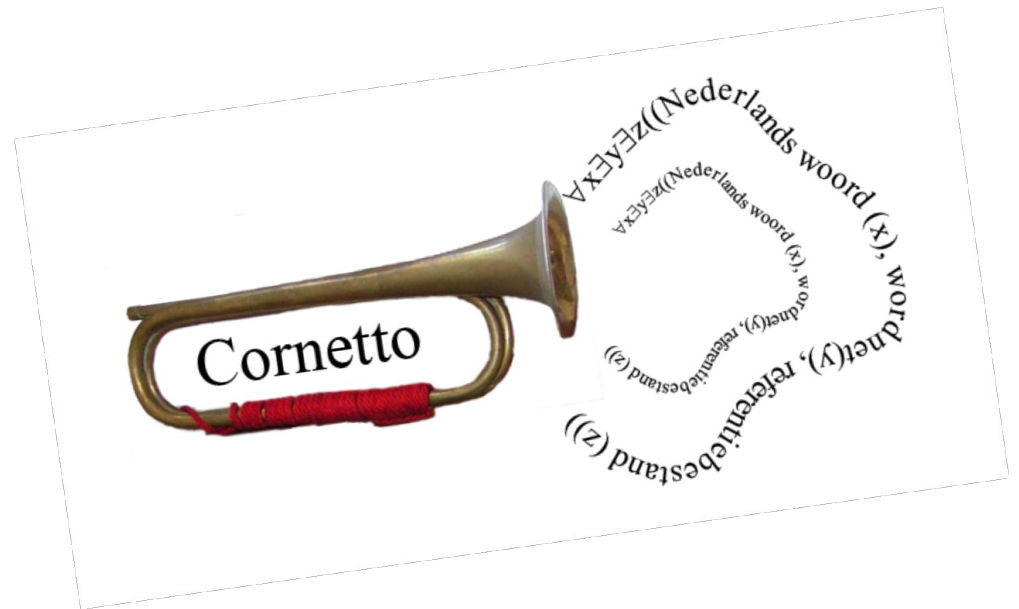


CORNETTO

(TENDER)

Combinatorial and Relational Network as
Toolkit for Dutch Language Technology



Project Cornetto

Consortium partners

VUA (Vrije Universiteit Amsterdam, General Linguistics Department)

UvA (University of Amsterdam, Informatics Institute)

K.U. Leuven (Katholieke Universiteit Leuven, Department of Computer Science)

Irion Technologies BV Delft

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

Project overview

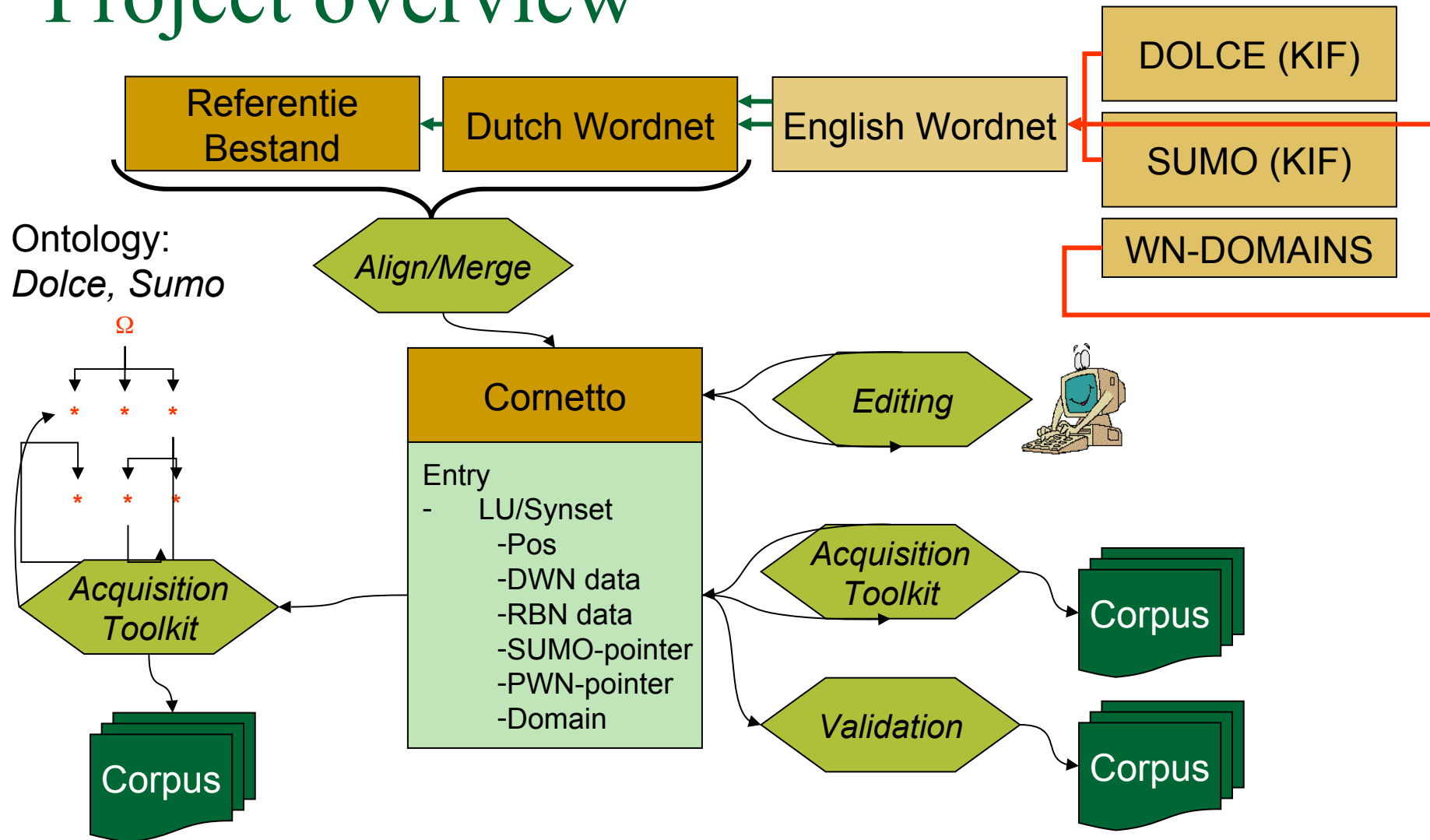
Goals of the Cornetto project

- 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

WP 01: Database

WP 01: Database

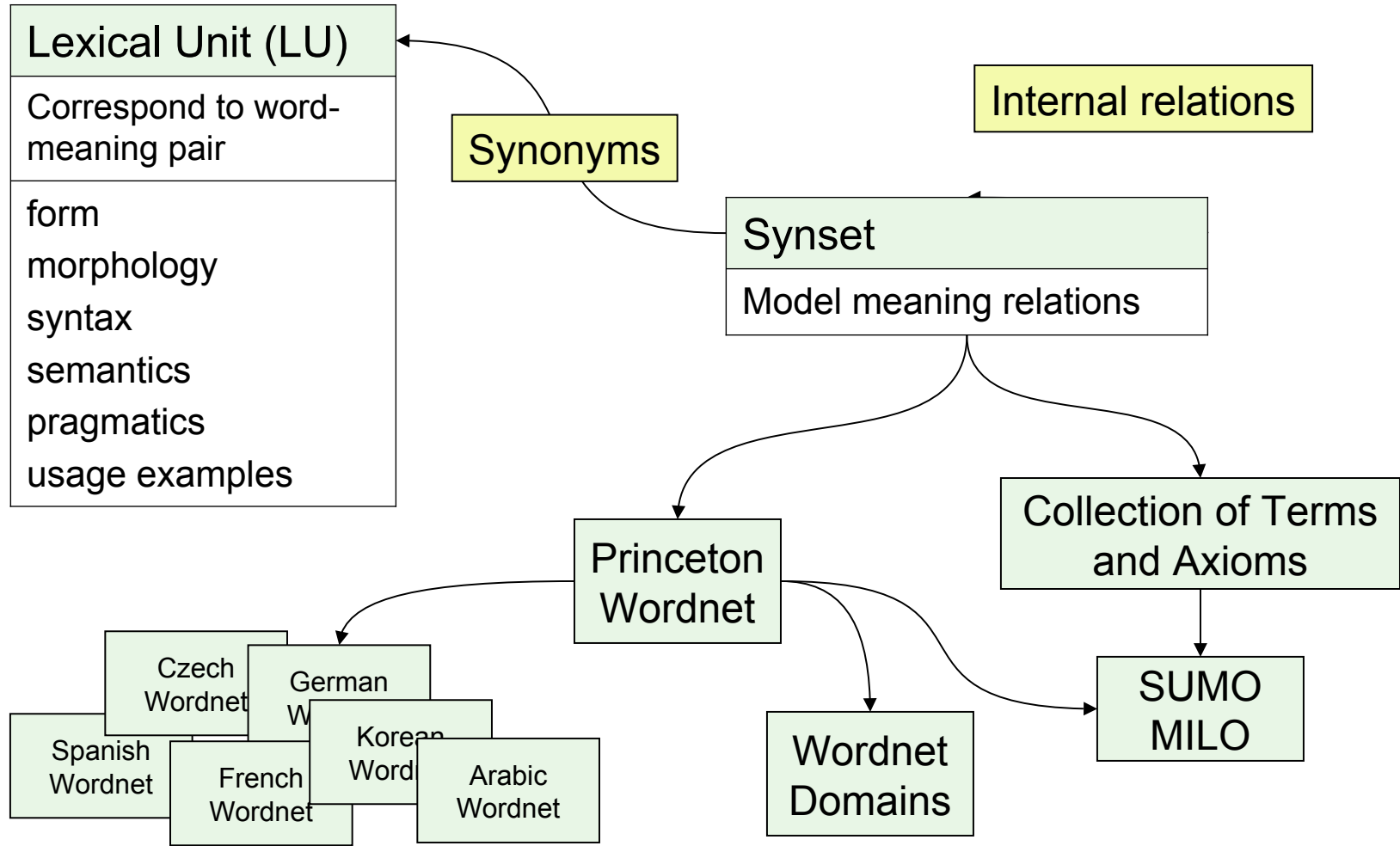
■ Collections:

- 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<-> SY
- SY <-> SY (within Dutch and from Dutch to English)
- SY <-> TE
- SY <-> 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

Cornetto Synset

theewater Search Search in all

d_n-20059 [noun] theewater:1 CUL water om thee te zetten

Concept Relations Tree Revtree Edit Xml

C_SY_ID: d_n-20059
Synonyms: theewater:1/r_n-37570

PoS specific: NOUN_HET
Definition: water om thee te zetten
Differentiae: om thee mee te zetten
VLIS domains: CUL
SUMO: (instance, 0, Water), (instance, 1, Tea), (instance, 2, Making), (resource, 0, 2), (result, 1, 2), (component, 0, 1)

-->> [HAS_HYPERONYM] majem:1, water:1/d_n-206

Equivalence relations:
[EQ_HAS_HYPERNYM] water:1, H2O:1/ENG20-14000

Querying a dictionary is complete.

Cornetto SUMO ontology

Water Search Search in all

Water

Preview Tree Revtree Xml

- Item
 - Physical
 - Object
 - SelfConnectedObject
 - Substance
 - PureSubstance
 - Compound
 - Water
 - Ice

binary is complete.

Cornetto SUMO o...

Tea Search Search in all

Tea

Preview Tree Revtree Xml

- Item
 - Physical
 - Object
 - SelfConnectedObject
 - Food
 - Apple
 - Avocado
 - Beverage
 - Tea
 - CerealGrain
 - Coconut
 - DateFruit
 - Fodder
 - Honey
 - LemonFruit
 - Meat
 - OrangeFruit
 - Pea
 - PearFruit
 - PreparedFood
 - Radish
 - RawFood

dictionary is complete. Found 16 item(s)

DEBVisDic English WordNet

water: Search Search in all

[n] body of water:1, water:2
[n] urine:1, piss:1, pee:1, piddle:1, weewee:1, water:5
[n] water system:2, water supply:1, water:3
[n] water:1, H2O:1
[n] water:4
[n] water:6
[v] water:1, irrigate:1
[v] water:2
[v] water:3
[v] water:4

Preview Tree Revtree Edit Query Xml

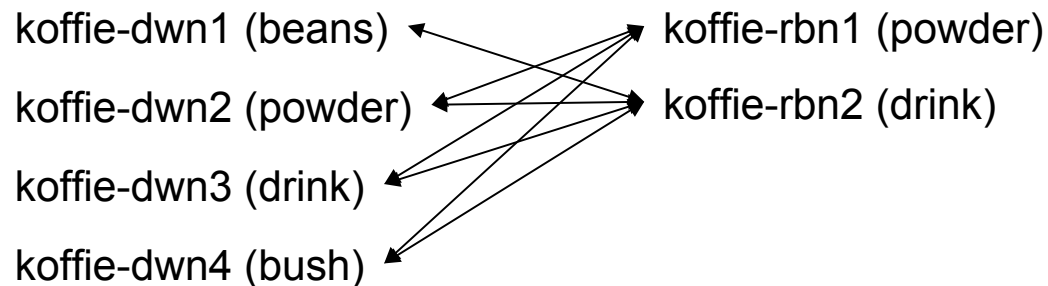
POS: n ID: ENG20-14000512-n BCS: 2
Synonyms: water:1, H2O:1
Definition: binary compound that occurs at room temperature as a clear colorless odorless tasteless liquid; freezes into ice below 0 degrees centigrade and boils above 100 degrees centigrade; widely used as a solvent
Domain: chemistry
SUMO/MILO: Water
-->> [hypernym] binary compound:1
-->> [hypernym] liquid:3
-->> [eng_derivative] water:1, irrigate:1
-->> [eng_derivative] water:2
-->> [holo_portion] tear:1, teardrop:2
-->> [holo_portion] perspiration:1, sweat:1, sudor:1
-->> [holo_portion] body of water:1, water:2
-->> [holo_portion] snowflake:1, flake:1
-->> [holo_portion] ice crystal:1, snow mist:1, diamond dust:1, poudrin:1, ice needle:1, frost snow:1, frost mist:1
-->> [holo_portion] dewdrop:1
-->> [holo_portion] ice:1, water ice:1
<<-- [hyponym] tap water:1
<<-- [hyponym] bath water:1

Querying a dictionary is complete. Item(s): 10

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.

Cornetto Synset

Search: hoed

Results:

- d_n-11620 [noun] hoed:1 KLED hoof
- d_n-17238 [noun] hoed:3 op een ho

Concept Relations Tree Revtree Edit Xml

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

Search: hoed

Results:

- r_n-16889 noun, hoed:1, hoofddeksel
- r_n-16890 noun, hoed:2, bovenste deel v....
- d_n-110164 noun, hoed:3

Form Examples Edit Xml

C_LU_ID: r_n-16890
C_SEQ_NR: 2
[noun] hoed

Morphology:
type: simpmorph
plurforms: hoeden

Syntax:
gender: m article: de

Cornetto ID

Search: hoed

Results:

- [*] hoed:1, NOUN, LU=r_n-16889, SY=d_n-11620; 79.95, -C
- hoed:2, NOUN, LU=r_n-16890, SY=; -1.0, -C
- hoed:3, NOUN, LU=d_n-110164, SY=d_n-17238; -1.0, -C

Preview Edit Xml

CID_ID: 19764 SEQ_NR: 1
Form: hoed
PoS: NOUN

Score: 79.95
Status:
LU ID: r_n-16889
Synset ID: d_n-11620

Results of the macro alignment

	LUS	LINKED		NOT-LINKED
RBN	66.024	47.250	72%	18.774
VLIS	106.504	46.924	44%	59.580
RBN-VLIS LINKS		58.053		

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 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)

KIF Expressions vs triplets

- Axioms in Sumo are written in SUO-KIF
- Cornetto: replaced by triplets, based on first order logic

SUMO

(and
(exists ?L ?W)
(instance, ?W, Water)
(instance, ?L, Liquid)
(Attribute, ?L, ?W))

Cornetto triplet

(instance, 0, Water)
(instance, 1, Liquid)
(Attribute, 1, 0)

Results

Quality labelling in the Cornetto database

- Cornetto database: **all data** in four subsets:
 - **Core Cornetto**: high quality, manually revised, central part of the Dutch language
(3.000 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**: 3+ meanings automatically aligned (approximately 6.000 words)
 - **Cornetto C**: 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, Universiteit van Tilburg.
- More information:

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

Overview of the results

Synsets	70571
Lexical Units	118467
Synonyms	104068
InternalRelations	132681
EquivalenceRelations	87413
Authors	171984
Definitions	35278
Mappings to WordnetDomains	94177
Mappings to SUMO	71048

Overview of the results

Cornetto Mappings Synset to LU	104258	
Manual	44978	43.14%
Automatic	59280	56.86%
DWN LU only	54987	52.74%
RBN LU only	3223	3.09%
DWN to RBN	35485	34.04%
Status field	44978	43.14%

Semantics for LUs

SEMANTICS			
semantics_adj	59245	sem_resume	40395
semantics_noun	45281	sem_selrestriction	59224
semantics_verb	84855	sem_selrestrictions	11125
semantics_example	15731	sem_shift	4117
sem_countability	5550	sem_spec_collocator	13063
sem_def	40395	sem_specificaes	13875
sem_defSource	18639	sem_subclass	36577
sem_def_noun	57352	sem_subtype_argument	7655
sem_definition	39032	sem_synonym	11359
sem_gc_compl	60171	sem_synonyms	17389
sem_gc_gramword	11567	sem_type	11786
sem_genus	11400	sem_caseframe	9545
sem_hyponym	39032	selrestriction	30698
sem_hyponyms	10530	selrestrole	8631
sem_lc_collocator	10530	caseframe	5550
sem_meaningdescription	25267	caserole	9545
sem_reference	22331		

Syntax for LUs

SYNTAX			
syntax_adj	16287	sy_gender	16248
syntax_noun	87319	sy_number	40395
syntax_verb	84855	sy_peraux	2982
syntax_example	15731	sy_position	10630
sy_advusage	17881	sy_reflexiv	8223
sy_article	8223	sy_separ	10630
sy_class	40395	sy_subject	10598
sy_combi	10630	sy_subtype	10630
sy_combicat	79740	sy_trans	45468
sy_combipair	99386	sy_type	10630
sy_combiword	99520	sy_valency	87319
sy_comp	99520	synset_list	10630
sy_compl_text	17868		
sy_complementation	2245		

Morphology for LUs

MORPHOLOGY

morphology_adj	59246	flex_conjugation	100733
morphology_noun	15731	flex_conjugationtype	10535
morphology_verb	84855	flex_mode	10535
mor_base	87319	flex_number	10630
mor_comparative	1709	flex_pastpart	10630
mor_comparis	4333	flex_pasttense	3518
mor_comparison	8147	flex_person	3518
mor_declinability	8147	flex_tense	10630
mor_flectional_type	8223		
mor_superlative	2059		
morpho_plurform	4333		
morpho_plurforms	30610		
morpho_structure	29003		
morpho_type	16691		

Other data for LUs

PRAGMATICS		EXAMPLES	
prag_chronology	17881	example	87319
prag_connotation	4238	canonicalform	87319
prag_domain	4886	textualform	39284
prag_frequency	25964	text_category	17881
prag_geography	4387	category	82447
prag_origin	4713		
prag_socGroup	6283		
prag_style	4093		
prag_subj_gen	6027		

Semantic relations for synsets

BE_IN_STATE	432	HAS_HOLO_LOCATION	278
CAUSES	1220	HAS_HOLO_MADEOF	145
CO_AGENT_INSTRUMENT	54	HAS_HOLO_MEMBER	233
CO_AGENT_PATIENT	40	HAS_HOLO_PART	1247
CO_AGENT_RESULT	46	HAS_HOLO_PORTION	84
CO_INSTRUMENT_AGENT	52	HAS_HOLONYM	276
CO_INSTRUMENT_PATIENT	282	HAS_HYPERONYM	67691
CO_INSTRUMENT_RESULT	88	HAS_HYPONYM	42743
CO_PATIENT_AGENT	39	HAS_MERO_LOCATION	279
CO_PATIENT_INSTRUMENT	282	HAS_MERO_MADEOF	148
CO_RESULT_AGENT	46	HAS_MERO_MEMBER	229
CO_RESULT_INSTRUMENT	86	HAS_MERO_PART	1240
CO_ROLE	24	HAS_MERO_PORTION	84
FUZZYNYM	6	HAS_MERONYM	275
XPOS_NEAR_SYNONYM	2013	HAS_SUBEVENT	409
XPOS_NEAR_ANTONYM	14	IS_SUBEVENT_OF	406

Semantic relations for synsets

HAS_XPOS_HYPERONYM	92	ROLE	116
HAS_XPOS_HYPONYM	89	ROLE_AGENT	647
IN_MANNER	19	ROLE_DIRECTION	9
INVOLVED	116	ROLE_INSTRUMENT	1544
INVOLVED_AGENT	641	ROLE_LOCATION	176
INVOLVED_DIRECTION	8	ROLE_PATIENT	968
INVOLVED_INSTRUMENT	1538	ROLE_RESULT	303
INVOLVED_LOCATION	173	ROLE_SOURCE_DIRECTION	13
INVOLVED_PATIENT	972	ROLE_TARGET_DIRECTION	30
INVOLVED_RESULT	301	STATE_OF	438
INVOLVED_SOURCE_DIRECTION	13		
INVOLVED_TARGET_DIRECTION	30		
IS_CAUSED_BY	1147		
MANNER_OF	18		
NEAR_ANTONYM	1592		
NEAR_SYNONYM	1044		

Equivalence relations to English

EQ_NEAR_SYNONYM	81671
EQ_HAS_HOLONYM	131
EQ_INVOLVED	46
EQ_HAS_HYPERNYM	571
EQ_BE_IN_STATE	35
EQ_IS_SUBEVENT_OF	6
EQ_HAS_HYPERONYM	1392
EQ_IS_STATE_OF	7
EQ_SYNONYM	3031
EQ_IS_CAUSED_BY	33
EQ_CAUSES	23
EQ_HAS_SUBEVENT	6
EQ_CO_ROLE	8
EQ_ROLE	53
EQ_HAS_MERONYM	60
EQ_HAS_HYPONYM	196

WordnetDomain relations

factotum	16518	linguistics	1154	literature	624
building_industry	2863	military	1149	biology	606
psychology	2722	sport	1021	geology	598
quality	2596	commerce	1008	time_period	583
gastronomy	2534	furniture	939	enterprise	579
transport	2257	money	922	agriculture	572
zoology	2102	physiology	921	metrology quality	563
economy	2049	administration	907	telecommunication	531
medicine	1886	physics	875	number	493
botany	1828	anthropology	848	town_planning	468
chemistry	1755	geography	846	exchange	466
anatomy	1726	merchant_navy	818	electricity	462
person	1564	pedagogy	752	pharmacy	461
law	1534	school	746	play	453
music	1482	mechanics	738	tourism	425
fashion	1414	publishing	711	history	374
industry	1348	university	673	theatre	356
politics	1278	art	651	computer_science	332
				medicine	