

Integrating lexical units, synsets and ontology in the Cornetto Database

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



Project Cornetto

Financed by

NTU Dutch Language Union

STEVIN: Dutch Flemish Research Programme for Dutch Language and

Speech Technology (2004-2011)

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

- Goals of the project
- What's in the Cornetto database?
- Integrating the ontology: Sumo terms and new axioms

Goals of the Cornetto project

- **CO**mbinatorial **R**elational **NE**twork voor **T**aal **IO**epassingen
- 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
- Partners:
 - Vrije Universiteit Amsterdam
 - Universiteit Amsterdam
 - Universiteit Leuven
 - Irion Technologies



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

Two types of semantic resources

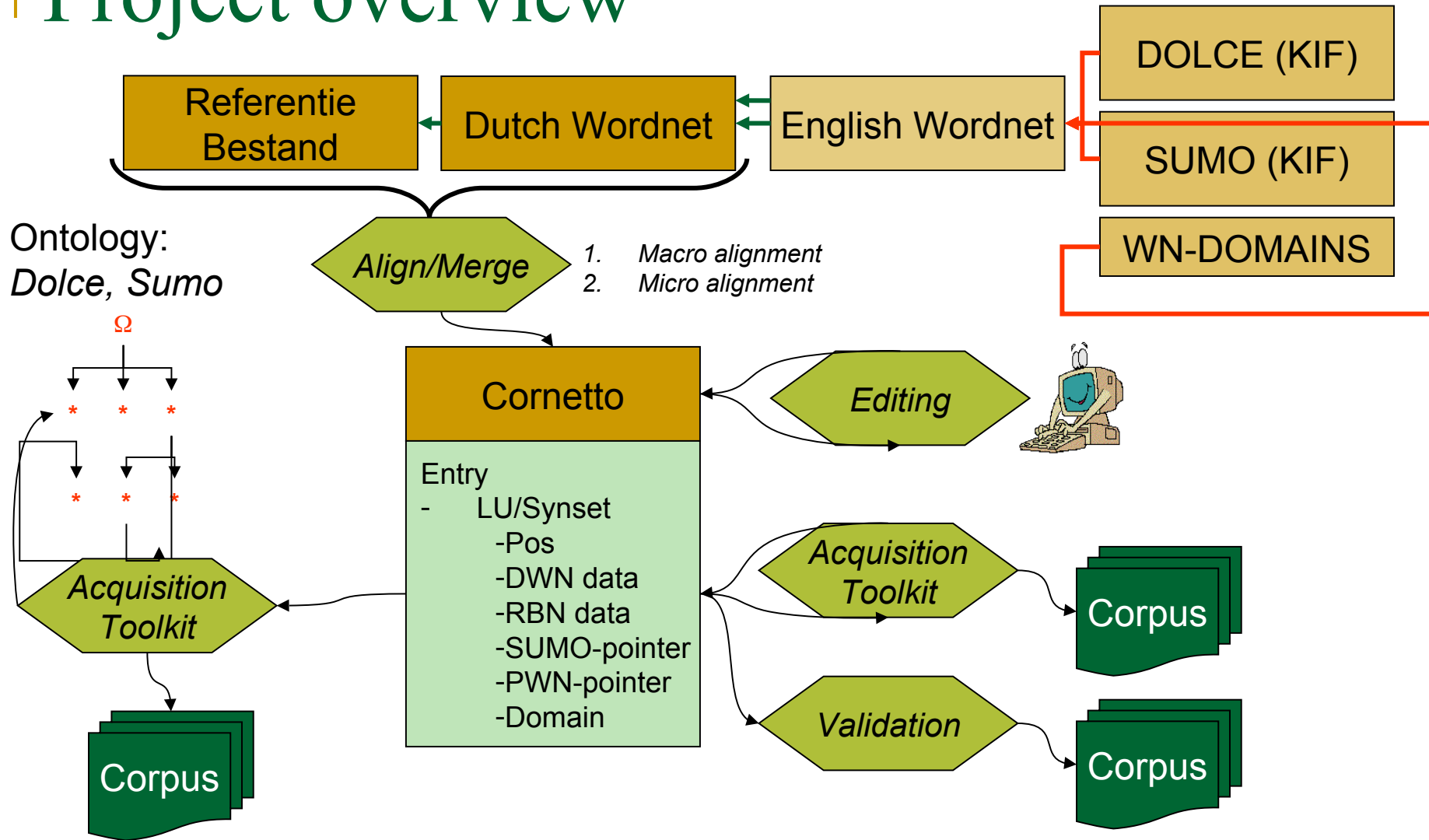
■ RBN:

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

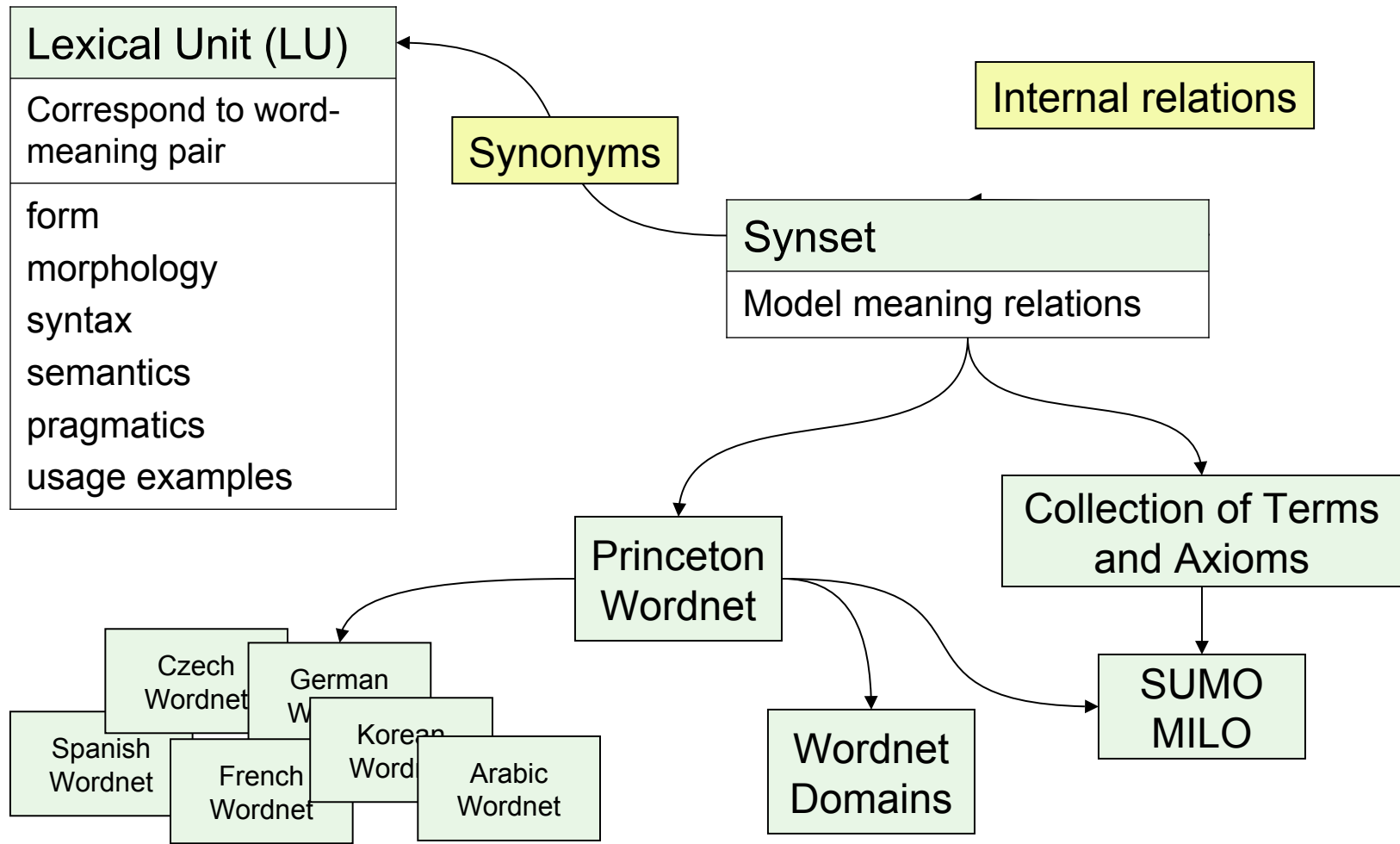
■ DWN:

- ❑ 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

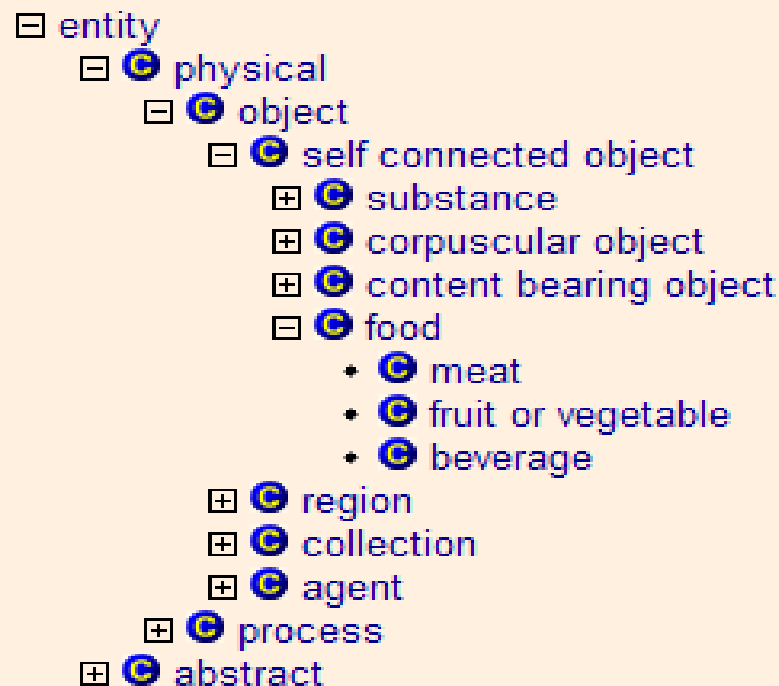
Project overview



Data Organization



Integrating the ontology: Sumo terms and new axioms



Rationale for an ontological layer

- Formal and fundamental model of meaning
- Detection of inconsistencies
- Formal reasoning
- Global semantic grid

SUMO/MILO as ontological framework

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

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)

Mapping to SUMO

- Subsumption, equivalence, instance

Tea (drink) (+,, Tea)

Tea (shrub) (+,, FloweringPlant)

Date (fruit) (=,, Datefruit)

Marrakech (instance,, City)

Ontology mapping: female/male variants

Teacher (a person whose occupation is teaching)

SUMO: equivalent to Teacher

In Dutch: no neutral form

leraar (male teacher)

(+,,Teacher), (+,, Man)

lerares (female teacher)

(+,,Teacher), (+,, Woman)

Synsets versus Ontology Types

- Many Synsets are lexicalizations that can name instances of the same Sumo Type in different contexts:
 - water used for a purpose (dishwater)
 - water occurring somewhere or originating from (tap water)
 - water being the result of a process (meltwater)
- The latter do not grant the introduction of new Types in the ontology

Complex ontology mapping

- theewater (*for making tea*)
 - (exists (?A ?W)
 (and
 (instance ?W Water)
 (hasPurposeForAgent ?W
 (exists (?T)
 (and
 (instance ?T Tea)
 (part ?W ?T))))))
- Simplified representation as list of triplets:
 - (instance, 0, Water) (instance, 1, Tea) (instance, 2, Making)
 (component, 0, 1) (resource, 0,2) (result,1, 2)

Some more triplets for water

regenwater (*rain water*)

- (instance, 0, Water) , (instance, 1, Raining) , (result, 0, 1)

smeltwater, dooiwater (*water from melting snow and ice*)

- (instance, 0, Water) , (instance, 1, Melting) , (instance, 2, Ice)
(patient, 2,1) , (result, 0,1)

kwelwater (*groundwater coming to the surface by the pressure of water, especially occurring close to a dike*)

- (instance, 0, GroundWater) , (instance, 1, StationaryArtifact (=Dike)) , (instance, 2, StreamWaterArea) (instance, 3, MotionUpward)

But what to do with...

- Grondwater (*groundwater*)

Sumo term: GroundWater ("Groundwater is the subclass of Water that is found in deposits in the earth.")

But is ground water a subclass of Water, or is it an instance of water with a certain place, usage or origin?

‘The groundwater got polluted.’

‘They used groundwater for crop irrigation’

SUMO and detection of inconsistencies

- Synsets with many equivalence relations get many different SUMO labels:
 - *godsdiensstofening* (religious service):
 - avail:1, help:3, service:13;military service:1, armed service:1, service:6;overhaul:1, inspection and repair:1, service:8;religious ceremony:1, religious ritual:1;rite:1, religious rite:1;serve:1, service:10;service:1;service:2;service:3, religious service:1, divine service:1;service:4;service:5;service:7;service:9, table service:1;service:11, serving:2, service of process:1;service:15;servicing:1, service:14;worship:1;worship:2, adoration:1;
 - Serving, Repairing, Giving, Impacting, SubjectiveAssessmentAttribute, IntentionalProcess, employs, SexualReproduction, ReligiousProcess, ReligiousService, Cooperation, Tableware, instrument, EmotionalState, ServiceOrganization, MilitaryService
 - The highest scores for translations are 3226.0 for service:3, religious service:1, divine service:1 and 2821.0 for worship:1

SUMO and detection of inconsistencies

■ Degree of consistency:

- Average distance in SUMO Hierarchy between multiple SUMO labels:
 - ReligiousAttribute:RelationalAttribute#Attribute#Abstract#Entity#
 - SoundAttribute:RelationalAttribute#Attribute#Abstract#Entity#
 - FormOfAdaptationAttribute:Attribute#Abstract#Entity#
 - *godsdiensstoefening* (religious service): average distance is 9,13
 - Cooperation:6.75.EmotionalState:11.06.Giving:7.5.Impacting:9.31.IntentionalProcess:5.5.MilitaryService:11.19.ReligiousProcess:7.ReligiousService:10.19.Repairing:6.12.ServiceOrganization:9.44.Serving:10.12.SexualReproduction:10.19.SubjectiveAssessmentAttribute:9.31.Tableware:9.94.employs:9.94.instrument:12.56.
 - dekkleed:14 (cover cloth): average distance is 0.5
 - Blanket:0.5.CottonFabric:0.5.Device:0.75.Fabric:0.25.0,5

Reducing equivalence mappings and average distance

Average distance	NOUN MEANINGS			VERB MEANINGS		
	ALL EQUI	TOP EQUI 80%	%	ALL EQUI	TOP EQUI 80%	%
0	1195	1278	106,9%	350	372	106,3%
1	1081	1136	105,1%	674	703	104,3%
2	1103	1112	100,8%	716	711	99,3%
3	1312	1280	97,6%	621	557	89,7%
4	2090	2037	97,5%	565	493	87,3%
5	1695	1533	90,4%	361	283	78,4%
6	930	743	79,9%	197	143	72,6%
7	375	232	61,9%	36	25	69,4%
8	107	56	52,3%	1		0,0%
9	14	8	57,1%			



The end....