

Welcome to the Cornetto workshop



June, 7th, 2007

Vrije Universiteit Amsterdam

ru | STEVIN

Program

- 09:00-09:15 - Coffee and welcome
- 09:15-10:15 - Project overview (Piek Vossen)
- 10:15-10:45 - Alignment of resources (Piek Vossen)
- 10:45-11:15 - Database design (Katja Hofmann)
- 11:15-11:30 - Coffee break
- 11:30-12:30 - Strategies of work and progress (Hennie van der Vliet
Isa Maks, Roxane Segers)
- 12:30-13:00 - User-scenario's (Erik Tjong Kim Sang)
- 13:00-14:30 - Lunch
- 14:30–16:00 - Presentation by Adam Pease on SUMO, entitled:
“Formal Ontology”
- 16:00 - Drinks

Goal

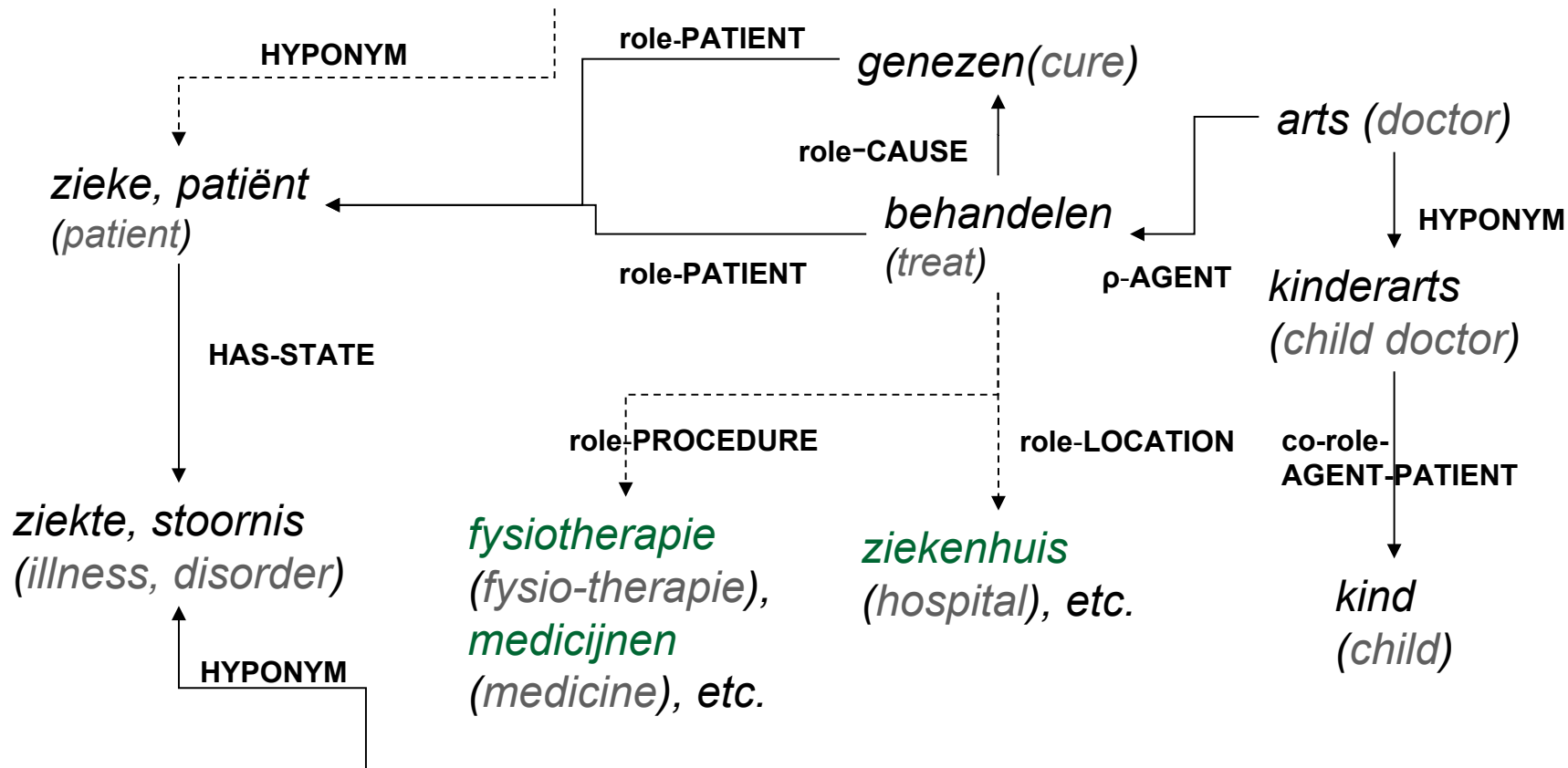
- <http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm>
- Stevin tender project (STE05039) to develop a lexical semantic database for Dutch:
 - 40K Entries
 - Generic and central part of the language
- Data:
 - Combination of WordNet and FrameNet
 - Vertical and horizontal semantic relations
 - Combinatorial lexical constraints
 - Aligned with the English Wordnet
 - Extended with an ontology
- Automatic acquisition toolkit
- Started April 2006, ends March 2008
- Licensed from TST-centrale, Nederlandse Taalunie

Consortium

- Vrije Universiteit Amsterdam, Faculteit der Letteren, Language, Cognition, and Communication:
 - Willy Martin, Isa Maks, Hennie vd Vliet, Roxane Segers, Piek Vossen
- Universiteit van Amsterdam, Instituut voor Informatica:
 - Maarten de Rijke, Erik Tjong Kim Sang, Katja Hofmann
- Katholieke Universiteit Leuven, Interdisciplinair Centrum voor Recht en Informatica (ICRI):
 - Sien Moens, Jan de Beer
- Irion Technologies BV:
 - Joop van Gent, Hetty van Zutphen, Piek Vossen

Horizontal & vertical semantic relations

chronisch zieke (chronical patient), *langdurig zieke* (long-term patient),
psychisch/geestelijk zieke (mental patient)



maagaandoening (stomach disorder)
nieraandoening (kidney disorder), *keelpijn* (sour throat).

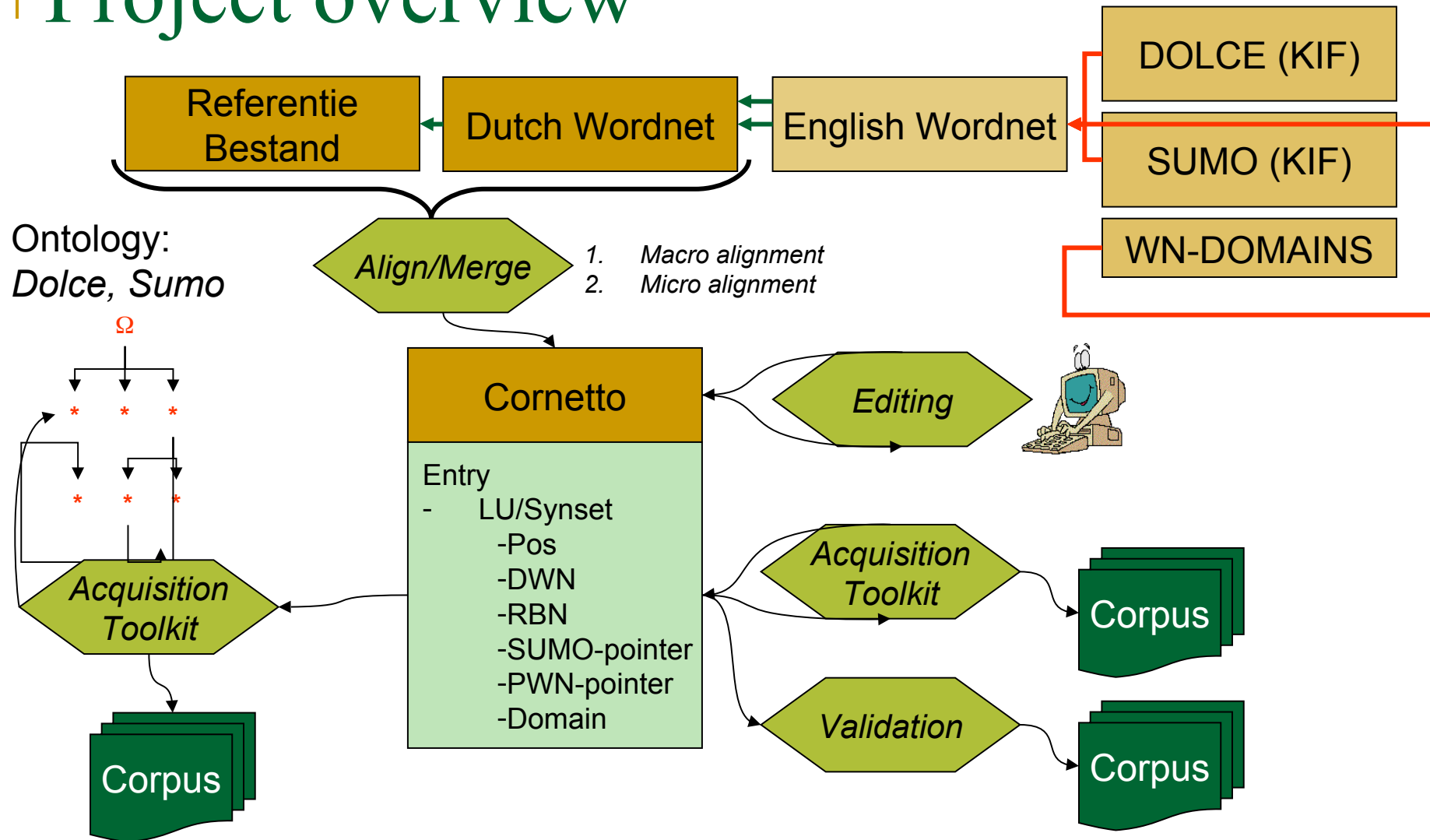
Combinatorics

slots	fillers (lex/conc)	fillers (coll)
action	<i>behandelen</i> (someone treat)	<i>iem. behandelen</i>
theme	<i>patiënt</i> (a patient treat)	<i>een patiënt behandelen</i>
state	<i>ziekte</i>	<i>iem. behandelen <u>voor</u> een ziekte</i> (someone treat for a disease)
	<i>behandelen</i>	<i>iem. <u>aan</u> zijn verwondingen</i> (somene at his injuries treat)
		<i>een ziekte behandelen (a disease treat)</i>

Approach

- Combine the information from two existing Dutch lexical resources:
 - The Dutch wordnet: synsets and lexical semantic relations
 - The Referentiebestand Nederlands: morpho-syntactic information, semantic information, pragmatic information, frame structures, lexical functions and combinatorics
- Macro level alignment
- Micro level alignment
- Populate with an ontology

Project overview



Global planning

- Two year project:
 - Month 1-6: design and database
 - Month 1-6: automatically aligned data
 - Month 7-10: ontology assignment
 - Month 7-22: editing
 - Month 7-15: acquisition
 - Month 16-17, 23-24: task-based evaluation

Database content overview

Two types of semantic resources

■ RBN:

- ❑ Lexical units
- ❑ Form-meaning relation
- ❑ Linguistic knowledge to use forms in a meaning
- ❑ Semantic-syntactic combinatoric frames

■ DWN:

- ❑ Based on WordNet (Fellbaum 1998): relational conceptual model of the lexicon
- ❑ Synsets: set of synonyms that represent a concept
- ❑ Semantic relations between synsets, i.e. abstracts from the specific form of particular synonyms

Lexical Unit & Synsets

- Lexical Unit = form-meaning relation, such that:
 - form = abstract representation of certain realizations;
 - part-of-speech is the same;
 - meaning is the same, where meaning is defined by a reference to a unique Synset;
- Synset = Set of synonyms (LUs) that refer to the same entities in most contexts.
 - Defined by lexical semantic relations;
 - Defined by reference to ontology Terms or KIF expressions involving Terms from the ontology;

Data structure overview

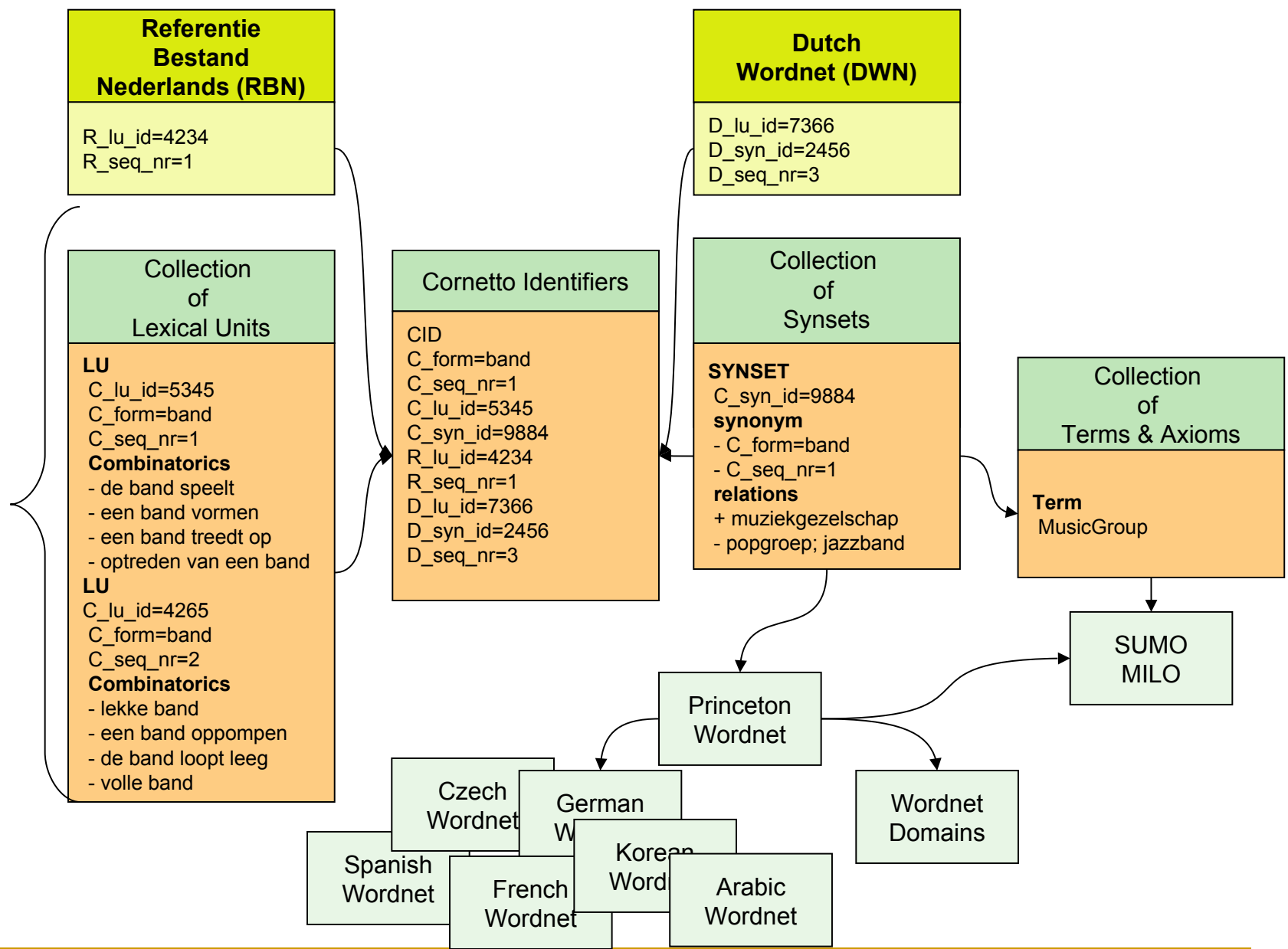
■ Collections:

- ❑ Lexical units (LU): -> mainly derived from RBN
- ❑ Synsets (SY): -> mainly derived from DWN
- ❑ Terms (TE): -> based on SUMO/MILO, linked to PWN
- ❑ Domains (DM): -> based on Wordnet domains

■ Mappings:

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

Cornetto Database (CDB)



Relation between form & meaning

Lexical Unit: form variants

- Inflectional variants, *appel, appels, appelen*
- Spelling variants
 - Meaning is identical
 - Pronunciation is mostly identical (*droppel, druppel*)
 - Spelling is different but the morphology is mostly the same: spelling variation can be systematic or incidental but it is in both cases conventional.
- Shortening:
 - Meaning is identical
 - Pragmatics is usually different
 - Pronunciation and spelling are different
 - Reduction in length for efficiency:
 - short forms (bus vs autobus)
 - abbreviation
 - contractions
 - acronyms

Lexical Unit: subtypes

- ❑ Different data structures for nouns, verbs & adjectives
- ❑ Single word unit:
 - simple morphological structure
 - complex morphological structure (compounds, derivations)
- ❑ Multiword unit (MWU, 85.000):
 - fixed (41 000)
 - free combinations (44 000): e.g. *morgen is er een vergadering over het bestemmingsplan van het industrieterrein* (tomorrow's meeting will be about the zoning plan for the industrial park))
- ❑ Excluded are **productive, compositional** derivations, compounds and multiword expressions => possibly included as examples take from RBN

Fixed expressions

- **Grammatical collocations** (10.000):
 - frequently co-occurring words;
 - noun, verb or adjective with a function word.(e.g. *met betrekking tot* (*with regard to*), *te voet* (*by foot*), *uit woede* (*out of anger*)).
- **Lexical collocations** with lexical function (15.500):
 - frequently co-occurring lexical words;
 - set of lexical functions (loosely based upon Melcuk).(e.g. *zware roker* (*heavy smoker*); *een vergadering schorsen* (*adjourn a meeting*))
- **Lexical collocations** – not specified (4.500):
 - cannot be expressed with a Melcuk function;
 - reduced semantic transparency
 - ontological type often differs from the type of the head of the phrase, i.e. it is subsumed by the head of the phrase.
(*systematische catalogus* (*systematic catalogue*), *open breuk* (*compound fracture*), *enkelvoudige breuk* (*simple fracture*))

Lexical Unit: polysemy

- Metonymy and metaphors:
 - lexical rules, encoded as shifts in RBN: school (institute) -> school (place)
- Complex Types (Pustejovsky):
 - Book incorporates different meaning layers: physical object & content
- Position in Cornetto is to maintain separate LUs, but to make the relations between these LUs explicit => shifts in RBN
- Motivation:
 - Semi-productive:
 - World-knowledge/culture-bound

walnut: refer to a wide range of objects and products that are related and derived from the walnut tree: fruit, wood, colour, taste, ice-cream, furniture, etc.

chestnut. similar object but less productive in our culture and therefore has less conventionalized meanings, even though anybody could productively create them at any time.

Relation between lexical meaning & ontology

Ontology versus lexicon

- Ontology is based on logical observation within a culture => relatively independent of language
- Terms represent disjoint Types, organized in a Type hierarchy where:
 - a Type represents a class of entities that share the same essential properties.
 - Instances of a Type belong to only a single Type: => disjoint (you cannot be both a *cat* & a *dog*)
- Terms can be combined in a knowledge representation language to form expressions of axioms (*you can be a watch dog & a bull dog*)

Ontology versus lexicon

- Ontology provides an independent anchoring of the relational meaning in DWN
- Ontology can be used to validate the lexical semantic structures, both for lexical units and synsets
- Ontology provides a mapping of vocabulary to formal representations for semantic web applications
- Lexical units map onto single synsets, but synsets can be related to the ontology in various ways

Ontology versus lexicon

- Synsets can be names for terms that are disjoint Types:
 - $\text{water}_{\text{NL}} = \text{Water}_{\text{SUMO}}$
 - $\text{zand}_{\text{NL}} = \text{Sand}_{\text{SUMO}}$
- Many Synsets are lexicalizations that can name instances of the same Type in different contexts:
 - water used for a purpose
 - water occurring somewhere or originating from
 - water being the result of a process
- The latter do not grant the introduction of new Types in the ontology

Comparing the DWN with SUMO

- Difficulties distinguishing the types:
 - $\{\text{water:1;majem=liquid appearance}\}_{\text{DWN}} \cong ?$
 $\text{Water}_{\text{SUMO}}$
 - $\{\text{water:2}\} = \text{WaterArea}_{\text{SUMO}}$
 - $\{\text{water:4=element}\} = \text{Water}_{\text{SUMO}} [\text{chemical element}]$

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

Preview Tree Revtree Edit Query Xml

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
 --> [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
 --> [eng_derivative] water:1, irrigate:1
 --> [eng_derivative] water:2
 <<-- [eng_derivative] water:1, irrigate:1
 <<-- [eng_derivative] water:2
 <<-- [derived] aqueous:1
 <<-- [derived] aquatic:1
 <<-- [derived] hydraulic:1
 <<-- [hyponym] tap water:1
 <<-- [mero_portion] hydrogen:1, H:1, atomic number 1:1
 <<-- [mero_portion] oxygen:1, O:2, atomic number 8:1
 <<-- [hyponym] bath water:1

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

Water used for a purpose

- ❑ theewater (*for making tea*)
 - (instance, 0, Water) , (instance, 1, Tea) , (component, 0, 1) (instance, 2, Making) , (component, 0, 2)
- ❑ koffiewater (*for making coffee*)
- ❑ bluswater (*for extinguishing fire*)
 - (instance, 0, Water) , (instance, 1, Fire) , (instance, 2, Process) , (instrument, 0, 2) , (finishes, 2, 1)
- ❑ scheerwater (*for shaving*)
- ❑ afwaswater (*for cleaning dishes*)
- ❑ waswater (*for washing*)
 - (instance, 0, Water) , (instance, 1, Washing) , (instrument, 0, 1)
- ❑ badwater (*for bading*)
- ❑ koelwater (*for cooling*)
- ❑ spoelwater (*for flushing*)
- ❑ drinkwater (*for drinking*)
 - (instance, 0, Water) , (instance, 1, Drinking) , (patient, 0, 1)



Water occurring somewhere or originating from

- ❑ putwater (*in a well*)
 - (instance, 0, Water) , (instance, 1, MineOrWell) , (origin, 0, 1)
- ❑ slootwater (*in a ditch*)
 - (instance, 0, Water) , (instance, 1, WaterArea) , (origin, 0, 1)
- ❑ oppervlaktewater (*at the surface*)
- ❑ welwater (*out of a spring*)
 - (instance, 0, Water) , (instance, 1, Creek) , (origin, 0, 1)
- ❑ leidingwater, gemeentepils, kraanwater (*out of the tap*)
 - (instance, 0, Water) , (instance, 1, Device) , (origin, 0, 1)
- ❑ gootwater (*in the kitchen sink or gutter*)
- ❑ grachtwater (*in a canal*)
 - (instance, 0, Water) , (instance, 1, Canal) , (origin, 0, 1)
- ❑ kwelwater (*coming from underneath a dike*)
 - (instance, 0, Water) , (instance, 1, StationaryArtifact) , (origin, 0, 1)
- ❑ grondwater, grondwater (*in the ground*)
- ❑ buiswater (*on a ship*)

Being the result of a process

- ❑ pompwater (*being pumped away*)
 - (instance, 0, Water) , (instance, 1, Pump) , (instance, 2, Transportation) , (instrument, 1, 2) , (patient, 0, 2)
- ❑ smeltwater, dooiwater (*melting snow and ice*)
 - (instance, 0, Water) , (instance, 1, Melting) , (result, 0, 1)
- ❑ afvalwater (*waste water*)
- ❑ condens, condensatiewater, condenswater (*from condensation*)
- ❑ lekwater (*leaking water*)
- ❑ regenwater (*rain water*)
 - (instance, 0, Water) , (instance, 1, Raining) , (result, 0, 1)
- ❑ spuiwater (*being drained for water maintenance*)



Combining resources

Semantics for frame structures

- Event structure for verbs from RBN:
 - ❑ E: behandelen <e0> action
 - ❑ A1: <?a1> pers
 - ❑ A2: <?a2>
 - ❑ C3: <?c3> prep
 - ❑ iemand aan [zijn verwondingen] behandelen
 - ❑ een patiënt voor [een nieraandoening/puistje/keelpijn] behandelen
 - ❑ iemand met [fysiotherapie/medicijnen]_{Instrument} behandelen
- DWN:
 - ❑ [causes] [v] genezen:2, beteren:1, herstellen:1
 - ❑ [involved_agent] [n] arts:1; dokter:1 <?a1>
 - ❑ [involved_patient] [n] zieke:1; patiënt:1 <?a2>
 - ❑ [involved_instrument] [n] hart-longmachine:1 <?c3>
 - ❑ [involved_instrument] [n] mitella:1, draagdoek:1 <?c3>
 - ❑ [involved_instrument] [n] geneesmiddel:1; medicijn:1 <?c3>
 - ❑ etc...

Advantage of combining RBN & DWN

- Missing information can be derived from the other resource;
- Argument structures & frames of RBN can be enriched with semantics => WSD;
- Add hierarchy to combinatorics;
- Synsets in DWN are enriched with contextual information: => WSD;
- Synsets in DWN are enriched with lexical & grammatical selection criteria: => text generation;

Identity criteria applied to DWN

- (Semi-)rigid type hierarchy in the ontology:
 - Canine => PoodleDog; NewfoundlandDog; DalmatianDog, etc.
- Wordnet consists of names for (semi-)rigid dog-types and other words for dogs with roles:
 - poedel = PoodleDog
 - jachthond (?CAN)
 - ⇔ (and
(instance ?CAN Canine)
(instance ?EV Hunting)
(agent ?CAN ?EV)))
- Type hierarchy remains compact and pure

Advantage of ontologizing Cornetto?

- More explicit definition of hierarchical structures & roles: => **better expansion**;
- Apply semantic implications through relations results in more exact database;
- Enables inferencing;
- Enables the development of domain specific database: => **required for semantic web applications**;

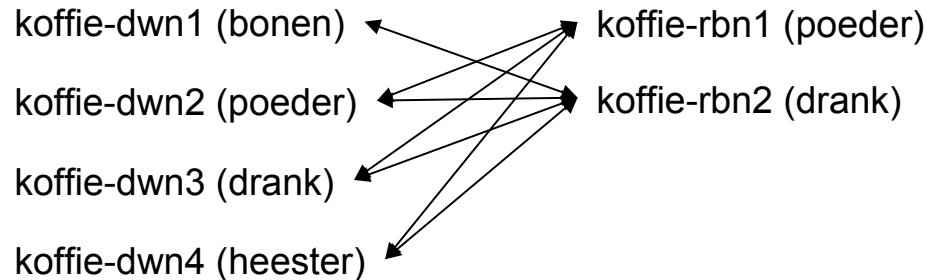
Alignment of lexical resources

Alignment

- Macro level alignment:
 - Lemma+pos
 - Word meanings
- Micro level alignment:
 - For each word meaning:
 - Co-index DWN and RBN information
 - Derive a new fused structure

Alignment

- Generate all weighted combinations:



- Produce merged output with mappings above probability threshold:

- New structure of word meanings

- koffie-cbn1(bonen) (source dwn1)
- koffie-cbn2 (poeder) (source dwn2, rbn1)
- koffie-cbn3 (drank) (source dwn3, rbn2)
- koffie-cbn4 (heester) (source dwn4)

Macro Alignment RBN - DWN

SYNSET	DEFINITION	DIFFERENTIAE	DOMAIN
baspartij_1 bas_1		die de bas zingt of speelt	#MUZ
bas_2 basstem_1	laagste mannenstem	laag, bij mannen	#MUZ
bas_3	baszanger, basspeler	met de basstem	#MUZ
contrabas_1 bas_4 basviool_1	het grootste en diepst gaande strijkinstrument	grootste en laagst klinkend	#MUZ

bas (noun) (bassen)

- 1 (count; nondynamic) [<gen-muz> zangstem] ⊢ <laagste> zangstem (BVD)
- 2 (count; human) [<gen-muz> zanger] ⊢ man met de stem van een bas (AA)
- 3 (count; artefact) [<gen-muz> contrabas] ⊢ strijkinstrument dat het grootst is en dat het laagste speelt (AA): *contrabas*
- 4 (count; artefact) [<gen-muz> basgitaar] ⊢ basgitaar (BVD-1)

6 Senses out of 8 candidates

- 1: lowest singing voice, RBN-1 & DWN-2
- 2: man with the voice of a bass, RBN-1 & DWN-3
- 3: biggest and lowest string instrument, RBN-3 & DWN-4
- 4: bass guitar, RBN-4
- 5: part of the music for the bass, DWN-1
- 6: bass singer or player, DWN-3

Strategies for the macro-alignment

	Conf.	Dev.	Factor	LINKS	
1: 1 RBN & 1 DWN meaning, no synonyms	97.1	4,9	3	9936	8,1%
2: 1 RBN & 1 DWN meaning	88.5	8,6	3	25366	20,8%
3: 1 RBN & >1 DWN meaning	53.9	8,1	1	22892	18,7%
4: >1 RBN & 1 DWN meaning	68.2	17,2	1	1357	1,1%
5: overlapping hyperonym word	85.3	23,3	2	7305	6,0%
6: overlapping hyponyms	74.6	22,1	2	21691	17,7%
7: overlapping domain-clusters	70.2	15,5	2	11008	9,0%
8: overlapping definition words	91.6	7,8	3	22664	18,5%

- 8 reviewers
- 100 random links per strategy
- nouns, verbs, adjectives, adverbs
- single confidence score per link based on all weighted strategies

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		

Cornetto Mapping Record

- CID unique pointer to bind them all, assigned by IRION
- C_LU_ID LU id to be assigned to each LU in CDB
- C_SY_ID SYNSET id to be assigned to each synset in CDB
- C_FORM lexical form
- C_SEQ_NR sequence number in CDB
- R_LU_ID LU id currently used in RBN
- R_SEQ_NR sequence number currently used in RBN
- D_LU_ID LU id currently used in DWN (original Vlis ID)
- D_SEQ_NR sequence number currently used in DWN
- D_SY_ID synset id currently used in DWN
- Score confidence score assigned by algorithm
- Status manually confirmed
- Name editor

Creation of Cornetto LUs and Synsets

- No mapping for a LU in RBN to a synonym in DWN:
 - create unique LU in Cornetto based on RBN LU. We do not create a synset for the LU in Cornetto;
- No mapping for a synonym in DWN to an LU in RBN:
 - create unique synonym in a unique synset in Cornetto
 - create corresponding Cornetto LU with the information from DWN;
- If there is a best scoring mapping between an LU in RBN and a synonym in DWN:
 - create single unique LU and a single unique synonym in Cornetto that point to each other and to both RBN and DWN;
- All remaining mappings:
 - do not create LUs and/or synsets;
 - stored as additional mappings (as weighted alternatives);

Different starting points RBN & DWN

- RBN tends to abstract from meaning specialization: => explained in examples, e.g. **helder** (*bright*) applied to liquids, light and colours.
- DWN tends to specialize meaning due to synonyms that represent more specific concepts: => explained by synsets, **helder** (bright) as a synonym of **onbewolkt** (without clouds) raises a different meaning.
- Cornetto will result in a new meaning structure based on both view points
- This will result in merging and splitting word meanings and synsets

Micro-alignment

- ❑ Separate layers with co-indexing:
 - DWN:
 - ❑ gitarist<0> -co_agent_instrument-> gitaar<1>
 - RBN:
 - ❑ gitarist <0> speelt **op** een gitaar <1>
- ❑ Unified CBN structure:
 - Event structure:
 - ❑ E: gitaarspelen <e0>
 - ❑ A1: gitarist <a1>
 - ❑ A2: gitaar <a2>
- ❑ Conceptual information shared by all synonyms
- ❑ Lexical information unique per synonym



The end....