

CORNETTO

(TENDER)

Combinatorial and Relational Network as Toolkit for Dutch Language Technology

Piek Vossen^{1, 2}, Isa Maks¹, Roxane Segers¹, Hennie van der Vliet¹

¹: Faculty of Arts, Vrije Universiteit Amsterdam

²: Irion Technologies, Delft

ru | STEVIN



Overview

- Project overview
- WP 01: database
- WP 02: alignment
- WP 03: ontology
- WP 04: editing
- *(WP 05: acquisition toolkit)*
- *(WP 06: domain acquisition)*
- *(WP 07: evaluation)*
- Results so far

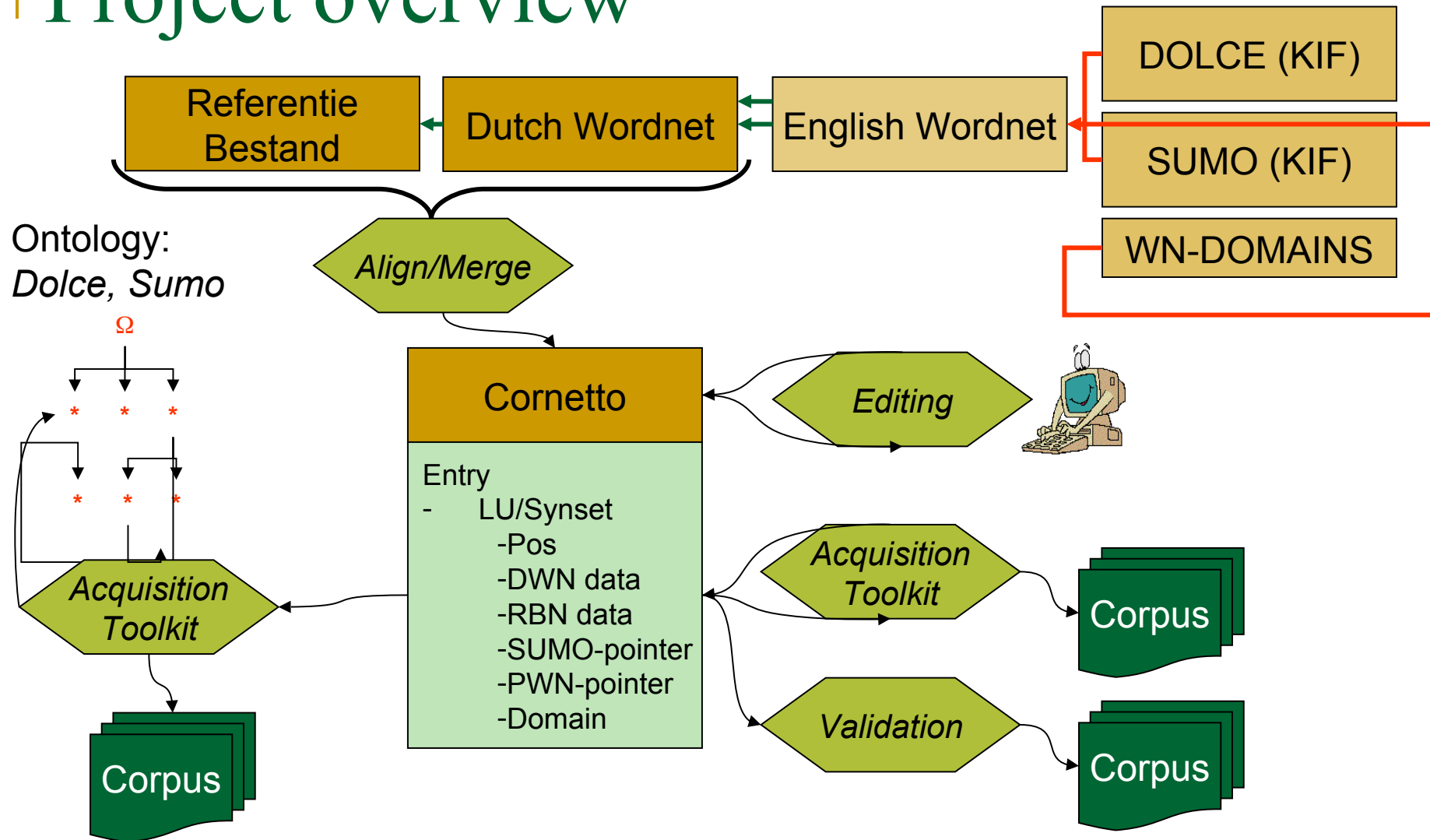
Goals of the Cornetto project

- Goal: to develop a lexical semantic database for Dutch:
 - 40K Entries: generic and central part of the language
 - Rich horizontal and vertical semantic relations
 - Combinatoric information
 - Ontological information
- April 2006-March 2008, extended to July 2008
- The data of the final results of the Cornetto project will be available through the TST-centrale of the Nederlandse Taalunie (free for research).

Approach

- Combine the information from two existing Dutch lexical resources:
 - The Dutch wordnet (DWN): synsets and lexical semantic relations
 - The Referentiebestand Nederlands (RBN): morpho-syntactic information, semantic information, pragmatic information, frame structures, lexical functions and combinatorics
- Link to English WordNet
- Link to Wordnet Domains
- Link to SUMO

Project overview



Data overview

- **Form Units in the Cornetto database:**
(nouns, verbs and adjectives)

■ FUs both in RBN and DWN	37.000
■ FUs in RBN only	6.000
■ FUs in DWN only	47.000
■ Total number FUs in Cornetto	90.000

Polysemy distributions after

align

Nr. of senses after alignment	Nr. of words (nouns)	Nr. of senses
10+	62	620
9	41	369
8	75	600
7	126	882
6	212	1272
5	389	2235
4	1026	4104
3	2507	7521
2	7293	14.586
1	64.298	64.298



WP 01: Database

WP 01: Database

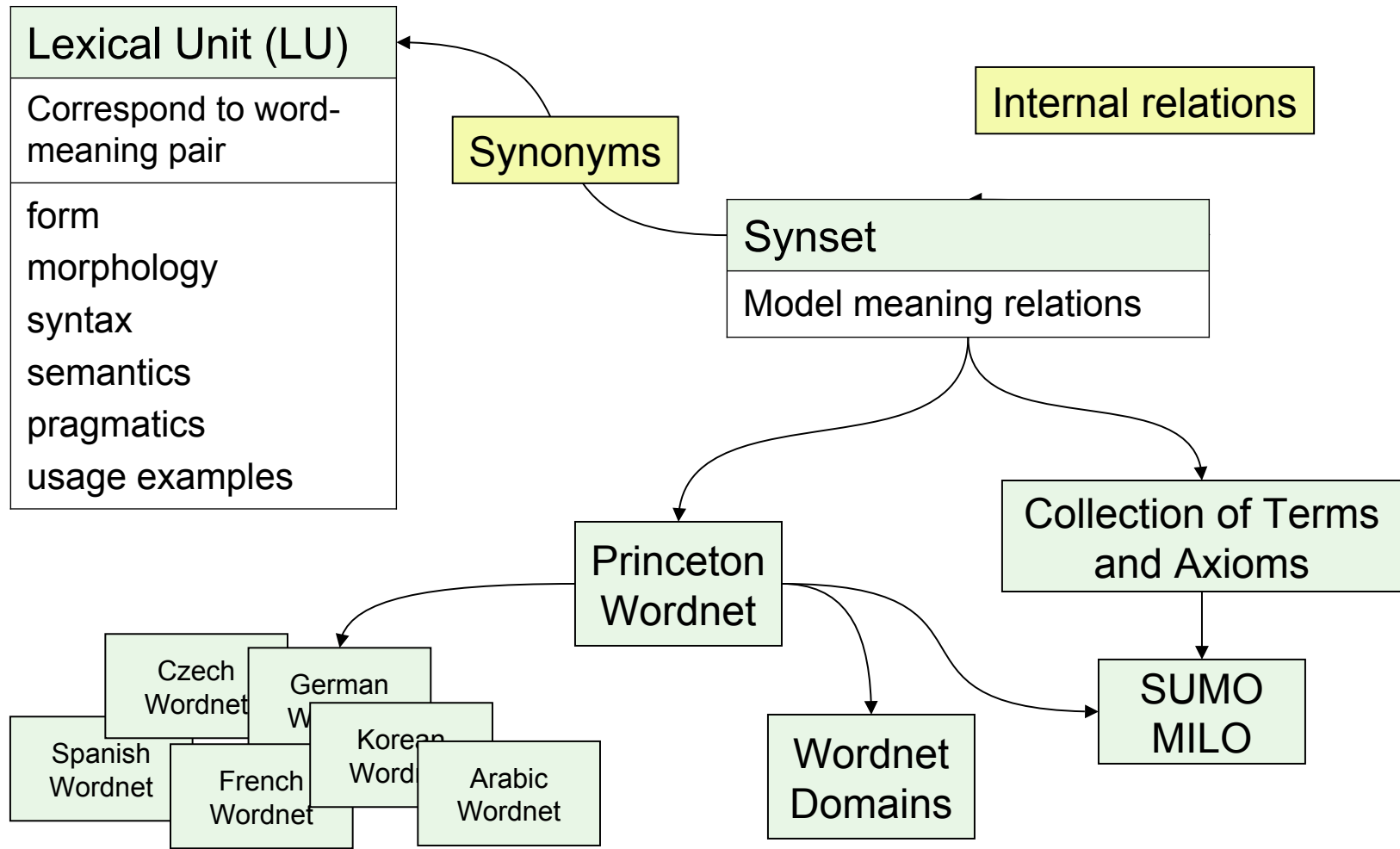
■ Collections:

- Lexical Units (LU): mainly derived from the RBN
- Synsets (SY): mainly derived from DWN
- Terms (TE) and axioms: mainly derived on SUMO and MILO
- Domains (DM): based on Wordnet domains

■ Mappings:

- LU $\leftrightarrow$ SY
- SY $\leftrightarrow$ SY (within Dutch and from Dutch to English)
- SY $\leftrightarrow$ TE
- SY $\leftrightarrow$ DM

WP 01: Data Organization



- Implemented in DebVisDic:

- <http://deb.fi.muni.cz/index.php>

- Demo version already available:

[http://www.let.vu.nl/onderzoek/projectsites/cornetto/
demo.html](http://www.let.vu.nl/onderzoek/projectsites/cornetto/demo.html)

water

Search

Search in all

d_n-20623 [noun] water:1, majem:1 SCHEIK

d_n-26238 [noun] water:2 AARDR rivier, s

d_n-44037 [noun] water:3 in algemene z

d_n-34850 [noun] water:4, lichaamsvocht:1

Concept

Relations

Tree

Revtree

Edit

Xml

C_SY_ID: d_n-20623

Synonyms: majem:1/d_n-166370,

water:1/r_n-42415

PoS specific: NOUN_HET

Definition: kleur- en smaakloze

vloeistof waarvan de moleculen

uit 2 atomen waterstof en 1

atoom zuurstof bestaan

Differentiae: 0

VLIS domains: SCHEIK

SUMO: (+, , Water)

-->> [HAS_HYPERONYM]

vloeistof:1/d_n-20474

<<-- [HAS_HYPONYM]

regenwater:1/d_n-19186

water

Search

Search in all

r_n-42415 noun, water:1, vocht

r_n-42416 noun, water:2, waterloop

d_n-330970 noun, water:3, H2O als eleme...

d_n-330971 noun, water:4, afgescheiden ...

Form

Examples

Edit

Xml

C_LU_ID: r_n-42415

C_SEQ_NR: 1

[noun] water

Morphology:

type: simpmorph

Syntax:

gender: n article: het

Semantics:

reference: common

countability: uncount

type: substance

subclass: natuurlijke stof

resume: vocht

definition: XVD: die helder en

water

Search

Search in all

water:1, NOUN, LU=r_n-42415, SY=d_n-20623,, MC

water:2, NOUN, LU=r_n-42416, SY=d_n-26238,, MC

water:3, NOUN, LU=d_n-330970, SY=d_n-44037,, MC

water:4, NOUN, LU=d_n-330971, SY=d_n-34850,, MC

Preview

Edit

Xml

CID_ID: 156793 SEQ_NR: 1

Form: water

PoS: NOUN

Score:

Status: manual

LU ID: r_n-42415

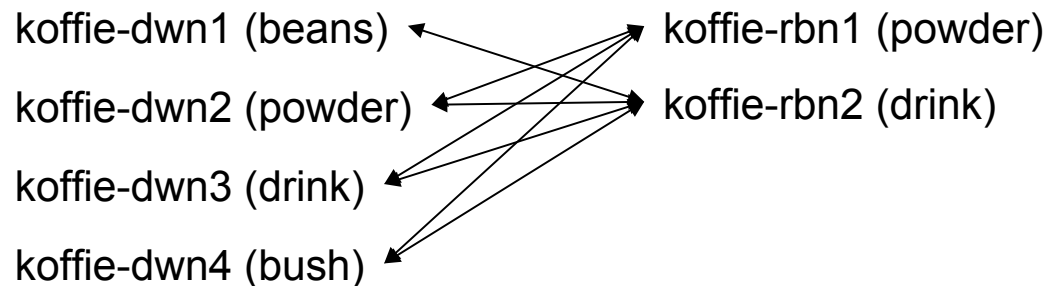
Synset ID: d_n-20623

WP 02: alignment



Alignment of Lexical Units NOT synsets

- Generate all weighted combinations, e.g. koffie (coffee):



Weighted scores and dummy LU

The image displays three windows from the Cornetto software interface, illustrating the relationship between a Synset, a Lexical Unit (LU), and an Instance (ID) for the word 'hoed'.

Cornetto Synset: Shows the synset for 'hoed'. The selected entry is 'd_n-17238 [noun] hoed:3 op een hoed'. The interface includes tabs for Concept, Relations, Tree, Revtree, Edit, and Xml. The details pane shows:
C_SY_ID: d_n-17238
Synonyms: hoed:3/d_n-110164
PoS specific: NOUN_MASCULINE
Definition: op een hoed lijkend voorwerp
Differentiae: dat op een hoed lijkt
SUMO: (+, , BodyPart)

Cornetto LU: Shows the Lexical Unit for 'hoed'. The selected entry is 'r_n-16890 noun, hoed:2, bovenste deel v...'. The interface includes tabs for Form, Examples, Edit, and Xml. The details pane shows:
C_LU_ID: r_n-16890
C_SEQ_NR: 2
[noun] hoed
Morphology: type: simpmorph, plurforms: hoeden
Syntax: gender: m article: de

Cornetto ID: Shows the Instance for 'hoed'. The interface includes tabs for Preview, Edit, and Xml. The details pane shows:
CID_ID: 19764 SEQ_NR: 1
Form: hoed
PoS: NOUN
Score: 79.95
Status:
LU ID: r_n-16890
Synset ID: d_n-11620

WP 04: Editing of the database

Main focus/strategies

- Focus on high-polysemous and high-frequent words
- 1800 most polysemous nouns (4 or more meanings)
- 1000 frequent nouns with a semantic shift
- 1000 frequent verbs
- 250 most frequent adjectives
- Extra: words with many equivalence relations to EWN.

Basic editing steps

- Checking and editing the mapping of RBN and DWN
- Checking and editing the mapping of DWN to English Wordnet 2.0
- Checking and editing the mapping of DWN to Wordnet Domain labels
- Checking and editing the mapping of DWN to SUMO and MILO

- The total of 3250 words in these sets equal to approximately 10.000 word meanings and cover the most frequent part of the Dutch language.
- Core Cornetto: high quality, manually checked, edited and confirmed.

WP 03: Ontology



SUMO/MILO as ontological framework

- Based on pragmatic grounds:
 - availability, size, coverage
 - linking to English Wordnet
 - mapping to other Wordnet-like projects

Mapping to SUMO

- Subsumption, equivalence, instance

Tea (drink) (+,, Tea)

Tea (shrub) (+,, FloweringPlant)

Date (fruit) (=,, Datefruit)

Rotterdam (instance,, City)

Some results



Quality labelling in the Cornetto database

- Cornetto database: **all data** in three subsets:
 - **Core Cornetto**: high quality, manually revised, central part of the Dutch language
(at least 2000 words, corresponding to 10.000+ meanings)
 - **Cornetto A**: automatically inserted quality labels in monoseme and biseme records that are not manually revised. (approximately 27.000 words)
 - **Cornetto B**: words that were not in RBN or DWN and therefore could not be aligned.
(approximately 53.000 words)

Preliminary database is already used

- Licenses for Rijksuniversiteit Groningen, Universiteit van Utrecht, Princeton University.
- More information:

<http://www.let.vu.nl/onderzoek/projectsites/cornetto/index.html>



The end....