense 1	Transitive direct (Subj+ obj)	Agent (+ animate), Patient (+ animate)	NP+VP+NP	Prototypical meaning
1.b spatial pos. (static) – sense 2	Intransitive (Subj)	Instrument (-animate)	NP+VP + <i>naar</i>	Transitive -> intransitive
2. cause to be/ happen – senses 3, 6	Transitive direct/ indirect (Subj + PP)	Agent, Patient	NP+VP+NP+ <i>in verzoeking</i> NP+VP+ <i>tot</i>	Meaning extension, metaphor
3. leadership – senses 4, 5	Transitive direct	Agent (+ animate), Patient	NP+VP+NP	Meaning restriction, metonymy

dutchsemcor

Free

List

Lemma:

Category: adj noun verb ?

Context: chars

Search

Help

#	Examples	Morphology	Resume/Def	Domains	Synonyms	Relations	Tagged	
1	een kind aan de hand <u>leiden</u>	v-tr-onsch-nrefl	in een richting doen gaan	factotum	voeren leiden	meevoeren meenemen		
2	de weg <u>leidt</u> naar Alkmaar	v-intr-onsch-nrefl	voeren naar	factotum	voeren leiden lopen	zijn vormen uitmaken		
3	iemand in verzoeking <u>leiden</u>	v-tr-onsch-nrefl	in een toestand brengen	factotum	veroorzaken teweegbrengen leiden	passeren afspeelen gebe		
4	een vergadering <u>leiden</u>	v-tr-onsch-nrefl	de leiding hebben	factotum	leiden leidend	handelen		
5	het peloton <u>leiden</u>	v-tr-onsch-nrefl	de voorste zijn	factotum	leiden	zijn vormen uitmaken		
6	[nergens toe/tot niets] <u>leiden</u>	v-tr-onsch-nrefl	resulteren in	factotum	uitdraaien uitlopen uitpakken	veranderen kenteren ke		

6 rows

Co-occurring:

Clear

Filter

Usage:

N

F

I

U

Prev 1 ... 25 of 1874

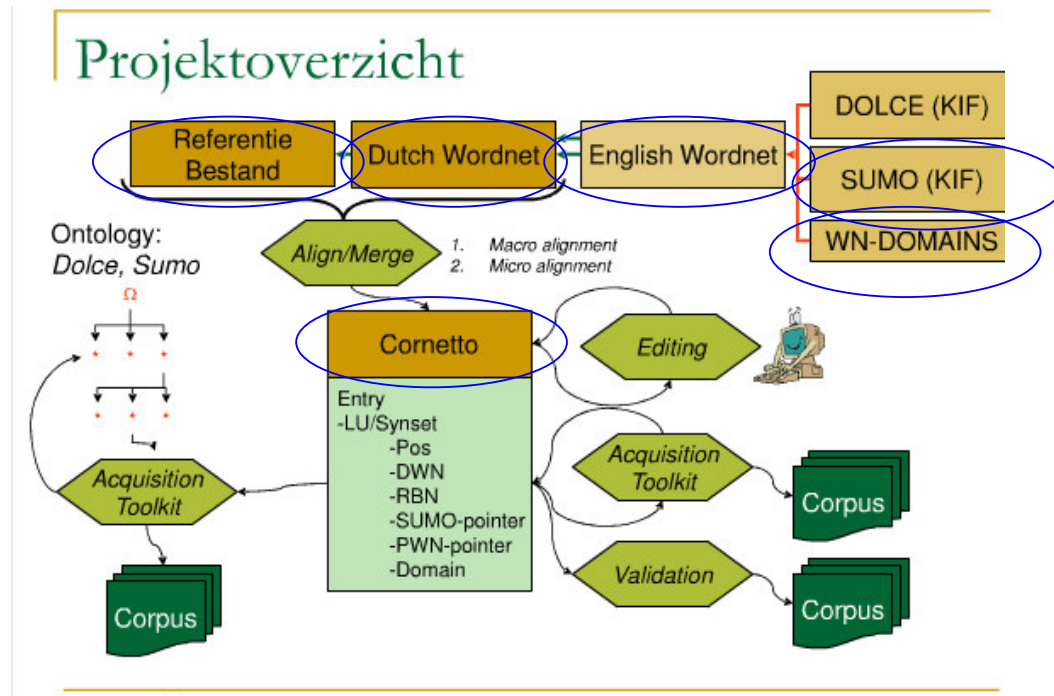
Next

Tag senses

#	tfeL	Left	Sense	Word	Right	
1	gelrev	indexatie . Gabriël Perl , die het overleg		<u>leidt</u>	, acht het eenzijdige opzeggen juridisch ongegrond .	
2	setev e	voor de parlementaire dioxinecommissie . Die interne vetes		<u>leidden</u>	finaal tot de recente afrekeningen waarin nu ook	
3	laz nek	de omkering van bewijslast , tot misbruiken zal		<u>leiden</u>	. ' Het kan mensen tot chantage of	
4	nenab	De ornitholoog die het project in goede banen		<u>leidde</u>	, Ernst Schaefer , was van Duitse origine	
5	eID . ne	konden speciale quantummechanische effecten oproepen worden . Die		<u>leidden</u>	ertoe dat zwakke stroomstoten telkens één foton uitstootten	
6	maahci	nog erger . Hormonale veranderingen in een moederlichaam		<u>leiden</u>	ertoe dat eerstgeborenen meer kans hebben om allergieën	
7	esaf et	de Europese Unie . In een eerste fase		<u>leidde</u>	dat nochtans tot een verstrakking van de standpunten	
8	nehplai	in Parijs voor grote opschudding gezorgd . Halphen		<u>leidde</u>	gedurende zeven jaar het onderzoek naar het HLM-schandaal	
9	, elano	leider van de post-fascistische partij Alleanza Nazionale ,		<u>leidt</u>	op zich ook tot heel wat controverse .	
10	', tned	' , zo zegt de president , '		<u>leidde</u>	tot armoede voor miljoenen landgenoten , vernietigde de	
11	taD . g	oprispingen bij leden van zijn regering . Dat		<u>leidde</u>	tot het ontslag van minister van Buitenlandse Zaken	
12	nak eib	komst van Braeckman tot een kleine revolutie kan		<u>leiden</u>	bij Ahrom . Er dreigt ' indien al	
13	laz nek	eurocraten dat de recente Verklaring van Laken zal		<u>leiden</u>	tot de verdere uitbouw van Europa . Als	
14	et nen	om de monetaire unie in goede banen te		<u>leiden</u>	nieuwe conflicten veroorzaken . En dat zo'n conflict	
15	et nevi	dan dat ze tot een confrontatie hoeven te		<u>leiden</u>	. Beide behoren tot dezelfde maatschappelijke cultuur .	

3.2 Cornetto Database

As we have said earlier, the Cornetto database is our lexical resource from which different data fields are presented in the SAT. This enables annotators to link the senses of a given lemma onto corpus occurrences. In this chapter, we will describe the Cornetto database: its creation and the most important Cornetto data used in the DutchSemCor project.

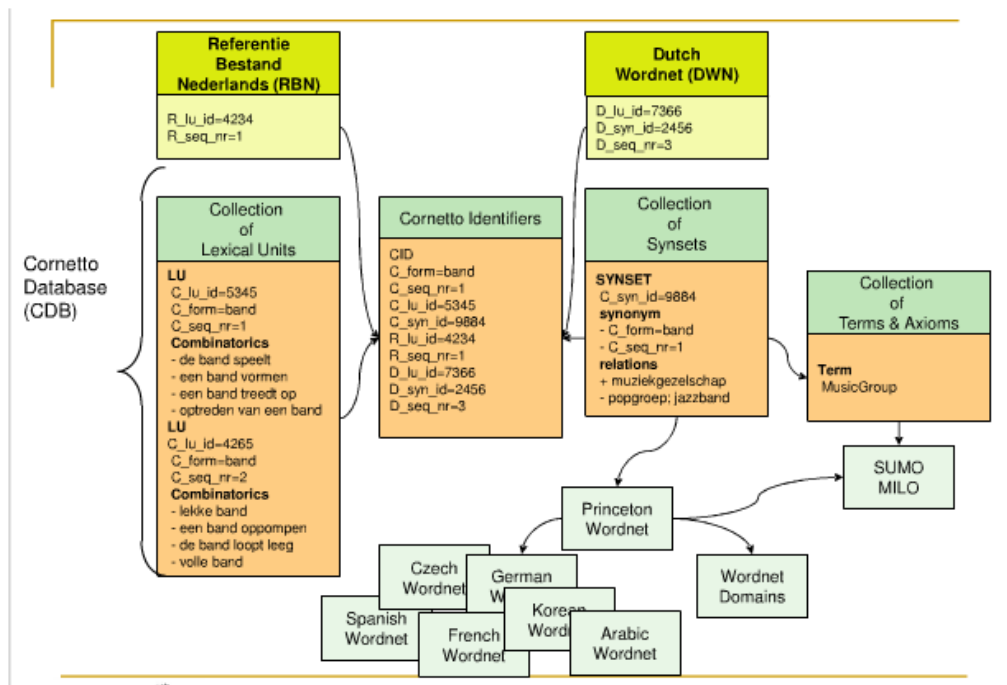


The making of Cornetto

(Based on: [Cornetto Deliverable – D16, Version 7 Januari 2009](#). Please check this document for a more detailed account of the Cornetto lexical database).

The Cornetto database was created by merging 4 main data collections:

- Collection of Lexical Units, mainly derived from the **Referentiebestand Nederlands (RBN)**
- Collection of Synsets, mainly derived from **Dutch Wordnet (DWN)**
- Collection of Domain labels derived from **Princeton Wordnet 2.0 & 3.0 (PWN)**
- Collection of Terms and axioms, mainly derived from **SUMO**



- **RBN:** RBN uses a traditional structure of form-meaning pairs, so-called Lexical Units (Cruse1986). Lexical Units are word senses in the lexical semantic tradition. They contain all the necessary linguistic knowledge that is needed to properly use the word in a language. Word meanings that are synonyms are separate structures (records) in RBN. They have their own specification of information, including morpho-syntax and semantics. For more information about RBN, please refer to RBN-Documentatie (Martin e.a., 2005).

voor een toelichting op de velden en

item	licentiaatsverhandeling
sorteerlemma	licentiaatsverhandeling
volgnummer	

lexicale informatie

woordklasse	nw
type item	w
id betekenisid	77
soort belgicisme	cultuurgebonden
definitie	verhandeling die men schrijft om de graad van licentiaat te bekommen
bron definitie	gvd
selectie BW	1

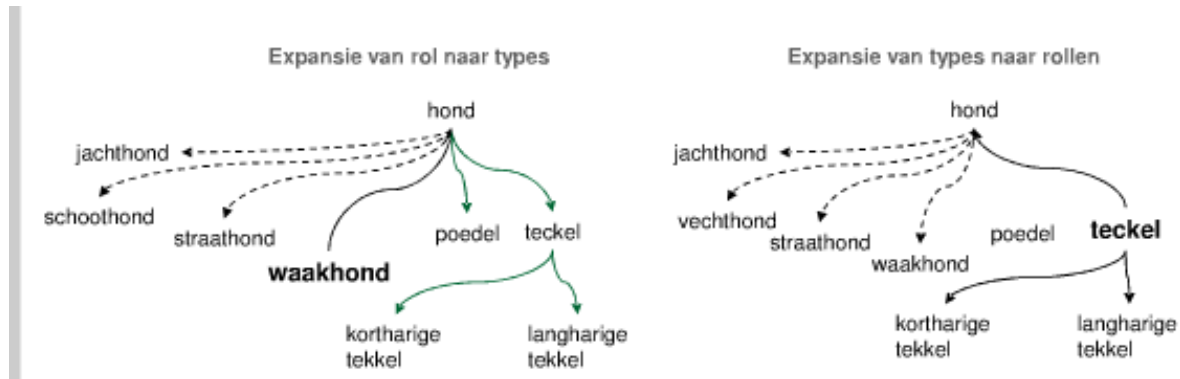
methodologische informatie

NRC freq	0	cultuurgebonden	ja	variant	scriptie
roul freq	9	domein	ond.		
Google.nl	0	restricties			
Google.be	40	type restrictie		lexicalisatie van variant	ja
CGN- freq		waarde restrictie		syntaxis	0
gvd label	b	informanten bron		roul freq van variant	11
restr gvd		a score trefwoord		a score variant	
verschueren		b score trefwoord		b score variant	
verschueren label					

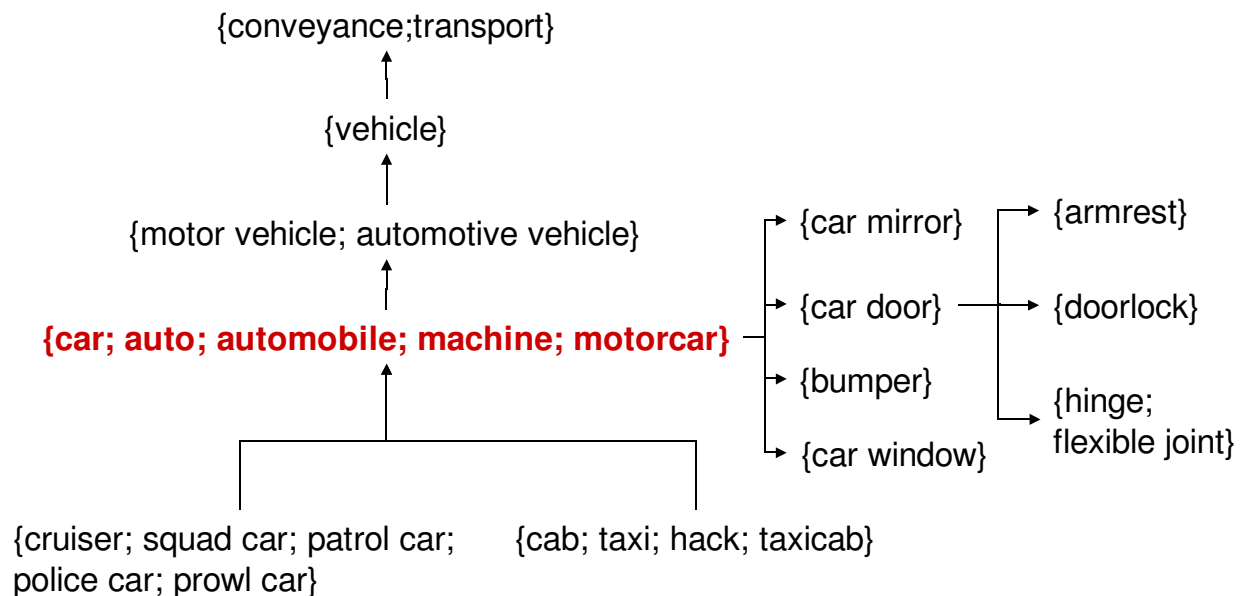
opmerkingen

selectie BW	1	commentaar 2004	
opmerkingen F			
opmerkingen R	'licentiaatsverhandeling' is de officiële term		

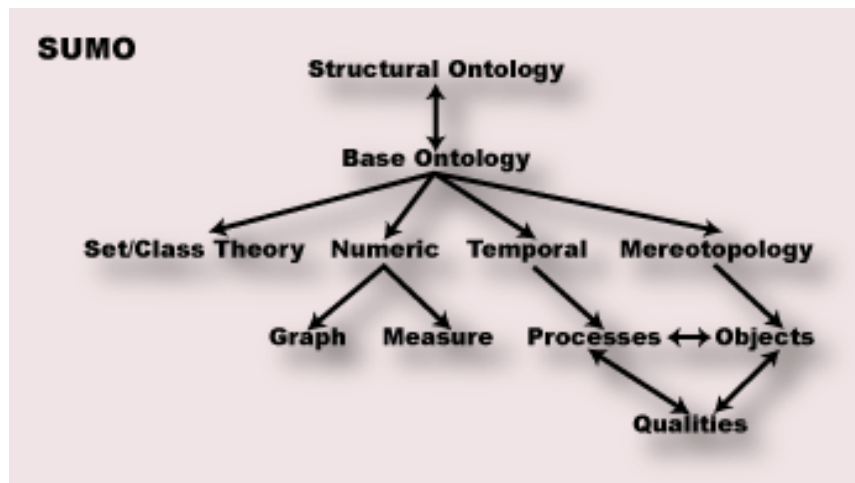
- **DWN:** DWN is organized around the notion of synsets. Synsets are a set of synonyms that represent a single concept. Typically in Wordnet, information is provided for the synset as a whole and not for the individual word meanings. For example, in Wordnet the synset has a single gloss but the different lexical units in RBN each have their own definition. From a Wordnet point of view, the definitions of lexical units that belong to the same synset should thus semantically be compatible or synonymous. For more information about DWN, please refer to EuroWordnet General Document, Vossen (2002).



- **PWN:** The Cornetto synsets have been mapped automatically to PWN which provides domain labels for Cornetto (for more information on the Cornetto-domains and how they are used in the SAT, see later in this section).



- **SUMO:** Outside the lexicon, an ontology provides a third layer of meaning. The Terms in an ontology represent the distinct types in a formal representation of knowledge. Terms can be combined in a knowledge representation language to form expressions of axioms. In principle, meaning is defined in the ontology independently of language but according to the principles of logic. In Cornetto, the ontology represents an independent anchoring of the relational meaning in Wordnet. The ontology is a formal framework that can be used to constrain and validate the implicit semantic statements of the lexical semantic structures, both the lexical units and the synsets. In addition, the ontology provides a mapping of a vocabulary to a formal representation that can be used to develop semantic web applications. For more information about SUMO please refer to <http://www.ontologyportal.org/>.



3.2.2 The design of Cornetto: Synsets & Lexical Units

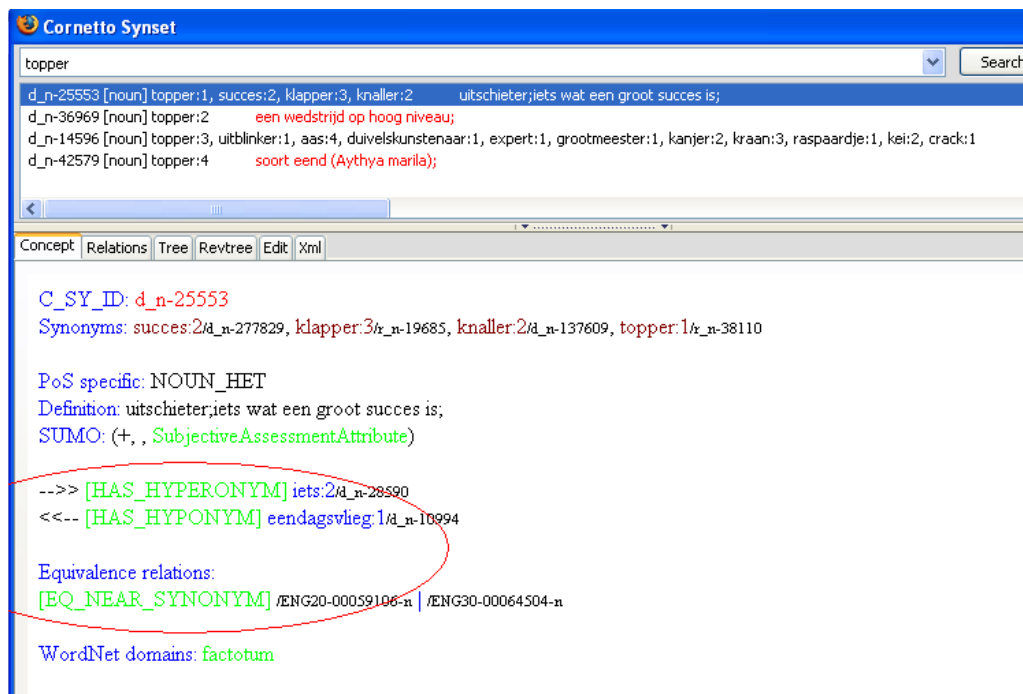
The Cornetto database is made up of two main databases: the Synset database and the Lexical Units database. Since the two databases have been merged automatically with each other and with other resources (such as Wordnet domains and SUMO) and only some of the data have been corrected manually, Cornetto still contains several mistakes: incorrect domain labels, 'empty' or redundant senses as a result of wrong mapping between synsets and LUs, examples linked with the wrong sense, mistakes concerning synonymy and other semantic relations.

3.2.2.1 Cornetto Synsets

(Synonyms – Relations – Domains)

The Cornetto synset database is organised around synsets and their semantic relations. A synset is a set of synonyms, i.e. words with the same part of speech that can be interchanged in a certain context. For example {car, motorcar, automobile, machine} form a synset because each word can be used to refer to the same concept. The synsets are related to each other by semantic relations like hyperonymy, hyponymy. Note that some semantic relations are used across part of speech.

The meaning of the synset is established by the set of synonyms and the relations to other synsets. Nevertheless, synsets can also have definitions (about 50% of the database). These definitions are amalgamated from the resume-fields of the lexical units that make up the synonyms. Other ways of defining the meaning are provided by the domain label and the mapping to Sumo. Finally, some synsets are marked as Base Concepts: the set of concepts that are used to build many different wordnets in the world (see the general EuroWordNet documentation for further details, Vossen 2002).



Cornetto Synset

topper

d_n-25553 [noun] topper:1, succes:2, klapper:3, knaller:2 uitschieter;iets wat een groot succes is;

d_n-36969 [noun] topper:2 een wedstrijd op hoog niveau;

d_n-14596 [noun] topper:3, uitblinker:1, aas:4, duivelskunstenaar:1, expert:1, grootmeester:1, kanjer:2, kraan:3, raspaardje:1, kei:2, crack:1

d_n-42579 [noun] topper:4 soort eend (Aythya marila);

Concept Relations Tree Revtree Edit Xml

C_SY_ID: d_n-25553

Synonyms: succes:2/d_n-277829, klapper:3/r_n-19685, knaller:2/d_n-137609, topper:1/r_n-38110

PoS specific: NOUN_HET

Definition: uitschieter;iets wat een groot succes is;

SUMO: (+, , SubjectiveAssessmentAttribute)

--> [HAS_HYPERONYM] iets:2/d_n-28590

<<-- [HAS_HYPONYM] eendagsvlieg:1/d_n-10994

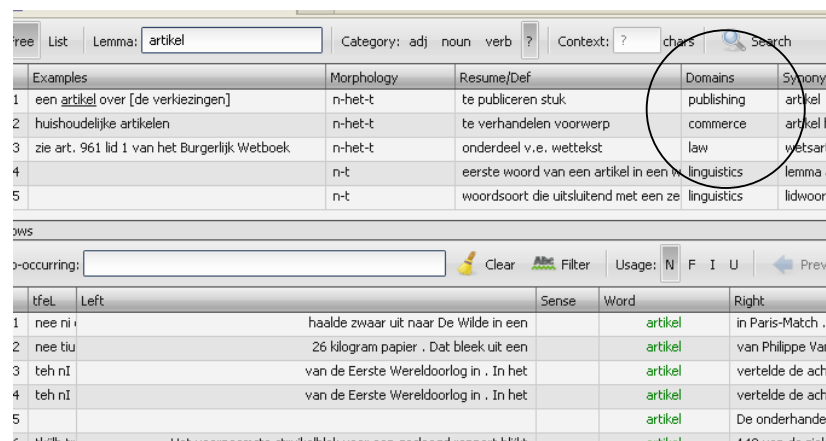
Equivalence relations:

[EQ_NEAR_SYNONYM] /ENG20-00059106-n | /ENG30-00064504-n

WordNet domains: factotum

Wordnet domains

[Wordnet Domain](#) labels have been assigned to Cornetto synsets. The assignment has been done by copying the domain labels from the English wordnet to the Dutch wordnet on the basis of the equivalence relations.

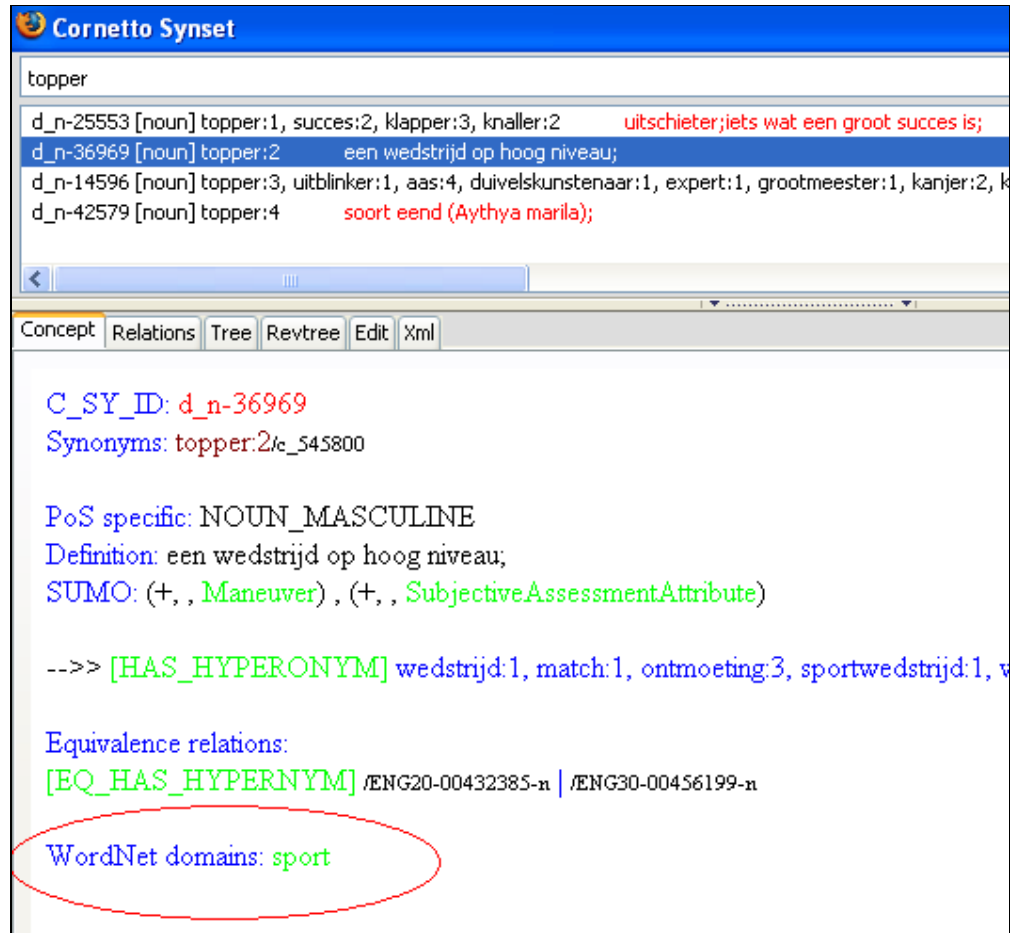


Examples	Morphology	Resume/Def	Domains	Synonym
1 een artikel over [de verkiezingen]	n-het-t	te publiceren stuk	publishing	artikel
2 huishoudelijke artikelen	n-het-t	te verhandelen voorwerp	commerce	artikel
3 zie art. 961 lid 1 van het Burgerlijk Wetboek	n-het-t	onderdeel v.e. wetttekst	law	wetsart
4	n-t	eerste woord van een artikel in een w.	linguistics	lemma
5	n-t	woordsoort die uitsluitend met een ze	linguistics	lidwoord

tfeL	Left	Sense	Word	Right
1	nee ni	haalde zwaar uit naar De Wilde in een	artikel	in Paris-Match
2	nee tiu	26 kilogram papier . Dat bleek uit een	artikel	van Philippe Var
3	teh ni	van de Eerste Wereldoorlog in . In het	artikel	vertelde de ach
4	teh ni	van de Eerste Wereldoorlog in . In het	artikel	vertelde de ach
5			artikel	De onderhande

The labels in Wordnet Domains come from the Dewey decimal system and form a hierarchy of 4 levels, which is shown in Appendix 5. We have also prepared a smaller list of domain labels in Dutch, mapping some of the English terms onto higher Dutch concepts. This resulted in a set which is a quarter in size of the original English set of labels.

The domain labels in Cornetto has been used to build domain classifiers. When the classifier is applied to the text, we can assign domain labels to text fragments. Typically, we apply the classifier to whole paragraphs.



Cornetto Synset

topper

d_n-25553 [noun] topper:1, succes:2, klapper:3, knaller:2 uitschieter;iets wat een groot succes is;

d_n-36969 [noun] topper:2 een wedstrijd op hoog niveau;

d_n-14596 [noun] topper:3, uitblinker:1, aas:4, duivelskunstenaar:1, expert:1, grootmeester:1, kanjer:2, k

d_n-42579 [noun] topper:4 soort eend (Aythya marila);

Concept Relations Tree Revtree Edit Xml

C_SY_ID: d_n-36969

Synonyms: topper:2/e_545800

PoS specific: NOUN_MASCULINE

Definition: een wedstrijd op hoog niveau;

SUMO: (+, , Maneuver) , (+, , SubjectiveAssessmentAttribute)

--> [HAS_HYPERONYM] wedstrijd:1, match:1, ontmoeting:3, sportwedstrijd:1, v

Equivalence relations:

[EQ_HAS_HYPERNYM] /ENG20-00432385-n | /ENG30-00456199-n

WordNet domains: sport

(See Appendix 5 for Domains in Cornetto + mapping onto a smaller number of Dutch domain labels)

3.2.2.2 Cornetto Lexical Units

The Cornetto Lexical Unit (LU) database is organized around lexical units: a word-meaning combination that represents a lexical form and a single meaning of this form. Each LU can have information on morphology, syntax, pragmatics, combinatorics, and example sentences that include collocations and idioms. For more information on the lexical units, see Documentatie Referentie Bestand Nederlands (RBN, Martin et al 2005).

Cornetto LU

topper

Search Search in all

r_n-38110 noun, topper:1, iets wat een groot succes is
 c_545800 noun, topper:2, een wedstrijd op hoog niveau
 r_n-38111 noun, topper:3, iemand die ergens in uitblinkt
 c_545801 noun, topper:4, soort eend (Aythya marila)

Form Examples Edit Xml

C_LU_ID: r_n-38110 C_SEQ_NR: 1
 [noun] topper

Morphology:
 type: derivation structure: top[*er] plurforms: toppers

Syntax:
 gender: m article: de

Semantics:
 reference: common countability: count type: nondynamic definition: XVD: dat veel belangstelling trekt, dat het hoogtepunt van iets vormt

Pragmatics:

Examples:
 Example: r_n-38110-1
 canonical form: die film/voorstelling is een absolute topper
 type: free

Querying a dictionary is complete. Item(s): 4

Morphology & Syntax

#	Examples	Morphology	Resume/Def	Domains	Synon
1	leren handschoenen	a	van leer	factotum schoo	onder
2	een vak leren	v-tr-onsch-nrefl	zich iets eigen maken	anatomy zoolog	leren
3	iemand leren typen	v-tr-onsch-nrefl	les geven	pedagogy	bijbre

Each lexical unit has values in the fields morphology & syntax. For the SAT we use the following values per word category:

Nouns

- **word category:** <form form-cat="noun" -> abbreviation: n
- **article:** <sy-article>de</sy-article>
- **count/ uncount:** <sem-countability>uncount</sem-countability> -> abbreviation: t = telbaar (count)/ nt = niet telbaar (uncount)

Adjectives

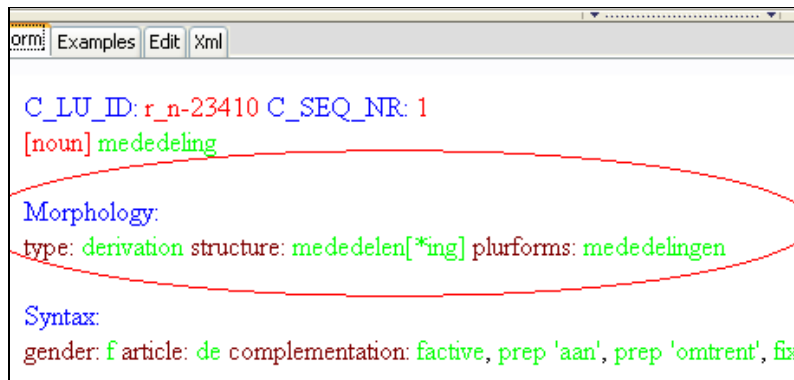
- **word category:** <form form-cat="adj" -> abbreviation: a

Manual for Semantic Annotation

Verbs

- **word category:** <form form-cat="verb" -> abbreviation: v
- **transitivity:** <sy-trans>tran</sy-trans> -> abbreviation: tr = transitief/ intr = intransitief
- **separability:** <sy-separ>onsch</sy-separ> -> abbreviation: sch = scheidbaar/ onsch = onscheidbaar
- **reflexivity:** <sy-reflexiv>nrefl</sy-reflexiv> -> abbreviation: refl = reflexief/ nrefl = niet reflexief

Of course, there is much more information on morphology and syntax in Cornetto as you can see in the following screenshots:



Examples Edit Xml

C_LU_ID: r_n-23410 C_SEQ_NR: 1

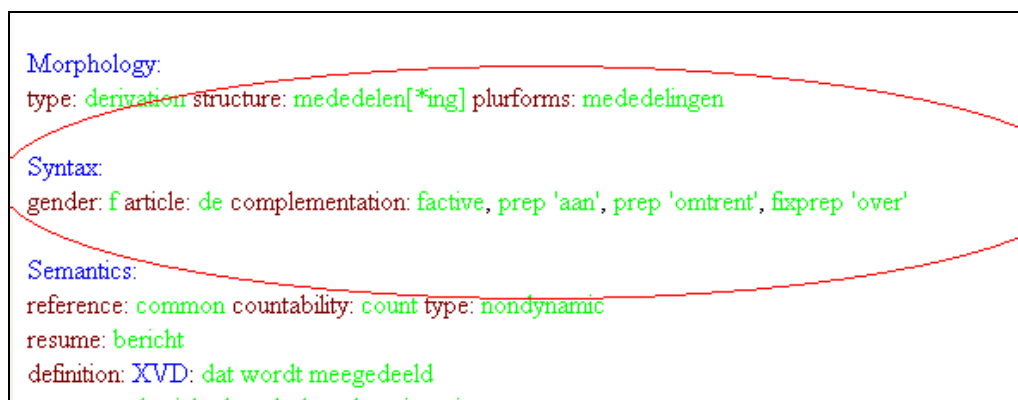
[noun] mededeling

Morphology:

type: derivation structure: mededelen[*ing] plurforms: mededelingen

Syntax:

gender: f article: de complementation: factive, prep 'aan', prep 'omtrent', fix



Morphology:

type: derivation structure: mededelen[*ing] plurforms: mededelingen

Syntax:

gender: f article: de complementation: factive, prep 'aan', prep 'omtrent', fix prep 'over'

Semantics:

reference: common countability: count type: nondynamic

resume: bericht

definition: XVD: dat wordt meegedeeld

Semantics

Examples	Morphology	Resume/Def	Domains	Synonyms	Relations
1 een kind aan de hand <u>leiden</u>	v-tr-onsch-nrefl	in een richting doen gaan	factotum	voeren leiden	meevoeren n
2 de weg leidt naar Alkmaar	v-intr-onsch-nrefl	voeren naar	factotum	voeren leiden lopen	zijn vormen u
3 iemand in verzoeking <u>leiden</u>	v-tr-onsch-nrefl	in een toestand brengen	factotum	veroorzaken teweeegbrengen lei	passeren afs
4 een vergadering <u>leiden</u>	v-tr-ansch-nrefl	de leiding hebben	factotum	leiden leidend	handelen
5 het peloton <u>leiden</u>	v-tr-ansch-nrefl	de voorste zijn	factotum	leiden	zijn vormen u
6 [nergens toe/tot niets]			factotum	uitdraaien uitlopen uitpakken	veranderen l

#	Resume/Def
1	in een toestand brengen

Left	Right
leidt	, acht het eenzijdige opzeggen juridisch ongegrond .
leidden	finiaal tot de recente afrekeningen waarin nu ook
leiden	. ' Het kan mensen tot chantage of
leidde	, Ernst Schaefer , was van Duitse origine
leidden	ertoe dat zwakke stroomstoten telkens één foton uitst
leiden	ertoe dat eerstgeborenen meer kans hebben om allerg
leidde	dat nochtans tot een verstrakking van de standpunter

Manual for Semantic Annotation

As far as the semantics in the SAT is concerned the most important data field is the **resume** or **definition**. Each edited LU has a resume that functions as a brief definition. These resumes are collided in the definitions of the synsets in which the LU takes part as a synonym.

Syntax:
gender: f article: de
Semantics:
reference: common countability: coll type: institut shift: 15:inst>hum subclass: institutie
resume: geheel aan ambtenaren v.d. openbare orde
definition: XVD: die toezicht houdt op het naleven van de wet, de veiligheid op straat en die overtreders van de wet opspoort
Examples:

Examples

Free	List	Lemma: leiden	Category: adj noun verb ?	Context: ?	chars	Search
#	Examples	Morphology	Resume/Def	Domains		
1	een kind aan de hand <u>leiden</u>	v-tr-onsch-nrefl	in een richting doen gaan	factotum		
2	de weg leidt naar Alkmaar	v-intr-onsch-nrefl	voeren naar	factotum		
3	iemand in verzoeking <u>leiden</u>	v-tr-onsch-nrefl	in een toestand brengen	factotum		
4	een vergadering <u>leiden</u>	v-tr-onsch-nrefl	de leiding hebben	factotum		
5	het peloton <u>leiden</u>	v-tr-onsch-nrefl	de voorste zijn	factotum		
6	[nergens toe/tot niets] <u>leiden</u>	v-tr-onsch-nrefl	resulteren in	factotum		

1 of 6 rows	Cornetto table events					
Co-occurring:	#	Example	Type	Usage: N F I U		
1	gelrevo	1 het peloton <u>leiden</u>	canonical		Word	Right
2	setev e	2 de competitie <u>leiden</u>	canonical		<u>leidt</u>	, acht h
3	laz nek	3 <u>leiden</u> met [twee tegen nul]	canonical		<u>leidden</u>	finaal to
4	nenab	4 de thuisclub leidt in de rust met twee tegen nul	textual		<u>leiden</u>	, 'Het k
5	eiD . ne				<u>leidde</u>	, Ernst S
6	maahci				<u>leidden</u>	ertoe de
					<u>leiden</u>	ertoe de
		4 rows				

The most important example features that will be discussed here are: **syntax example** and **semantics example**. The examples can be divided in **free examples** that intend to illustrate the meaning of the LU, and **fixed examples** like idioms, lexical collocations, grammatical collocations, proverbs and slogans.

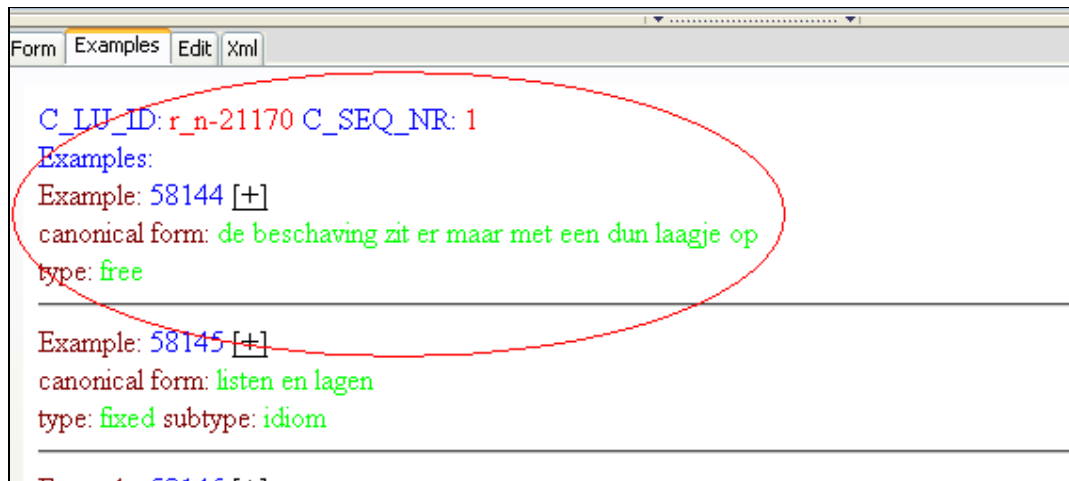
The feature 'syntax example' specifies if the syntactic type of example is free or fixed. For fixed examples the syntactic subtype specifies if the example is an idiom, lexical collocation etc. The first example shows a free example, the second is fixed.

1:laag/r_n-21170

(canonicalform: **de beschaving zit er maar met een dun laagje op**)

<syntax_example>

<sy-type>free</sy-type>
</syntax_example>



Form Examples Edit Xml

C_LU_ID: r_n-21170 C_SEQ_NR: 1

Examples:

Example: 58144 [+]

canonical form: de beschaving zit er maar met een dun laagje op

type: free

Example: 58145 [+]

canonical form: listen en lagen

type: fixed subtype: idiom

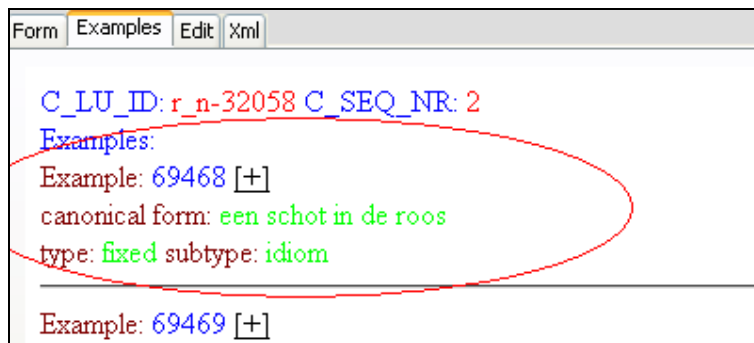
2:roos/r_n-32058

(canonicalform: een schot in de roos)

<syntax_example>

<sy-type>fixed</sy-type> <sy-
subtype>idiom</sy-subtype>

</syntax_example>



Form Examples Edit Xml

C_LU_ID: r_n-32058 C_SEQ_NR: 2

Examples:

Example: 69468 [+]

canonical form: een schot in de roos

type: fixed subtype: idiom

Example: 69469 [+]

Note that some of the features in the syntax example like ‘combinationword’ and ‘cat combinationword’ are not discussed here. See the RBN documentation (Martin e.a.(2005)) for all specifications.

Pre-processing and editing the Cornetto data

The editing work will be divided in various cycles. Each cycle is defined by a selection of LUs and synsets that is further divided among the editors. The selections represent different strategies of work, aiming at improving certain aspects of the data.

The editing process consists of the following main tasks:

1. Define a proper selection to work on;
2. Delineation of word meanings:
 - a. Apply ontological principles
 - b. Verify the alignment
 - c. Create new word meanings

Manual for Semantic Annotation

- d. Split word meanings
 - e. Merge word meanings
 - f. Remove word meanings
3. Specification of Lexical Units:
 - a. Morpho-syntactic information
 - b. Co-index frame elements
 - c. Examples
 - d. Semantic information
 - e. Add synonyms to synset
4. Specification of Synsets:
 - a. Ontological definition
 - b. Domain information
 - c. Synonyms
 - d. Lexical semantic relations
5. Specification of ontology terms:
 - a. Create a new Term
 - b. Provide the subclass relation for the Term
 - c. Provide axioms

These tasks are not necessarily carried out separately. In fact, editors could carry out any of these tasks at any point in the project cycle.

Separately of the editing, constraints will be formulated that restrict the possible combinations of information. These constraints will follow from the initial editing work. If the constraints are general enough, they can be implemented to either make suggestions to the editor (as scored values or relations) or excluded options.

Correcting sense-divisions or creating new senses for existing lemma in Cornetto

Based on the annotated occurrences in the corpus, our aim is to, if necessary, correct senses or create new additional senses in Cornetto. This happens using the following steps:

1. **Cluster senses based on corpus data** → using lexical-contextual clues and syntactic patterns within paragraph.
2. **Choose ‘prototypical sense’** (= ‘kernbetekenis’) → from cluster (based on frequency and intuition).
3. **Determine the different “shifts”** → This shows the meaning changes from ‘prototype sense’ to other senses (metaphor, metonymy etc.) and the sense divisions inside a sense inventory.
4. **If necessary *merge/ split* senses**

For further details on editing Cornetto refer to the Cornetto editing guidelines ([EditingGuideline_v1.doc](#)).

3.3 Snippet Tool

In addition to the existing text corpus, we use text fragments from the Internet (so-called “snippets”) in cases we lack a sufficient number of corpus examples for a sense. Copyright issues are avoided by restricting snippets to at most 155 characters that could be re-distributed, as declared in the legal document:

Manual for Semantic Annotation

<http://www.rechtspraak.nl/Gerechten/Rechtbanken/Alkmaar/Actualiteiten/Twee+uitspraken+in+kort+geding+betreffende+Jaap.nl.htm>

The snippets are especially needed to acquire sufficient examples for word meanings that are less frequent.

For the purpose of collecting Internet data and enriching the corpus, we have developed a so-called ‘Snippet Tool’ which queries the API of the search engine Webcorp (www.webcorp.org.uk). Webcorp has the advantage that it can leverage data in a text format, as a KWIC-index and with URL-links. The interface comes with simple and advanced features and the searches are highly customizable. Unfortunately no language specific search is supported by Webcorp at the moment. In order to filter out examples relating to other languages than Dutch, we limit our searches to .nl domains.

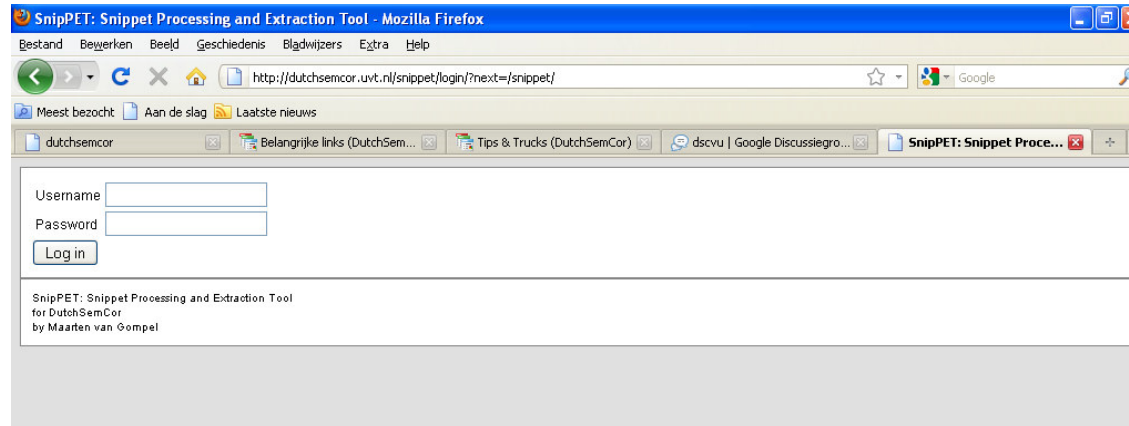


Webcorp – the linguistic search engine used in the Snippet Tool

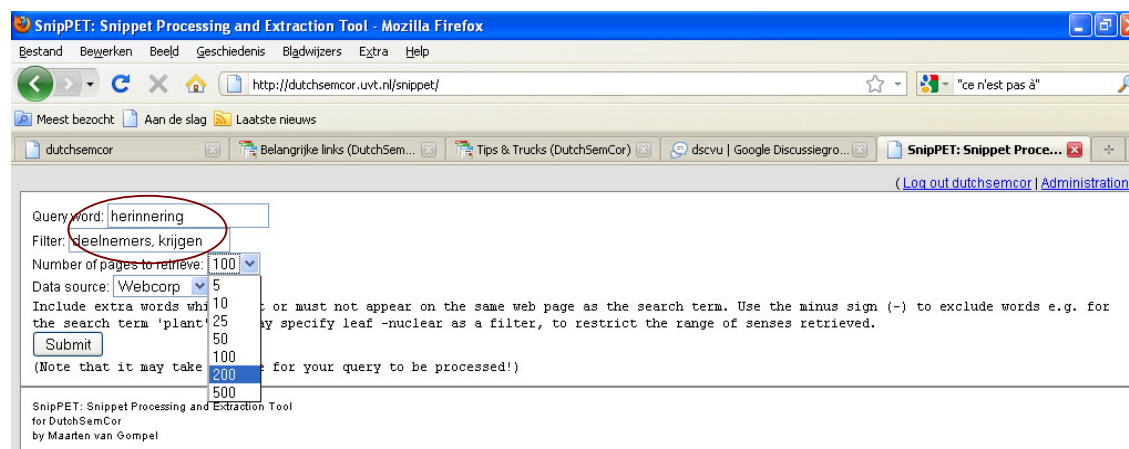
Another option is to choose for the data in the Open Taal project (see earlier). While the Webcorp option yields more data, the Open Taal project contains text which has been collected from the web in a directed way and has been linguistically pre-processed. After having made a selection, the annotators can export snippets from the Snippet Tool into the SAT.

In what follows we give a short introduction on how the Snippet Tool is used illustrating the different steps with screenshots. The tool can be found here: <http://dutchsemcor.uvt.nl/snippet/>

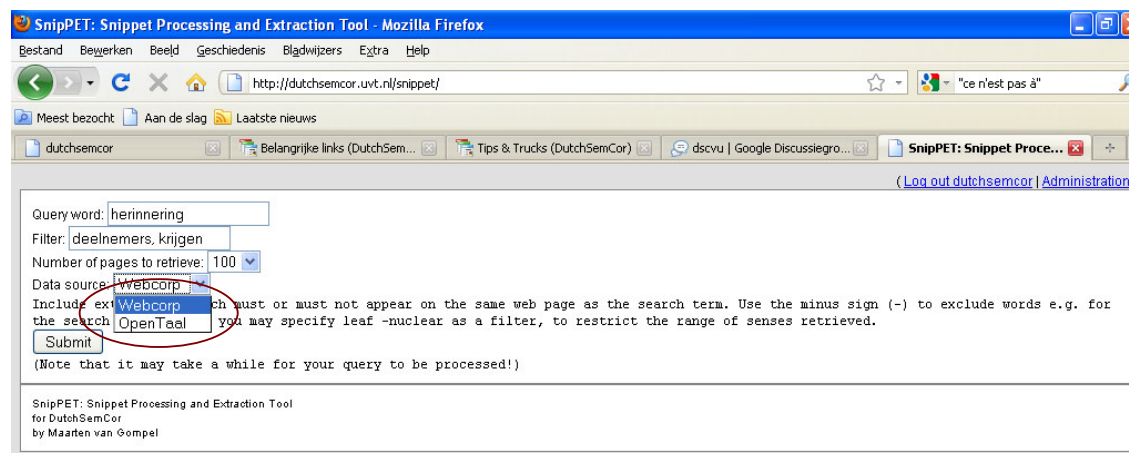
Manual for Semantic Annotation



Log in screen

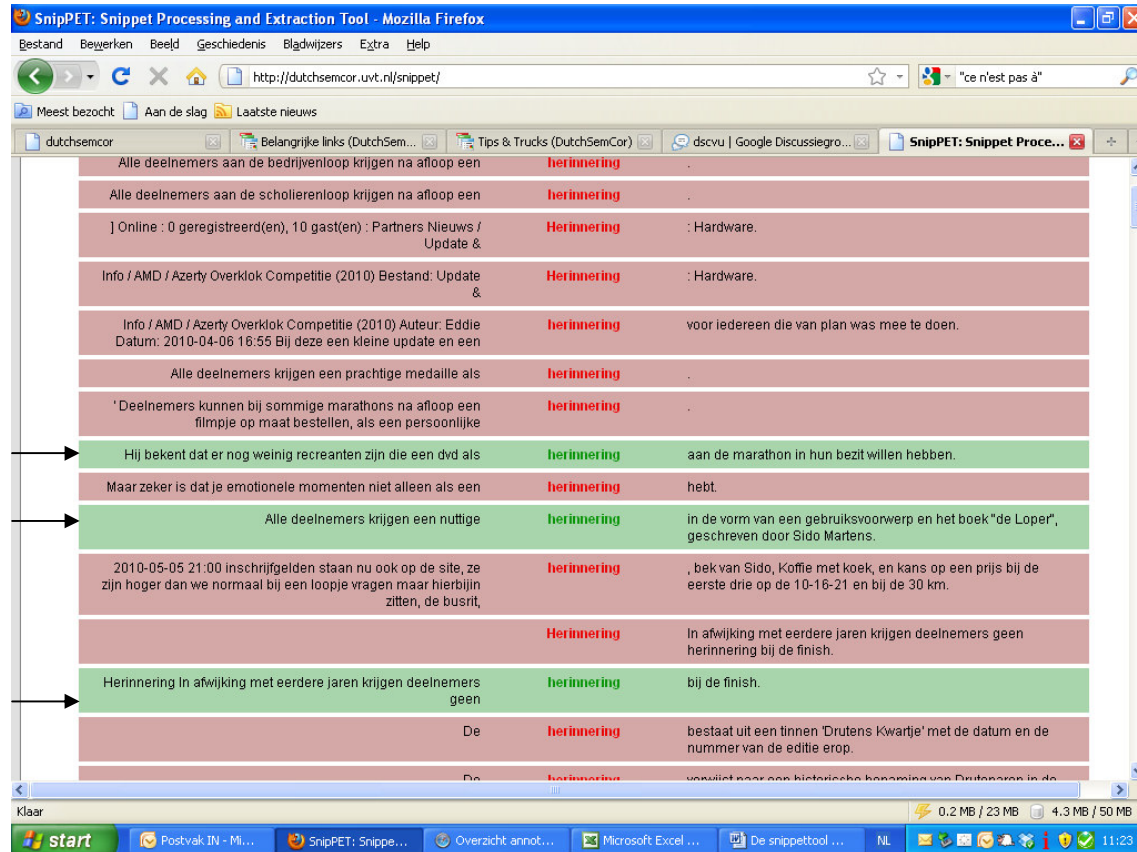


Typing in the target word and the filter word(s)

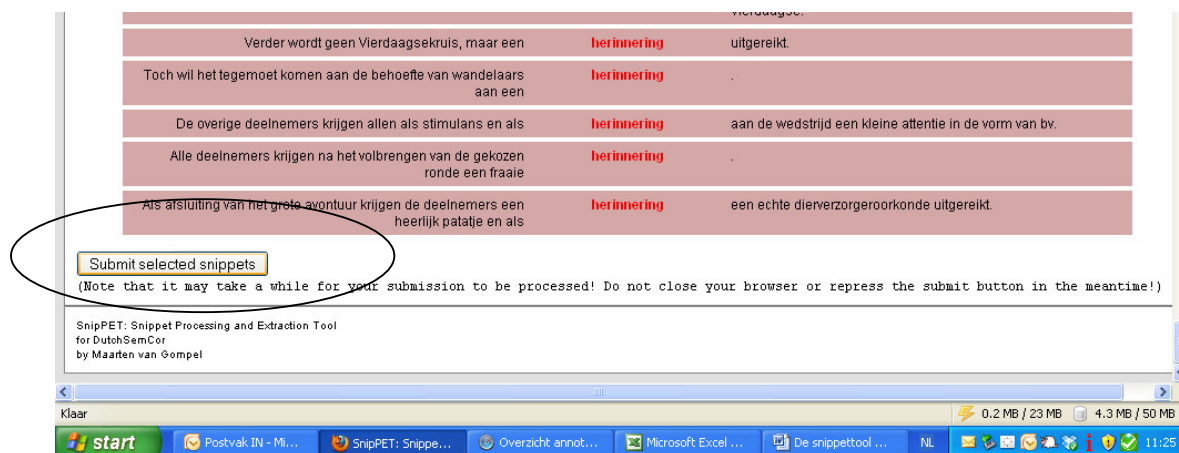


It is possible to search the whole internet (Webcorp) or a set of manually pre-processed examples from the web (Open Taal). Searching in Open Taal results in less noise but also in less data.

Manual for Semantic Annotation



The annotator selects the relevant examples (highlighted with green)



Clicking the "Submit selected snippets" button will export the examples into the SAT (i.e. into the corpus which is being annotated).

Manual for Semantic Annotation

Meest bezocht | Aan de slag | Laatste nieuws

dutchsemcor | Belangrijke links (DutchSemCor) | Tips & Tricks (DutchSemCor) | dscvu | Google Discussiegroep | SnipPET: Snippet Processing and Extraction Tool

(Log out dutchsemcor | Administratie)

Done processing, the following snippets have been successfully added. You can now [go back](#) and do another query if you want.

WR-U-E-S-0000000007.xml.il.k.tok.pos	Hij bekent dat er nog weinig recreanten zijn die een dvd als	herinnering	aan de marathon in hun bezit willen hebben.
WR-U-E-S-0000000008.xml.il.k.tok.pos	Alle deelnemers krijgen een nuttige	herinnering	in de vorm van een gebruiksvoorwerp en het boek "de Loper", geschreven door Sido Martens.
WR-U-E-S-0000000009.xml.il.k.tok.pos	Herinnering In afwijking met eerdere jaren krijgen deelnemers geen	herinnering	bij de finish.

SnipPET: Snippet Processing and Extraction Tool
for DutchSemCor
by Maarten van Gompel

The selected examples are exported from the Snippet Tool

dutchsemcor - Mozilla Firefox

Bestand | Bewerken | Beeld | Geschiedenis | Bladvijzers | Extra | Help

http://dutchsemcor.uvt.nl/sat/

Meest bezocht | Aan de slag | Laatste nieuws

dutchsemcor | Belangrijke links (DutchSemCor) | Tips & Tricks (DutchSemCor) | dscvu | Google Discussiegroep | SnipPET: Snippet Processing and Extraction Tool

Mode: Free | List | Buddy | Lemma: | Category: adj noun verb ? | Context: 75 chars | Source: All SoNaR | **Snippets** | CGN | DB Search | Config

| Examples | Morphosyntax | Resume/Def | Domains | Synonyms | Search in web snippets

0 rows

Co-occurring: Clear Filter Usage: N F I U | Paginate | ... of , chunk size: 50 | Tag senses

#	tfel	Left	Sense	Word	Right
0 rows					

After having logged in, the annotator chooses the corpus type option "Snippets"...

dutchsemcor - Mozilla Firefox

Bestand | Bewerken | Beeld | Geschiedenis | Bladvijzers | Extra | Help

http://dutchsemcor.uvt.nl/sat/

Meest bezocht | Aan de slag | Laatste nieuws

dutchsemcor | Belangrijke links (DutchSemCor) | Tips & Tricks (DutchSemCor) | dscvu | Google Discussiegroep | SnipPET: Snippet Processing and Extraction Tool

Mode: Free | List | Buddy | Lemma: | Category: adj noun verb ? | Context: 75 chars | Source: All SoNaR | **Snippets** | CGN | DB Search | Config

#	Examples	Morphosyntax	Resume/Def	Domains	Synonyms	Relations
1	ter herinnering aan [iets/iemand]	n-de-t	iets dat je je herinnert	psychology	reminiscentie	gedachte
2	zover de herinnering reikt	n-de-nit	vermogen om te herinneren	factotum	herinneringsvermogen	memorie
3	alle deelnemers aan de tocht krijgen een herinnering	n-de-t	aandenken	factotum	aandenken memorabilia	souvenir
4	Hij heeft al een aantal herinneringen van de bibliotheek gehad om het	n-t	schriftelijke aanmaning	administration	rappel	maning
5	Het herinneren gaat moeizaam na de schok	n-nit	het herinneren als proces	factotum	memorie geheugen	daad

5 rows

#	tfel	Left	Sense	Word	Right
1	sla dvd	Hij bekent dat er nog weinig recreanten zijn die een dvd als		herinnering	aan de marathon in hun bezit willen hebben.
2	egittun	Alle deelnemers krijgen een nuttige		herinnering	in de vorm van een gebruiksvoorwerp en het boek "de Loper", geschreven door
3	neeg si	Herinnering In afwijking met eerdere jaren krijgen deelnemers geen		herinnering	bij de finish.

3 rows

...enters the lemma and receives the three imported snippets to annotate.

3.4 Log file and the *loganalyser*

Each tagged sentence and every annotator action is recorded in a log file. Since every corpus fragment receives an ID it is possible to analyze the quality, diversity and quantity of the tagged instances.



SoNaR Tags, count is 147921, showing most recent 500

[Download all](#)

#	Word id	Word	Lemma	PoS	Source	Sense id	Annotator	Usage	Time stamp
147921	WR-U-E- S-0000012554.p.1.s.1.w.2	molen	molen	n	Snippet	c_545621	Laura	N	2010-07-29 01:24:08
147920	WR-U-E- S-0000012553.p.1.s.1.w.4	molen	molen	n	Snippet	c_545621	Laura	N	2010-07-29 01:24:08
147919	WR-U-E- S-0000012552.p.1.s.1.w.12	molen	molen	n	Snippet	c_545621	Laura	N	2010-07-29 01:24:08
147918	WR-U-E- S-0000012551.p.1.s.1.w.4	molen	molen	n	Snippet	c_545621	Laura	N	2010-07-29 01:24:08
147917	WR-U-E- S-0000012550.p.1.s.1.w.14	molen	molen	n	Snippet	c_545621	Laura	N	2010-07-29 01:24:08
147916	WR-U-E- S-0000012549.p.1.s.1.w.8	molen	molen	n	Snippet	c_545621	Laura	N	2010-07-29 01:24:01
147915	WR-U-E- S-0000012548.p.1.s.1.w.6	molen	molen	n	Snippet	c_545621	Laura	N	2010-07-29 01:24:01
147914	WR-U-E-	molen	molen	n	Snippet	c_545621	Laura	N	2010-07-29

Log file DutchSemCor

Data fields

#: number of annotation

Word id: identification code of the corpus example

Word: word form

Lemma: lemma

PoS: Part of Speech

Sense id: identification code of the sense as in the Cornetto LUs

Annotator: the name of the annotator

Usage: N= normal; F= figurative; I= Idiomatic; U= Unknown sense

Time stamp: date and time when the annotation was made

Analysis of the log file

Using a link (<http://ilk.uvt.nl/~mvgompel/dsclog.zip>) the coordinator of the project can download the log file in order to run an analysis program written in java called the *loganalyser*.

Microsoft Excel - dsc-tags2010.06.29 n. Chr. at 09_37_07 AM CEST.noun.overview [Alleen-lezen]

BestandBewerkenBeeldInvoegenOpmaakExtraDataVensterHelp

Typ een vraag voor hulp

Arial10

B*I*U

Output 1 of the loganalyser: overview

Microsoft Excel - dsc-tags2010.06.29 n. Chr. at 09_37_07 AM CEST.noun [Alleen-lezen]

Bestand Bewerken Beeld Invoegen Opmaak Extra Data Venster Help Typ een vraag voor hulp

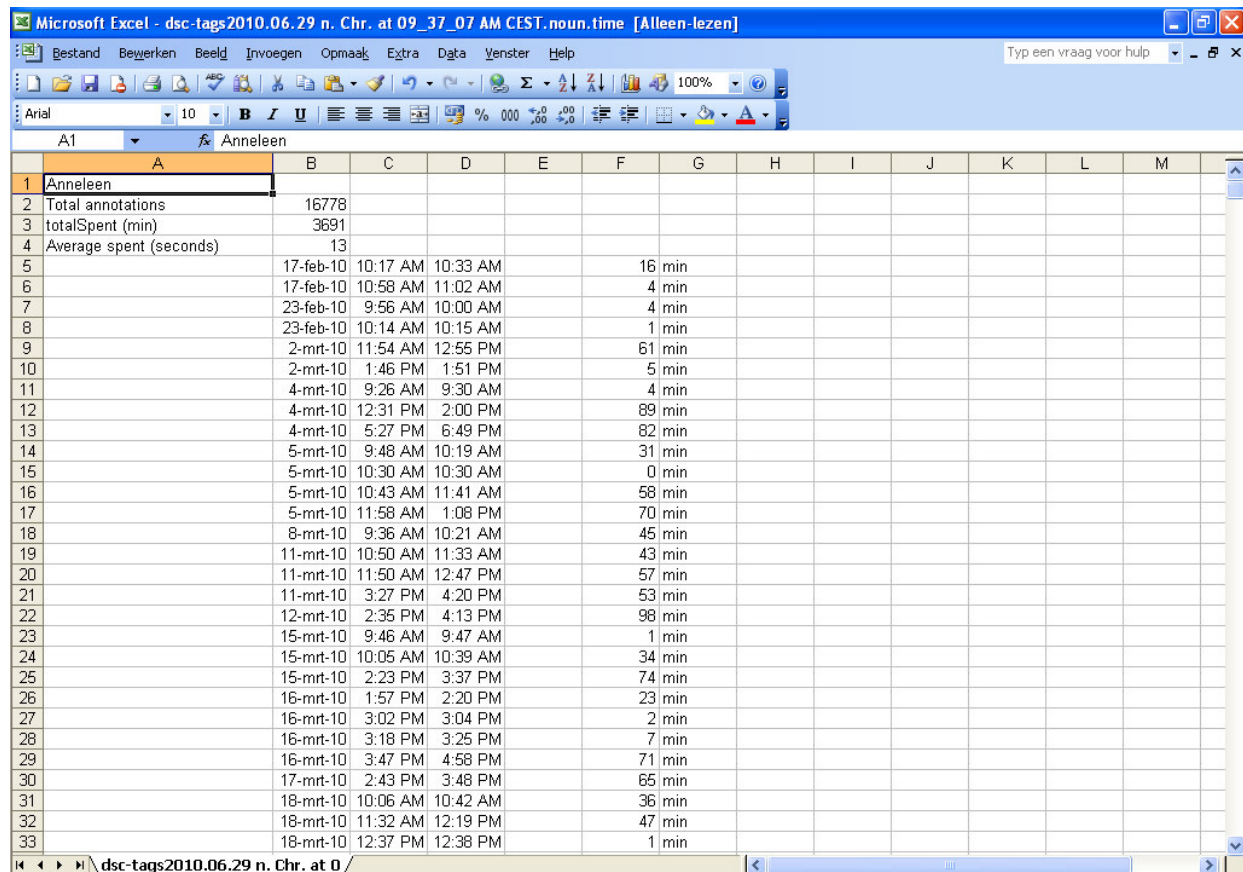
Arial 10 B I U % 000 0.00 0.00

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	president													
2	Nr of annot	2												
3	Nr of token	86												
4	Annotator	Nr tokens	Nr annos	Nr unique	Nr overlap	Nr Agree	Nr Disagree	Nr Partial	IAA exact	IAA weak	Start time	End time		
5	Anneleen	67	67	60	7	7	0	0	100	100	#####	#####		
6	Laura	26	26	19	7	7	0	0	100	100	#####	#####		
7														
8	Anneleen							Laura						
9	Nr tokens	Nr unique	Nr overlap	Nr Agree	Nr Disagree	Nr Partial		Nr tokens	Nr unique	Nr overlap	Nr Agree	Nr Disagree	Nr Partial	
10	r_n-29546	34	27	7	7	0	0	26	19	7	7	0	0	
11	r_n-29547	33	33	0	0	0	0	0	0	0	0	0	0	
12														
13	academie													
14	Nr of annot	2												
15	Nr of token	75												
16	Annotator	Nr tokens	Nr annos	Nr unique	Nr overlap	Nr Agree	Nr Disagree	Nr Partial	IAA exact	IAA weak	Start time	End time		
17	Lisanne	75	75	0	75	75	0	0	100	100	#####	#####		
18	Jonica	75	75	0	75	75	0	0	100	100	#####	#####		
19														
20	Lisanne							Jonica						
21	Nr tokens	Nr unique	Nr overlap	Nr Agree	Nr Disagree	Nr Partial		Nr tokens	Nr unique	Nr overlap	Nr Agree	Nr Disagree	Nr Partial	
22	r_n-2521	42	0	42	42	0	0	42	0	42	42	0	0	
23	r_n-2520	32	0	32	32	0	0	32	0	32	32	0	0	
24	c_545197	1	0	1	1	0	0	1	0	1	1	0	0	
25														
26	zoon													
27	Nr of annot	2												
28	Nr of token	46												
29	Annotator	Nr tokens	Nr annos	Nr unique	Nr overlap	Nr Agree	Nr Disagree	Nr Partial	IAA exact	IAA weak	Start time	End time		
30	Wilma	46	46	0	46	33	13	0	71	71	#####	#####		
31	Gratia	46	46	0	46	33	13	0	71	71	#####	#####		
32														
33	Wilma							Gratia						

dsc-tags2010.06.29 n. Chr. at 0

Output 2 of the loganalyser: detailed analysis

Manual for Semantic Annotation



The screenshot shows a Microsoft Excel spreadsheet titled "Microsoft Excel - dsc-tags2010.06.29 n. Chr. at 09_37_07 AM CEST.noun.time [Alleen-lezen]". The spreadsheet has columns A through M. Column A is labeled "Anneleen". Column B contains dates and times. Column C contains times. Column D contains times. Column E contains times. Column F contains times. Column G contains times. Column H contains times. Column I contains times. Column J contains times. Column K contains times. Column L contains times. Column M contains times. The data is organized into rows, with the first row (row 1) being the header row. The data rows start from row 2. The data is organized into rows, with the first row (row 1) being the header row. The data rows start from row 2. The data is organized into rows, with the first row (row 1) being the header row. The data rows start from row 2.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Anneleen												
2	Total annotations	16778											
3	totalSpent (min)	3691											
4	Average spent (seconds)	13											
5		17-feb-10	10:17 AM	10:33 AM			16 min						
6		17-feb-10	10:58 AM	11:02 AM			4 min						
7		23-feb-10	9:56 AM	10:00 AM			4 min						
8		23-feb-10	10:14 AM	10:15 AM			1 min						
9		2-mrt-10	11:54 AM	12:55 PM			61 min						
10		2-mrt-10	1:46 PM	1:51 PM			5 min						
11		4-mrt-10	9:26 AM	9:30 AM			4 min						
12		4-mrt-10	12:31 PM	2:00 PM			89 min						
13		4-mrt-10	5:27 PM	6:49 PM			82 min						
14		5-mrt-10	9:48 AM	10:19 AM			31 min						
15		5-mrt-10	10:30 AM	10:30 AM			0 min						
16		5-mrt-10	10:43 AM	11:41 AM			58 min						
17		5-mrt-10	11:58 AM	1:08 PM			70 min						
18		8-mrt-10	9:36 AM	10:21 AM			45 min						
19		11-mrt-10	10:50 AM	11:33 AM			43 min						
20		11-mrt-10	11:50 AM	12:47 PM			57 min						
21		11-mrt-10	3:27 PM	4:20 PM			53 min						
22		12-mrt-10	2:35 PM	4:13 PM			98 min						
23		15-mrt-10	9:46 AM	9:47 AM			1 min						
24		15-mrt-10	10:05 AM	10:39 AM			34 min						
25		15-mrt-10	2:23 PM	3:37 PM			74 min						
26		16-mrt-10	1:57 PM	2:20 PM			23 min						
27		16-mrt-10	3:02 PM	3:04 PM			2 min						
28		16-mrt-10	3:18 PM	3:25 PM			7 min						
29		16-mrt-10	3:47 PM	4:58 PM			71 min						
30		17-mrt-10	2:43 PM	3:48 PM			65 min						
31		18-mrt-10	10:06 AM	10:42 AM			36 min						
32		18-mrt-10	11:32 AM	12:19 PM			47 min						
33		18-mrt-10	12:37 PM	12:38 PM			1 min						

Output 3 of the loganalyser: time

The output of a loganalyser is basically a .csv file which can be opened in any spread sheet program like *excel* or *open office*. Running the program on the log file produces three separate .csv files with slightly different information.

The first overview (*Output 1*) is what we use the most often and which is the most useful for annotators: this file contains information on the annotated lemmas, the couples who annotated them, the weak and strong Inter-annotator Agreement (IA weak/strong) per lemma, the number of examples and annotations per senses, the coverage in percentage etc. Based on this overview, we can decide whether a lemma needs to be discussed and redone because of low IA or if some senses lacking a sufficient number of examples (coverage).

As opposed to the first overview, the second overview contains information on the annotations per annotator. This way one can examine the exact number of overlapping, unique etc. annotations as well as the number of *agree*, *disagree* and *partial agree* annotations.

Finally, the third overview provides useful information on the productivity of the annotators measuring the approximate time spent per annotation, a total since the beginning of the project, the total number of annotated examples per annotators and finally gives an overview of the exact time stamp in which the annotator has been busy annotating. With this output file we can get a more “humanistic” impression of the annotation process, examine habits and find ways to improve the quality of the annotations.

Evaluation of results: quantity, diversity and quality scores

Based on the Log file, annotation results can be evaluated and discussed with the annotators. Apart from measuring the number of tagged tokens (quantity), we are also focusing on the diversity of examples (diversity) and on the quality of annotations (IA minimum 80%).

Quantity-wise, we are aiming at 25 manually tagged tokens per sense which will be complemented with 75 tokens obtained by unsupervised techniques and bootstrapping. We are counting on an average of 3.4 senses per word (based on data in the Cornetto database) and take the 3000 most polysemous and most frequent words of the Dutch language (65% Nouns, 23% Verbs and 12% Adjectives). At the end of the three year project, we will deliver a semantically annotated Dutch corpus of circa 1 million tokens.

The **diversity** score is based on syntactic and semantic patterns which will be used to filter out frequently recurring constructions. Text fragments with a great syntactic and semantic diversity can better serve machine learning for WSD purposes. For instance, we wouldn't want to have 24 tagged examples of the lexical unit "lead" in the combination "lead the competition" and only one as in "lead the U.N.". Our aim is to offer as much diversity (syntactic and semantic) as possible. See also (Ng, 1997): "Recent work on intelligent example selection techniques suggest that the quality of the examples used for supervised learning can have a large impact on the classification accuracy of the induced classifier."

Finally, the IA-score is our **quality** measurement aid and is to show the percentage of Inter-annotator Agreement. It is also very useful for the discussions. Low agreement usually means difficulties either in linking the right examples to the existing Cornetto senses - "lumper" vs. "splitter" - or problems with the sense divisions of Cornetto itself. This latter will need to be corrected by the co-ordinator of the annotation project. As we have said, our aim is to reach a minimum of 80% IA.

4. Conclusion

In this guideline, we have tried to give a quick introduction to the tools & methods used in the DutchSemCor project. Our aim was to describe the different steps of the annotation process using screenshots in order to visualise and explain the different functions of the tools used. We realize that such a practical guide cannot contain a solid theoretical background. Still it is necessary to provide relevant literature which we ourselves have also used in the project bridging this way the gap between practice and theory. Please find for this purpose a short list of books, essays and links on WSD, Semantic Annotation and related matters.

Appendix 1 – Relevant links

DutchSemCor: <http://www2.let.vu.nl/oz/cltl/dutchsemcor/index.html>

Cornetto: <http://www2.let.vu.nl/oz/cltl/cornetto/index.html>

OntoNotes: <http://www.bbn.com/ontonotes/>

Senseval: <http://www.senseval.org/>

Semeval: <http://nlp.cs.swarthmore.edu/semeval/>

<http://semeval2.fbk.eu/semeval2.php>

Appendix 2 – Litterature

5.2.1 Books and articles

AGIRRE, E. AND EDMONDS, P., Eds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer, New York, NY.

AGIRRE, E. AND STEVENSON, M. 2006. Knowledge sources for WSD. In *Word Sense Disambiguation: Algorithms and Applications*, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 217–251.

AGIRRE, E. et al. 2006. A methodology for the joint development of the Basque WordNet and Semcor, In *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC)*

Bar-Hillel, Yehoshua. 1964. *Language and Information*. New York: Addison-Wesley

JURAFSKY, D. AND MARTIN, J. 2009. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, NJ.

KILGARRIFF, A. 1998. Senseval: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC, Granada, Spain)*. 1255–1258.

KILGARRIFF, A. 2006. Word senses. In *Word Sense Disambiguation: Algorithms and Applications*, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 29–46.

MIHALCEA, R. 2002a. Bootstrapping large sense tagged corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluations (LREC, Las Palmas, Spain)*.

MIHALCEA, R. 2002b. Word sense disambiguation with pattern learning and automatic feature selection. *J. Nat. Lang. Eng.* 8, 4, 348–358.

MIHALCEA, R. 2004. Co-training and self-training for word sense disambiguation. In *Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL, Boston, MA)*. 33–40.

MIHALCEA, R. 2006. Knowledge-based methods for WSD. In *Word Sense Disambiguation: Algorithms and Applications*, E. Agirre and P. Edmonds, Eds. Springer, New York, 107–131.

MIHALCEA, R. AND EDMONDS, P., Eds. 2004. *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain)*.

MIHALCEA, R. AND FARUQUE, E. 2004. Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3, Barcelona, Spain)*. 155–158.

MIHALCEA, R. AND MOLDOVAN, D. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI, Orlando, FL)*. 461–466.

NAVIGLI, R. 2009. Word Sense Disambiguation: a Survey. In *ACM Computing Surveys*, 41(2), (ACM Press). pp. 1-69.

Manual for Semantic Annotation

PALMER, M., NG, H. T., AND DANG, H. T. 2006. Evaluation of WSD systems. In *Word Sense Disambiguation: Algorithms and Applications*, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 75–106.

PIANTA, E. AND LUISA B. 2003. Translation as Annotation, In *Proceedings of the AI*IA 2003 Workshop 'Topics and Perspectives of Natural Language Processing in Italy'*, Pisa, Italy.

VOSSEN, P. 2006. Cornetto: Een lexicaal-semantische database voor taaltechnologie, Dixit Special Issue, Stevin.

VOSSEN, P., HOFMANN, K., DE RIJKE, M., TJONG KIM SANG, E., AND DESCHACHT, K. 2007. The Cornetto Database: Architecture and User-Scenarios. In *DIR*.pp.89-96.

VOSSEN, P., MAKES, I., SEGERS, R. AND H. VAN DER VLIET, H. 2008. Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database. In *Proceedings of LREC-2008, Marrakech, Morocco*.

Freely accessible manuals and articles on the Internet

Sense tagging

- [fellbaum_grabowski_landes_ANLP97.pdf](#)
- [Fellbaum_SemCorinvitedPaper.pdf](#)
- [HumbookCorpora.pdf](#)
- [Killgariff_and_Rosenzweig_athens.pdf](#)
- [LREC_BasqueWN_Semcor.pdf](#)
- [2005-K-Meaning-PCID.doc](#)
- [building_a_sense_tagged_corpus_with_open_306232.pdf](#): Semantic tagging with Open Mind Word Expert

WSD

- [Kilgarriff_Senseval_1998-K-CompSL.pdf](#)
- [WSDBOOK.zip](#)
- [Ng_W97-0201.pdf](#)
- [Resnik_Yarowsky_siglex97_perspective.pdf](#)
- [Veronis_2001-lancaster-sense.pdf](#)
- [Wilks_Stevenson_anlp97.pdf](#)
- [Stevenson_Wilks_J01-3001.pdf](#)
- [acl06.pdf](#): Word Sense and Subjectivity (Wiebe & Mihalcea) – the rol of subjectivity analysis in WSD
- [Using_Wikipedia_for_automatic_WSD.pdf](#): Wikipedia used for unsupervised WSD

TOOLS

- EAMT-2006-Baroni_WebBootCat.pdf: an introduction to the tools WebBootCat and SketchEngine
- Some background documents on Cornetto:
 - [DutchWordnet.doc](#)
 - [EWN_General.doc](#)
 - [EditingGuideline_v1.doc](#)

Appendix 3 - Domain labels

Domain labels in Cornetto		Domain labels in the SAT		Abbreviations
top				
doctrines		leerstellingen		wetensch.
archaeology		archeologie		gesch.
astrology		astrologie		ster.
history		geschiedenis		gesch.
	heraldry		heraldiek	gesch.
linguistics		taalkunde		taal.
	grammar		grammatica	taal.
literature		literatuur		lit.
	philology		filologie	lit.
philosophy		filosofie		fil.
psychology		psychologie		psy.
	psychoanalysis		psychoanalyse	psy.
art		kunst		kunst
	dance		dans	kunst
	drawing		tekening	kunst
			schilderij	kunst
	painting		filatelie	kunst
	philately			kunst
	music		muziek	kunst
	photography		fotografie	kunst
	plastic_arts		beeldende kunsten	kunst
			sieraden	kunst
	jewellery		numismatiek	kunst
	numismatics		sculptuur	kunst
	sculpture			kunst
	theatre		theater	kunst
religion		religie		rel.
	mythology		mythologie	rel.
	occultism		occultisme	rel.
	roman_catholic		katholicisme	rel.
	theology		theologie	rel.
free_time		vrijetijd		sp.

Manual for Semantic Annotation



applied_science	play		spelen	
	betting		wedden	sp.
	card		kaart	sp.
	chess		schaken	sp.
	sport		sport	sp.
	badminton		badminton	sp.
	baseball		baseball	sp.
	basketball		basketbal	sp.
	cricket		cricket	sp.
	football		Amerikaansvoetbal	sp.
	golf		golf	sp.
	rugby		rugby	sp.
	soccer		voetbal	sp.
	table_tennis		tafeltennis	sp.
	tennis		tennis	sp.
	volleyball		volleybal	sp.
	cycling		fietsen	sp.
	skating		schaatsen	sp.
	skiing		skiën	sp.
	hockey		hockey	sp.
	mountaineering		alpinisme	sp.
	rowing		roeien	sp.
	swimming		zwemmen	sp.
	sub		sub	sp.
	diving		duiken	sp.
	racing		racing	sp.
	athletics		atletiek	sp.
	wrestling		worstelen	sp.
	boxing		boksen	sp.
	fencing		schermen	sp.
	archery		boogschieten	sp.
	fishing		visserij	sp.
	hunting		jacht	sp.
	bowling		bowlen	sp.
	showjumping		springconcours	sp.
		toegepastewetenschap		wetensch.

Manual for Semantic Annotation



pure_science	agriculture		landbouw	landb.
	alimentation		voeding	voed.
				voed.
	gastronomy		gastronomie	bouwk.
	architecture		architectuur	bouwk.
				bouwk.
	town_planning		stedenbouw	bouwk.
	building_industry		bouwnijverheid	bouwk.
	furniture		meubilair	bouwk.
	computer_science		informatica	comp.
	engineering		techniek	tech.
				tech.
	mechanics		mechanica	tech.
	astronautics		ruimtevaart	tech.
	electrotechnics		electrotechniek	tech.
	hydraulics		hydraulica	tech.
	medicine		geneeskunde	med.
				med.
	dentistry		tandheelkunde	med.
	pharmacy		farmacie	med.
	psychiatry		psychiatrie	med.
	radiology		radiologie	med.
	surgery		chirurgie	med.
	veterinary		diergeneeskunde	med.
				med.
	zootchnics		zoötechniek	wetensch.
	astronomy		astronomie/ sterrenkunde	ster.
				ster.
	topography		topografie	biol.
	biology		biologie	biol.
				biol.
	biochemistry		biochemie	biol.
	ecology		ecologie	biol.
	botany		plantkunde	biol.
	zoology		zoölogie	biol.
		entomology		biol.
	anatomy		anatomie	biol.
	physiology		fysiologie	biol.
	genetics		genetica	biol.
	chemistry		chemie	schei.
	earth		geowetenschappen	geo.
				geo.
	geology		geologie	

Manual for Semantic Annotation



		meteorology				meteorologie		geo.
		oceanography				oceanografie		geo.
		paleontology				paleontologie		geo.
		geography				geografie		geo.
	mathematics					wiskunde		wisk.
		geometry				geometrie		wisk.
		statistics				statistieken		wisk.
	physics					natuurkunde		nat.
		acoustics				akoestiek		nat.
		atomic_physic				atoomfysica		nat.
		electricity				elektriciteit	elektronica	nat.
		electronics						nat.
		gas				gas		nat.
		optics				optica		nat.
social_science					socialewetenschappen			soc.
	anthropology					antropologie		soc.
		ethnology				etnologie	folklore	soc.
		folklore						soc.
body_care					lichaamsverzorging			lich.
military					militair			mil.
pedagogy					pedagogie			ped.
	school					school		ped.
	university					universiteit		ped.
publishing					uitgeverij			media
sociology					sociologie			soc.
telecommunication					telecommunicatie			media
	cinema					filmkunst		media
	post					post		media
	radio					radio		media
	telegraphy					telegrafie		media
	telephony					telefonie		media
	tv					televisie		media
artisanship					vakmanschap			kunst
commerce					commerce			handel
industry					industrie			ind.
	textiles					textiel		ind.

Manual for Semantic Annotation



transport

aeronautic
auto
merchant_navy
railway

economy

banking
book_keeping
enterprise
exchange
insurance
money
tax

administration

law

politics

diplomacy

tourism

fashion

sexuality

factotum

number
color
time_period
person
quality
metrology
state

vervoer

economie

administratie

recht

politiek

toerisme

mode

seksualiteit

algemeen

luchtvaart
auto
scheepvaart
spoorwegen
bankieren
boekhouding
onderneming
uitwisseling
verzekering
geld
belasting

diplomatie

nummer
kleur
tijd/periode
persoon
kwaliteit
metrologie
toestand

trans.

trans.

trans.

trans.

trans.

ec.

ec.

ec.

ec.

ec.

ec.

ec.

ec.

adm.

jur.

pol.

pol.

toer.

mode

sex

alg.

alg.

alg.

alg.

alg.

alg.

alg.

alg.