



Stevin programmaday

11 September 2006

Antwerpen

ru | STEVIN

Consortium

- Vrije Universiteit Amsterdam, Faculteit der Letteren, General Linguistics Department, Onderzoeksgroep Lexicologie/Terminologie:
 - 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

Other partners

■ User-group:

- ❑ Polderland
- ❑ Knowledge Concepts
- ❑ LibRT
- ❑ Irion Technologies
- ❑ Van Dale Lexicografie
- ❑ Larcier-De Boeck

■ Ontology-group:

- ❑ Dr. W. Ceusters, Office Line Engineering nv
- ❑ Prof. F. van Harmelen, Vrije Universiteit Amsterdam
- ❑ Dr. P. Buitelaar, DFKI
- ❑ Dr. P. Monachesi, Universiteit van Utrecht

Overview

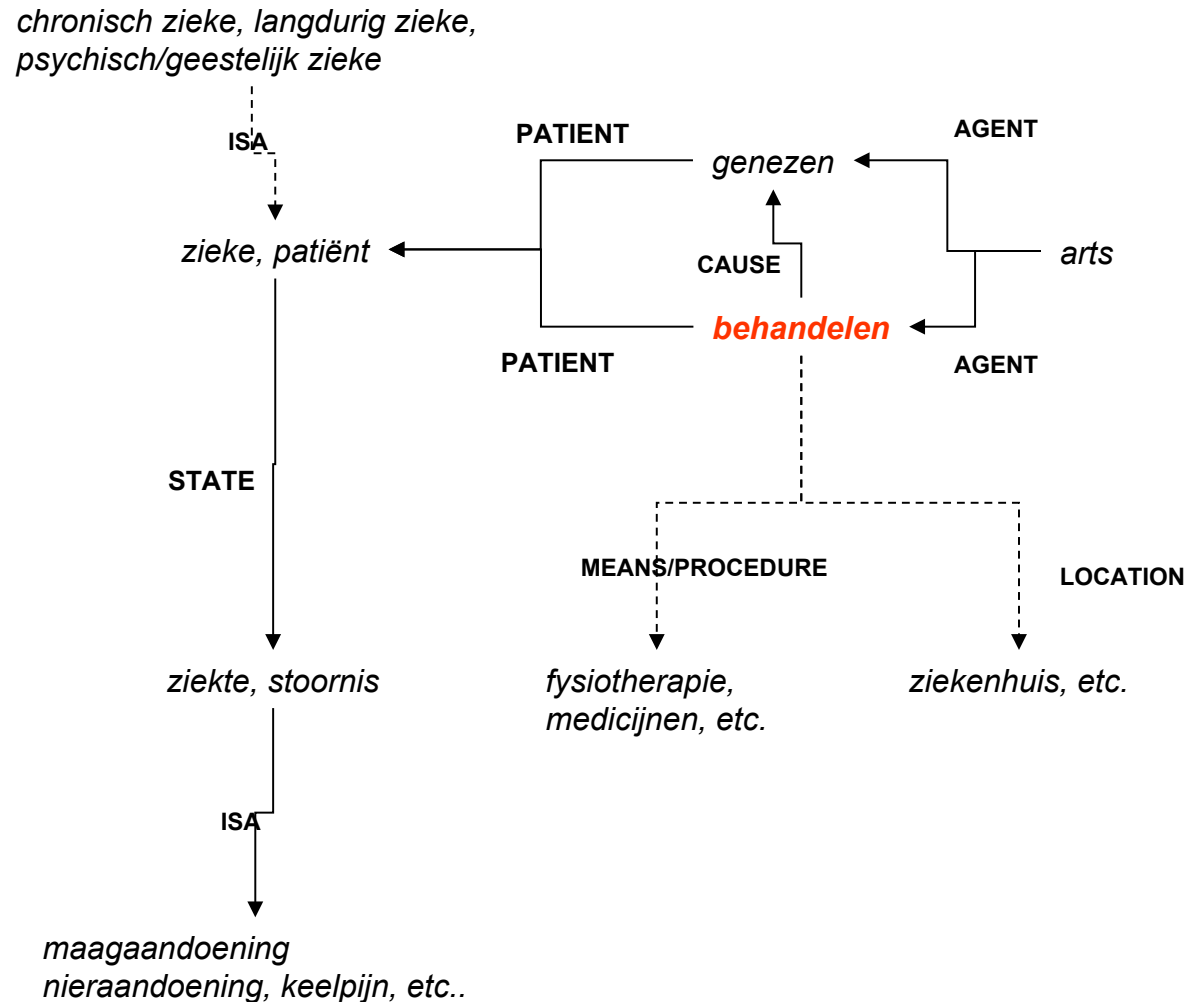
1. Project background information
2. Alignment of lexical resources
3. Database design
4. Next steps



Goal

- 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

Horizontal & vertical semantic relations



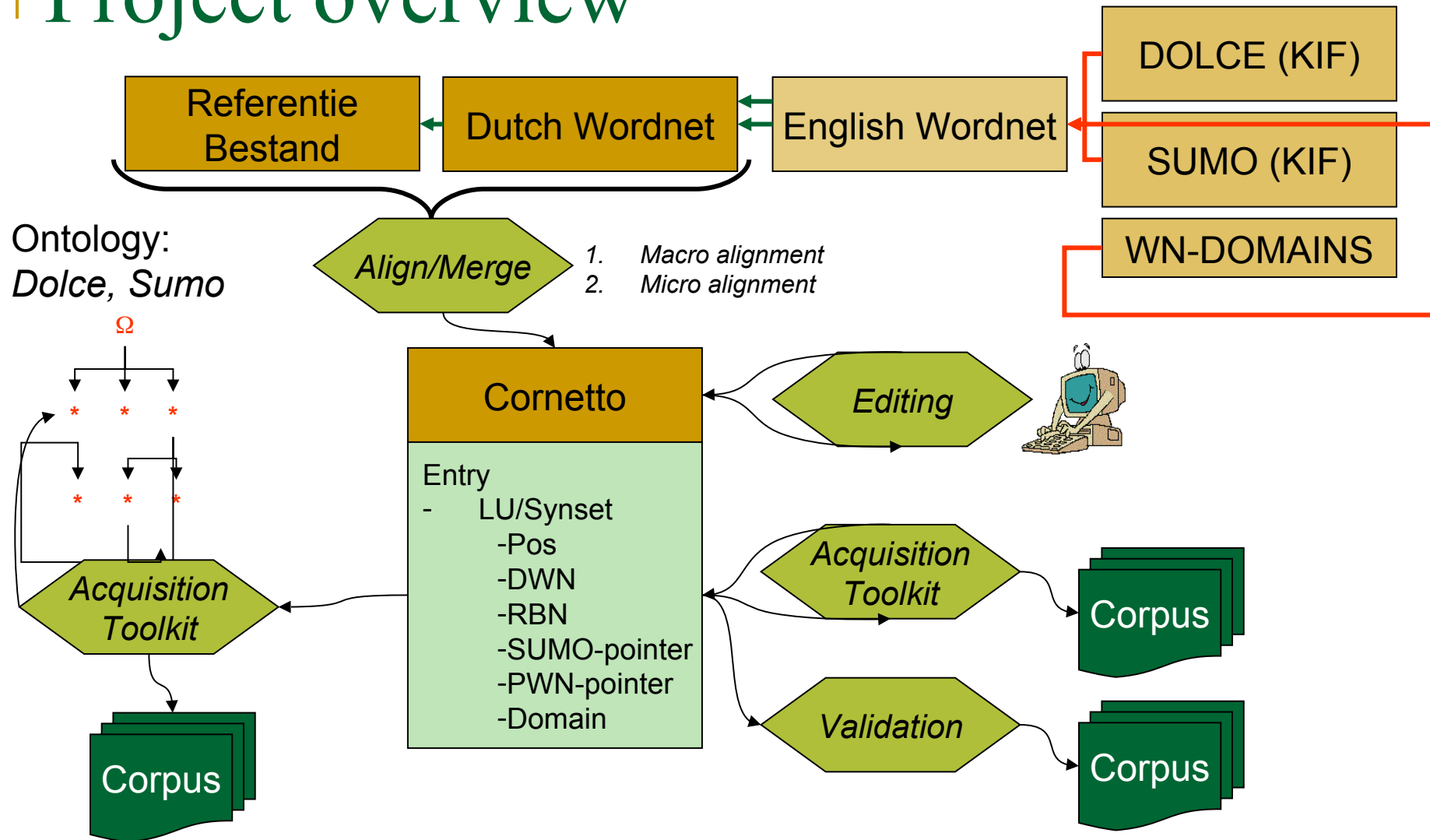
Combinatorics

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>treat)</i>	<i>een ziekte behandelen (a disease</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



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 the distinct Terms in the ontology or KIF expressions involving Terms from the ontology;
- 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;

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
- sms language

Lexical Unit: Meaning variants

- Roles, including male/female variants, e.g.:
 - theoloog/theologe
 - leraar/lerares
 - secretaresse vs. mannelijke secretaresse
 - kleuterleidster vs. ?
- Criteria for distinguishing different concepts:
 - If defined exhaustively as a role that is neutral with respect to male/femaleAND
 - If the male/female form can be derived with a regular and compositional derivationTHEN 1 LU for the abstract neutral form.
 - In all other cases: separate LUs, possibly related to different ontology terms or a KIF expressions, depending on the ontology.
- Consequences:
 - Single LU for *theoloog*:
 - neutral form (possibly zero-derivation) for *man*.
 - neutral form or apply a derivational rule to create *theologe* when applied to *woman*
 - Separate LUs for *leraar* (male teacher) and *lerares* (female teacher), there is no neutral form. The same applies to *verpleger* (male nurse) and *verpleegster* (female nurse).

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

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

Macro alignment approach

- Feature match across RBN & DWN
 - Shared features DWN-RBN, DWN-PWN
 - Lemma, POS, hyperonym, definitions, domain labels, synonyms, semantic features (+/-animate)
 - Dependent features:
 - Relations, e.g. instrument <-> theme's
 - Ontology <-> syntactic complements
 - Merge tables, domain labels across resources
- Implementation of heuristics
 - Benchmarking & Normalization:
 - Samples per heuristic
 - Multiple reviewers
 - Combined probability:
 - Overall score
 - Score per heuristic
 - 60.64.56.89.67.45.34.89

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

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

Data structure overview

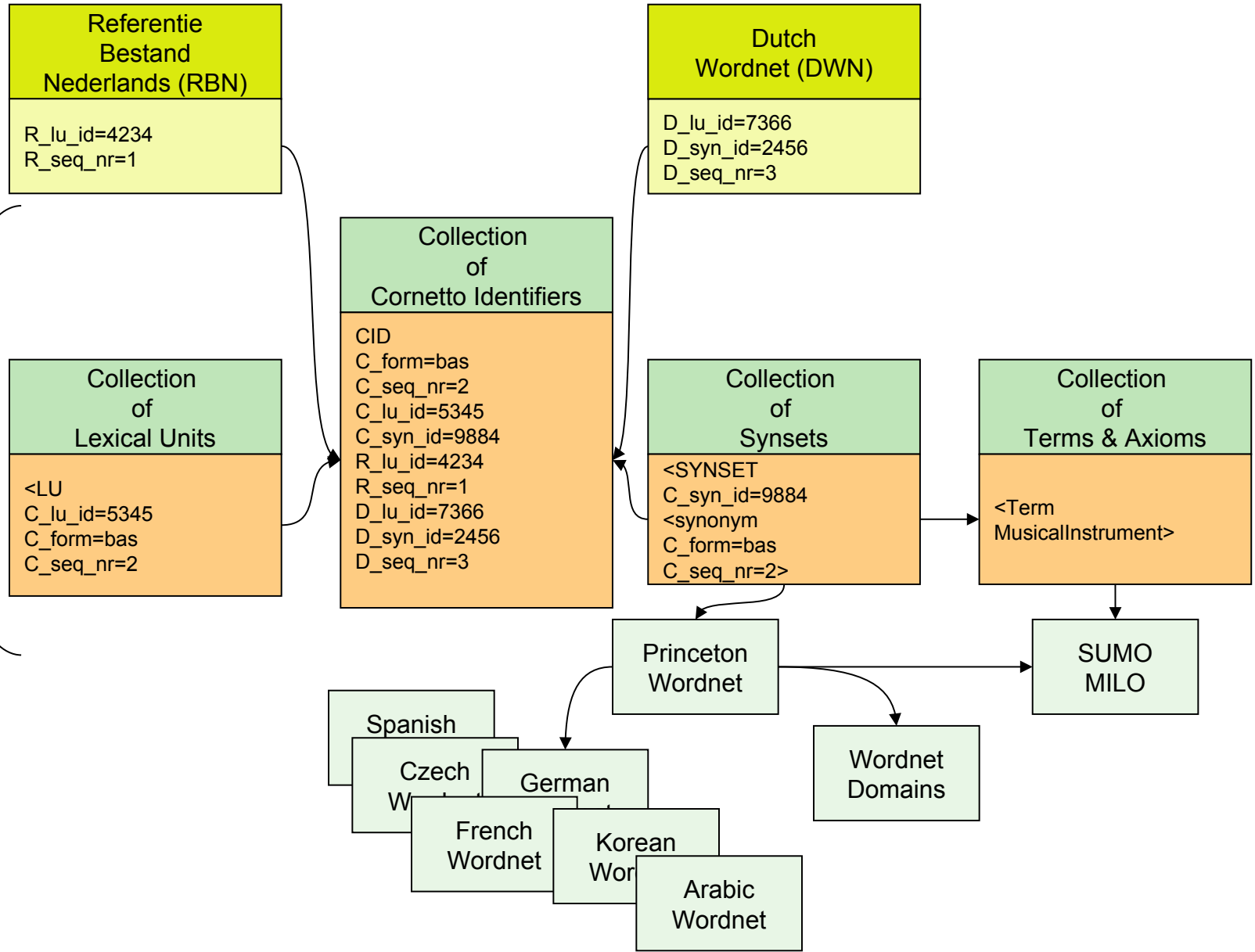
■ Collections:

- ❑ 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)



Cornetto Synset

koffie

Search

[noun] koffie:1
[noun] koffie:2
[noun] koffie:3, leut:2, slemp:1, slobber:3
[noun] koffieketel:1

Concept Relations Tree Revtree Edit Xml

Item

drank:2, drinken:2, noun
koffie:3, leut:2, slemp:1, slobber:3, noun

Querying a dictionary is complete. Found 2 item(s)

Cornetto ID

koffie

Search

koffie:1, NOUN, LU=d_139119, SY=d_17721; ...
koffie:2, NOUN, LU=d_139120, SY=d_23338; ...
koffie:3, NOUN, LU=d_139121, SY=d_28951; ...
koffie:4, NOUN, LU=d_139124, SY=d_34570; ...

Cornetto LU

koffie

Search

[noun] koffie:1; ,
[noun] koffie:2; ,
[noun] koffie:3; ,
[noun] koffie:4; ,
[noun] koffiebar:1; ,
[noun] koffiebes:1; ,
[noun] koffieblad:1; ,
[noun] koffieboom:1; ,
[noun] koffieboon:1; ,
[noun] koffiebranderij:1; ,

Form Syn-Sem Examples Edit Xml

C_LU_ID: d_139121 C_SEQ_NR: 3

Form:

spelling: koffie /

cat: noun

length: full

Morphology:

type:

structure:

plurform:

Querying a dictionary is complete.

Cornetto Synset

zetten

Search

[verb] inzetten:5, zetten:4
[verb] vastzetten:1, zetten:2
[verb] zetten:1, neerpoten:1, neerzetten:1, planten:2, poten:2
[verb] zetten:3

Concept Relations Tree Revtree Edit Xml

Item

veranderen:4, herscheppen:1, modifieren:1, wijzigen:1, verb
bewerken:1, verb
zetten:3, verb

Querying a dictionary is complete. Found 2 item(s).

DEBVisDic English WordNet

coffee

Search

[n] coffee urn:1
[n] coffee:1, java:2
[n] coffee:2, coffee tree:1
[n] coffeeberry:1, California buckthorn:1, California coffee:1, Rham...
[n] coffeecake:1, coffee cake:1

Preview Tree Revtree Edit Query Xml

POS: n STAMP: ID: ENG20-07452170-n BCS: 2

:1, java:2

verage consisting of an infusion of ground coffee beans
d a cup of coffee
tion
verage
beverage:1, drink:3, drinkable:1, potable:1
coffee substitute:1
Irish coffee:1
cafe au lait:1
cafe noir:1, demitasse:1
decaffeinated coffee:1, decaf:1
drip coffee:1
espresso:1
capuccino:1, capuccino coffee:1, coffee capuccino:1
iced coffee:1, ice coffee:1
mocha:3
Turkish coffee:1
coffee bean:1, coffee berry:1, coffee:3
cafe royale:1, coffee royal:1
caffeine:1, caffein:1

Querying a dictionary is complete.



Cornetto Synset

koffie

[noun] koffie:1
[noun] koffie:2
[noun] koffie:3, leut:2, slomp:1, slobber:3
[noun] koffieketel:1

Concept | Relations | Tree | Revtree | Edit | Xml

Synset Id ☐ Base concept

Synonyms: Cid Id, Lu Id, Status

Domain Relations

Semantic Relations: Target, Relation

Equivalence Relations: Target, Name, Version, Pos

Ontology

Querying a dictionary is complete. //

Cornetto LU

koffie

[noun] koffie:3; ,
[noun] koffie:4; ,

Form | Syn-Sem | Examples | Edit | Xml

Lu id Seq. nr. Type

Syntax

Gender Number
Article Complementation

Semantics

Definition: def, genus, specifica, defSource

Resume Def
Genus Specifica
DefSource

Reference Countability
Type Subclass
Shift Selfrestriction

Pragmatics

SocGroup Domain
Style Origin
Frequency Connotation

Querying a dictionary is complete. //

Current results & next steps

- Finalize macro alignment & database
- Finalize licenses
- Editing
 - Revising critical alignments
 - Defining ontology constraints
 - Revising word meanings based on ontology distinctions
 - Revising ontology assignment
 - Micro-level alignment
- Automatic acquisition
- Task-based evaluation



The end....

