WordNet, EuroWordNet and MCR

German Rigau i Claramunt

german.rigau@ehu.es

IXA group

Departamento de Lenguajes y Sistemas Informáticos

UPV/EHU

WordNet & EuroWordNet
Outline

- Introduction
- Wordnet
- EuroWordNet
- MCR
- WordNet extensions
- Building wordnets

About WordNet
Use WordNet online
Download
Frequently Asked Questions
Related projects
WordNet documentation
WordNet statistics
Publications
License and commercial use
Contact us, Report problems

About WordNet

WordNet® is a large lexical database of English, developed under the direction of George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

Over the years, many people have contributed to the development of WordNet. Currently, the WordNet team includes the following members of the Cognitive Science Laboratory:

- George A. Miller
- Christiane Fellbaum
- Randee Tengi
- Pamela Wakefield
- Helen Langone

WordNet has been supported by grants from the NSF, ARDA, DARPA, DTO, and REFLEX.

Click here for current research on WordNet being conducted at Princeton

WordNet News

New tools for researchers

We appreciate your comments and suggestions, especially when they are constructive and help us improve WordNet. Please contact us at [email].

Our staff examines all mail and tries to make appropriate changes, but we hope you understand that due to time constraints we cannot always respond to the sender.

Please note that changes made to the database are not reflected until a new version of WordNet is publicly released. Due to limited staffing, the date of the next release has not been scheduled. Check our website periodically for WordNet release information.

WordNet

- Princeton University (Miller et al. 1990, Fellbaum 98)
- Lexicalised concepts (words, compounds, multiwords)
- Synset: synonym set (of words)
- Large semantic net conecting synsets
  - synonymy, antonymy, hyperonymy, hyponymy
  - meronymy, implication, causation ...
- Structure
  - Noun hierarchy depth ~12
  - Verb hierarchy depth ~3
  - Adjective/adverb not in hierarchy, but in star structure
- Freely available: http://wordnet.princeton.edu/
- Extensively used in NLP
  - i.e. : disambiguated definitions for CLIR & WSD
- http://wordnet.princeton.edu

Word to search for: band Search WordNet
Display Options: (Select option to change) Change
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) set, circle, **band**, lot (an unofficial association of people or groups) *"the smart set goes there"; "they were an angry lot"*
- S: (n) **band** (instrumentalists not including string players)
  - *direct hyponym* / *full hyponym*
  - *part meronym*
    - S: (n) section (a division of an orchestra containing all instruments of the same class)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) musical organization, musical organisation, musical group (an organization of musicians who perform together)
      - S: (n) organization, organisation (a group of people who work together)
        - S: (n) social group (people sharing some social relation)
          - S: (n) group, grouping (any number of entities (members) considered as a unit)
            - S: (n) abstraction, abstract entity (a general concept formed by extracting common features from specific examples)
              - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
- S: (n) **band**, banding, stria, striation (a stripe or stripes of contrasting color) *"chromosomes exhibit characteristic bands"; "the black and yellow banding of bees and wasps"*
- S: (n) **band**, banding, stripe (an adornment consisting of a strip of a contrasting color or material)
- S: (n) dance band, **band**, dance orchestra (a group of musicians playing popular music for dancing)
- S: (n) **band** (a range of frequencies between two limits)
- S: (n) **band** (a thin flat strip of flexible material that is worn around the body or one of the limbs (especially to decorate the body))
- S: (n) isthmus, **band** (a cord-like tissue connecting two larger parts of an anatomical structure)
- S: (n) ring, **band** (jewelry consisting of a circlet of precious metal (often set with jewels) worn on the finger) *"she had rings on every finger"; "he noted that she wore a wedding band"*
- S: (n) **band** (a driving belt in machinery)
- S: (n) **band** (a thin flat strip or loop of flexible material that goes around or over something else, typically to hold it together or as a decoration)
- S: (n) **band**, ring (a strip of material attached to the leg of a bird to identify it (as in studies of bird migration))
- S: (n) **band** (a restraint put around something to hold it together)

Verb

- S: (v) **band** (bind or tie together, as with a band)
- S: (v) ring, **band** (attach a ring to the foot of, in order to identify) *"ring birds"; "band the geese to observe their migratory patterns"*

WordNet home page

WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: band Search WordNet
Display Options: (Select option to change) Change
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- {08240169} <noun.group>S: (n) set, circle, **band**, lot (an unofficial association of people or groups) "*the smart set goes there*"; "*they were an angry lot*"
- {08249038} <noun.group>S: (n) **band** (instrumentalists not including string players)
  - *direct hyponym* / *full hyponym*
  - *part meronym*
    - {08216900} <noun.group>S: (n) section (a division of an orchestra containing all instruments of the same class)
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - {08246613} <noun.group>S: (n) musical organization, musical organisation, musical group (an organization of musicians who perform together)
      - {08008335} <noun.group>S: (n) organization, organisation (a group of people who work together)
        - {07950920} <noun.group>S: (n) social group (people sharing some social relation)
          - {00031264} <noun.Tops>S: (n) group, grouping (any number of entities (members) considered as a unit)
            - {00002137} <noun.Tops>S: (n) abstraction, abstract entity (a general concept formed by extracting common features from specific examples)
              - {00001740} <noun.Tops>S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
- {04680465} <noun.attribute>S: (n) **band**, banding, stria, striation (a stripe or stripes of contrasting color) "*chromosomes exhibit characteristic bands*"; "*the black and yellow banding of bees and wasps*"
- {02784732} <noun.artifact>S: (n) **band**, banding, stripe (an adornment consisting of a strip of a contrasting color or material)
- {08249960} <noun.group>S: (n) dance band, **band**, dance orchestra (a group of musicians playing popular music for dancing)
- {06260518} <noun.communication>S: (n) **band** (a range of frequencies between two limits)
- {02784218} <noun.artifact>S: (n) **band** (a thin flat strip of flexible material that is worn around the body or one of the limbs (especially to decorate the body))
- {05267211} <noun.body>S: (n) isthmus, **band** (a cord-like tissue connecting two larger parts of an anatomical structure)
- {04092609} <noun.artifact>S: (n) ring, **band** (jewelry consisting of a circlet of precious metal (often set with jewels) worn on the finger) "*she had rings on every finger*"; "*he noted that she wore a wedding band*"
- {02785570} <noun.artifact>S: (n) **band** (a driving belt in machinery)
- {02785365} <noun.artifact>S: (n) **band** (a thin flat strip or loop of flexible material that goes around or over something else, typically to hold it together or as a decoration)
- {02785191} <noun.artifact>S: (n) **band**, ring (a strip of material attached to the leg of a bird to identify it (as in studies of bird migration))
- {02784998} <noun.artifact>S: (n) **band** (a restraint put around something to hold it together)

Verb

WordNet & EuroWordNet
WordNet Relations

- For **nouns** are:

@ Hypernym	~ Hyponym
@i Instance Hypernym	~i Instance Hyponym
#m Member holonym	%m Member meronym
#s Substance holonym	%s Substance meronym
#p Part holonym	%p Part meronym
! Antonym	
= Attribute	
+ Derivationally related form	
;c Domain of synset – TOPIC	-c Member of this domain – TOPIC
;r Domain of synset – REGION	-r Member of this domain – REGION
;u Domain of synset – USAGE	-u Member of this domain - USAGE

WordNet & EuroWordNet
WordNet Relations

- For **verbs** are:
  - @ Hypernym ~ Hyponym
  - * Entailment
  - > Cause
  - ! Antonym
  - ^ Also see
  - $ Verb Group
  - + Derivationally related form
  - ;c Domain of synset – TOPIC
  - ;r Domain of synset – REGION
  - ;u Domain of synset – USAGE

WordNet & EuroWordNet
WordNet Relations

- For **adjectives** are:
  - & Similar to
  - ! Antonym
  - = Attribute
  - ^ Also see
  - < Participle of verb
  - \ Pertainym (pertains to noun)
  - ;c Domain of synset - TOPIC
  - ;r Domain of synset - REGION
  - ;u Domain of synset - USAGE

WordNet & EuroWordNet
WordNet Relations

- For **adverbs** are:
  - ! Antonym
  - \ Derived from adjective
  - ;c Domain of synset - TOPIC
  - ;r Domain of synset - REGION
  - ;u Domain of synset - USAGE

WordNet Semantic Relations

- **Synonymy**
  - Between senses (or variants) in synsets
  - Weak notion of synonymy: from context
  - *Synset*: set of words expressing the same concept in a particular context (SemCor)
- **Hyperonymy / Hyponymy**
  - Class/subclass relation
    - {lion} -> {feline}

WordNet Semantic Relations

- **Meronymy relations**
  - Part /component
    - {hand}→{arm}
  - Element of a collectivity
    - {person}→{people}
  - Substance
    - {newspaper}→{paper}

WordNet Semantic Relations

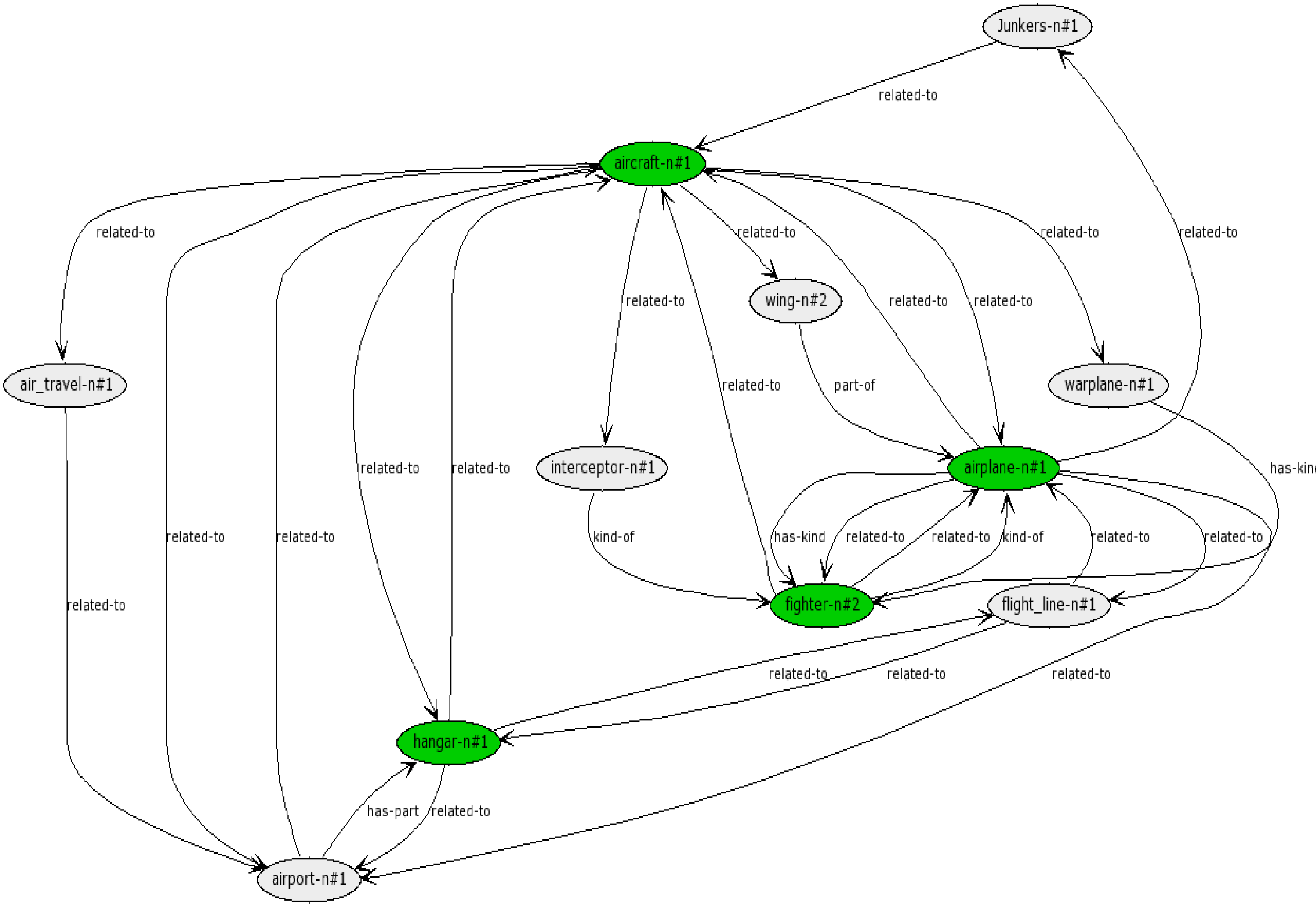
- **Antonymy**
  - {big}↔{small}
- **Causation**
  - {kill}→{die}
- **Implication**
  - {divorce}→{marry}
- **Derivation**
  - {presidential}→{president}
- **Similarity**
  - {good}↔{positive}

WordNet volumes

Version	nouns	verbs	adj	adv	total synsets	relations
WN1.5	51,253	8,847	13,460	3,145	76,705	103,445
WN1.6	66,025	12,127	17,915	3,575	99,642	138,741
WN1.7	74,488	12,754	18,523	3,612	109,377	151,546
XWN						551,551
WN1.7.1	75,804	13,214	18,576	3,629	111,223	153,781
WN2.0	79,689	13,508	18,563	3,664	115,424	204,074
WN2.1	81,426	13,650	18,877	3,644	117,597	232,916
WN3.0	82,115	13,767	18,156	3,621	117,659	235,402
WN3.0g						449,355

WordNet & EuroWordNet

WordNet graph



Welcome to:

Building a multilingual database with wordnets for several European languages.

EuroWordNet was a European resources and development project supported by the Human Language Technology sector of the Telematics Applications Programme (also see the description of all the funded projects).

Project Reference number:
LE-2 4003 & LE-4 8328
Application Area:
Language Resources, Language Engineering
Start Date:
March 1996
End Date:
June 1999

Abstract

EuroWordNet is a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The wordnets are structured in the same way as the American wordnet for English (Princeton WordNet, Miller et al 1990) in terms of synsets (sets of synonymous words) with basic semantic relations between them. Each wordnet represents a unique language-internal system of lexicalizations. In addition, the wordnets are linked to an Inter-Lingual-Index, based on the Princeton wordnet. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. The index also gives access to a shared top-ontology of 63 semantic distinctions. This top-ontology provides a common semantic framework for all the languages, while language specific properties are maintained in the individual wordnets. The database can be used, among others, for monolingual and cross-lingual information retrieval, which was demonstrated by the users in the project.

The EuroWordNet project was completed in the summer of 1999. The design of the database, the defined relations, the top-ontology and the Inter-Lingual-Index are now frozen. Nevertheless, many other institutes and research groups are developing similar wordnets in other languages (European and non-European) using the EuroWordNet specification. If compatible, these wordnets can be added to the above database and, via the index, connected to any other wordnet. The EuroWordNet format is defined by the EuroWordNet Database Editor Polaris. A specification can be found in the

EuroWordNet

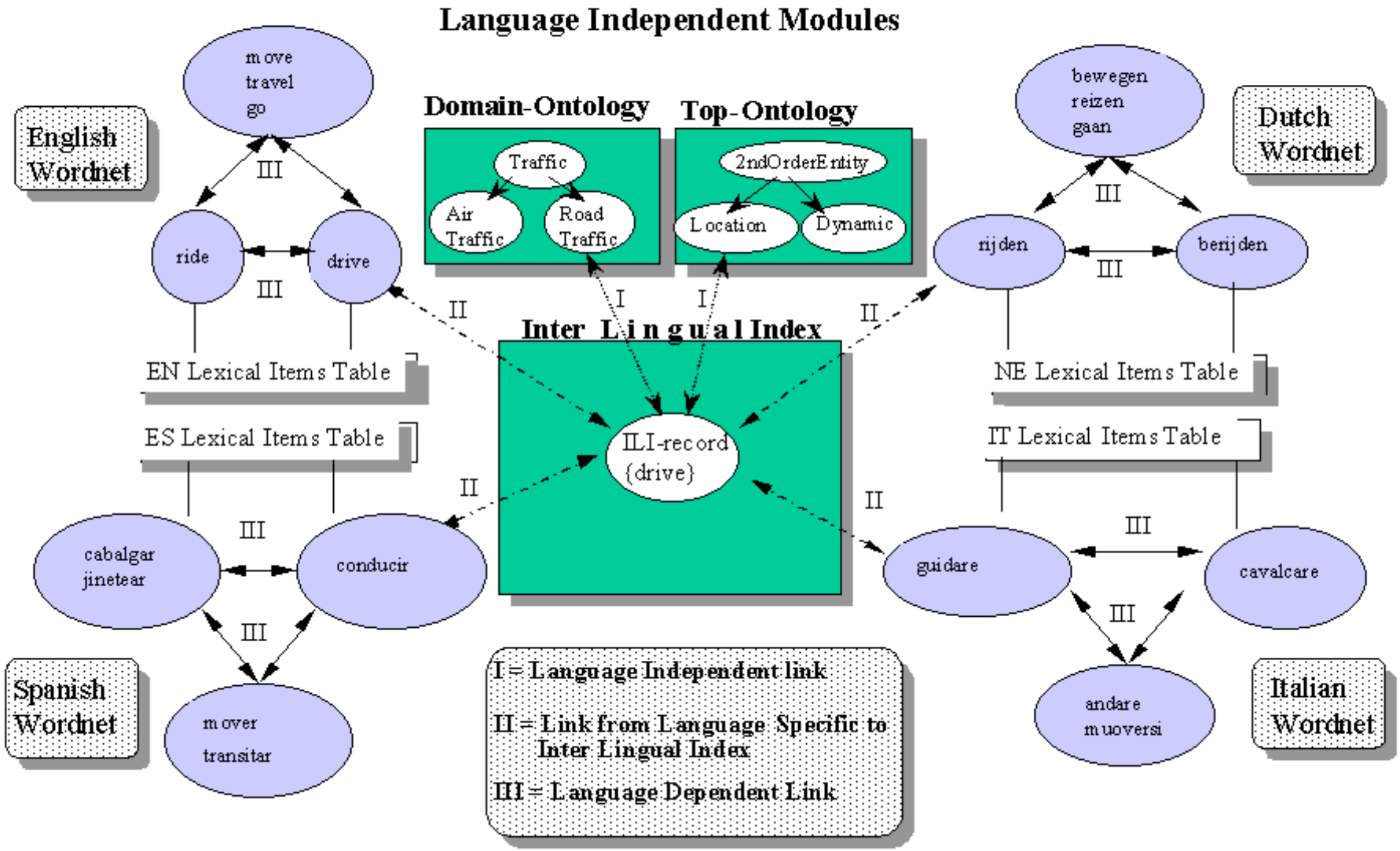
- Project LE-2 4003
  - Telematics Application Programme de la UE

- Integrated local wordnets in several languages
  - **English** Sheffield
  - **Dutch** Amsterdam
  - **Italian** Pisa
  - **Spanish** UB, UPC, UNED.

- Computers and the Humanities (Vossen 98)

- http://www.hum.uva.nl/~ewn/

Architecture of the EuroWordNet Data Structure



EuroWordNet Architecture

- Core
  - Inter-Lingual-Index (ILI)
  - Top Concept Ontology (TCO)
  - Domain Ontology (DO)
- Extensions
  - Local wordnets
  - Domain wordnets

Interlingual Index of EuroWordNet

- Set of synsets from WN 1.5
- Base concepts connected to TCO and DO

Top Concept Ontology of EuroWordNet

- Hierarchy of language independent concepts
  - Semantic distinctions: object, place, …
  - abstract (not lexical)
  - Connected to the ILI
- Three types of concepts:
  - First order: *entities*
  - Second order: *estatic or dinamic situations*
  - Third order: *abstract prepositions*

WordNet & EuroWordNet

Top Concept Ontology of EuroWordNet

Top[0]	
1stOrderEntity[1]	**2ndOrderEntity**[0]
Origin[0] Natural[21] Living[30] Plant[18] Human[106] Creature[2] Animal[123] Artifact[144] **Form**[0] Substance[32] Solid[63] Liquid[13] Gas[1] Object1[62] **Composition**[0] Part[86] Group[63] **Function**[55] Vehicle[8]	**SituationType**[6] Dynamic[134] BoundedEvent[183] UnboundedEvent[48] Static[28] Property[61] Relation[38] **SituationComponent**[0] Cause[67] Agentive[170] Phenomenal[17] Stimulating[25] Communication[50] Condition[62] Existence[27] Experience[43] Location[76] Manner[21] Mental[90]

WordNet & EuroWordNet

Domain Ontology of EuroWordNet

- Hierarchy of domains
  - Traffic:Road Traffic, Air traffic, etc.
  - Medicine
  - ...
- Domains label different parts of the hierarchies:
  - Medicine: doctor, nurse, operation, etc.
- Domains label different POS categories:
  - Medicine: doctor, to operate, etc.

EuroWordNet extensions

- EWN2
  - German (GermaNet), French, Chec, Swedish, Estonian
- ITEM, HERMES, KNOW, KNOW2, SKaTER
  - Spanish, Catalan, Basque and Galician
- CREL
  - Catalan (UB, UPC)
- EuroTerm, Jur-Wordnet
  - Extending EWN in particular domain
- Balkanet
  - Extending EWN for the Balkan languages
- Meaning
  - Enriching EWN automatically: MCR
- Kyoto
  - Web2.0 to enrich EWN for a domain

WordNet & EuroWordNet
MEANING

- Project IST-2001-34460

- Automatic extension of wordnets
  - **Spanish, Catalan** UPC
  - **Italian** Trento
  - **Basque** UPV/EHU
  - **English** Sussex

- http://www.lsi.upc.edu/~nlp/meaning

meaning

MEANING IST-2001-34460
Developing Multilingual Web-scale Language Technologies

New 2nd WorkShop February 3-4 2005

1st WorkShop Papers (April 2003)

Home

Participants

Background References

Contact Information

Links

Intranet (members only)

Documentation

Downloads

Demos

Project Summary

MEANING will be concerned with automatically collecting and analysing language data from the WWW on a large scale, and building more comprehensive multilingual lexical knowledge bases to support improved word sense disambiguation (WSD).

Current web access applications are based on words; MEANING will open the way for access to the Multilingual Web based on concepts, providing applications with capabilities that significantly exceed those currently available. MEANING will facilitate development of concept-based open domain Internet applications (such as Question/Answering, Cross Lingual Information Retrieval, Summarisation, Text Categorisation, Event Tracking, Information Extraction, Machine Translation, etc.). Furthermore, MEANING will supply a common conceptual structure to Internet documents, thus facilitating knowledge management of web content.

Progress is being made in Human Language Technology (HLT) but there is still a long way towards Natural Language Understanding (NLU). An important step towards this goal is the development of technologies and resources that deal with concepts rather than words. MEANING will develop concept-based technologies and resources through large-scale knowledge processing over the web, robust and fast machine learning algorithms, very large lexical resources and novel strategies for combining them. Small-scale, isolated experiments with limited infrastructure (such as Internet access, processing power, and storage space) have no chance of bridging the gap to understanding. Advances in this area can only be expected in the context of large-scale long-term research projects.

MEANING will treat the web as a (huge) corpus to learn information from, since even the largest conventional corpora available (e.g. the Reuters corpus, the British National Corpus) are not large enough to be able to acquire reliable information in sufficient detail about language behaviour. Moreover, most European languages do not have large or diverse enough corpora available.

MEANING

- From NLP to NLU
- Large-scale Semantic Processing dealing with concepts (senses) rather than words
- Two complementary OPEN problems:

- Acquisition bottleneck
  - Autonomous large-scale knowledge acquisition systems
- Ambiguity bottleneck
  - Highly accurate WSD systems

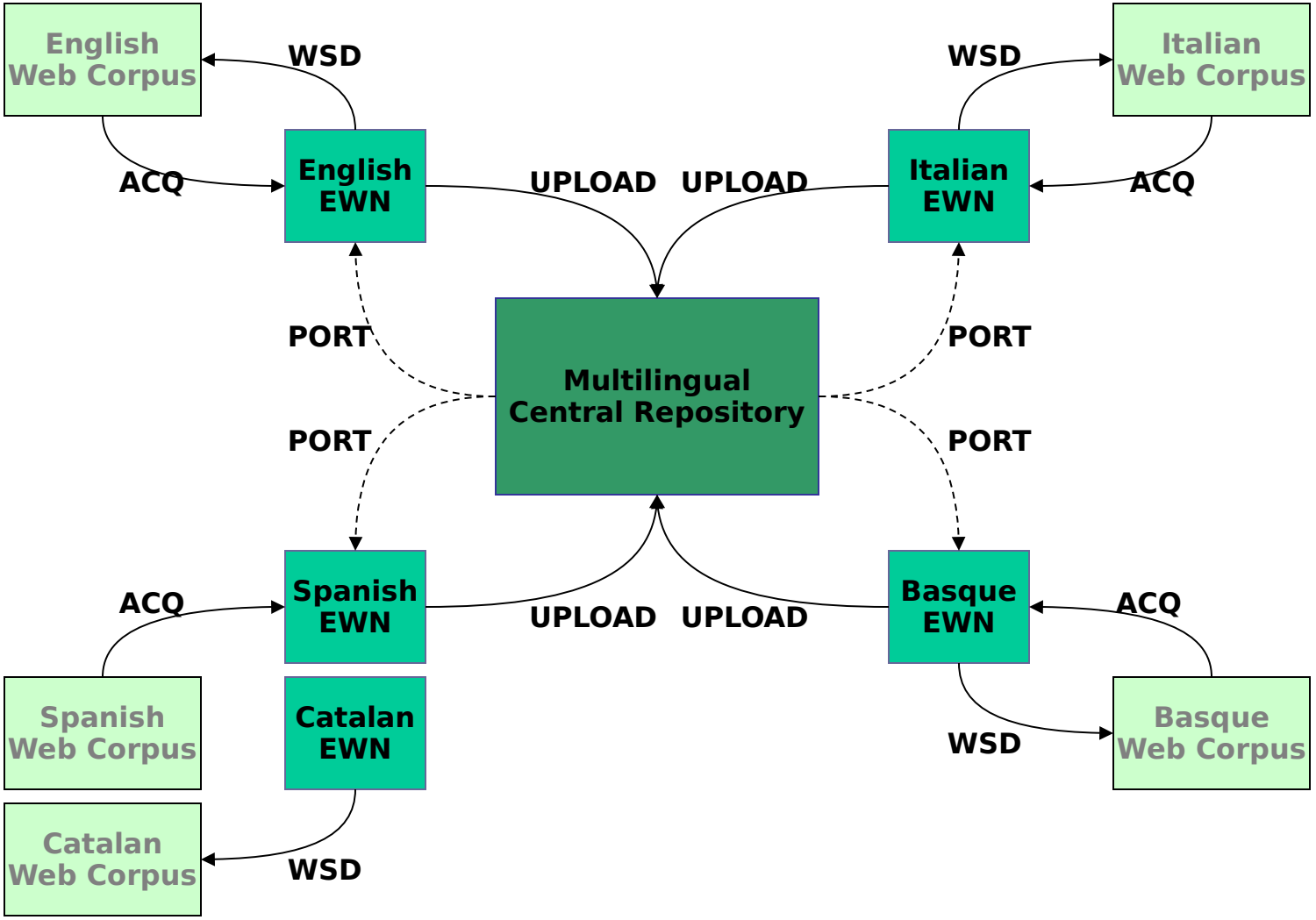
MEANING

Dealing with the ACQ/WSD deadlock

- Addressing Acquisition and WSD simultaneously
  - three consecutive MEANING cycles
- Language is highly polysemous
  - but also highly redundant
- Multilingualism
  - maybe is part of the solution using EuroWordNet
- Reuse of incompatible large-scale resources
  - Mapping technology to connect already available data
  - Cross-checking capabilities to detect inconsistencies

WordNet & EuroWordNet

MEANING



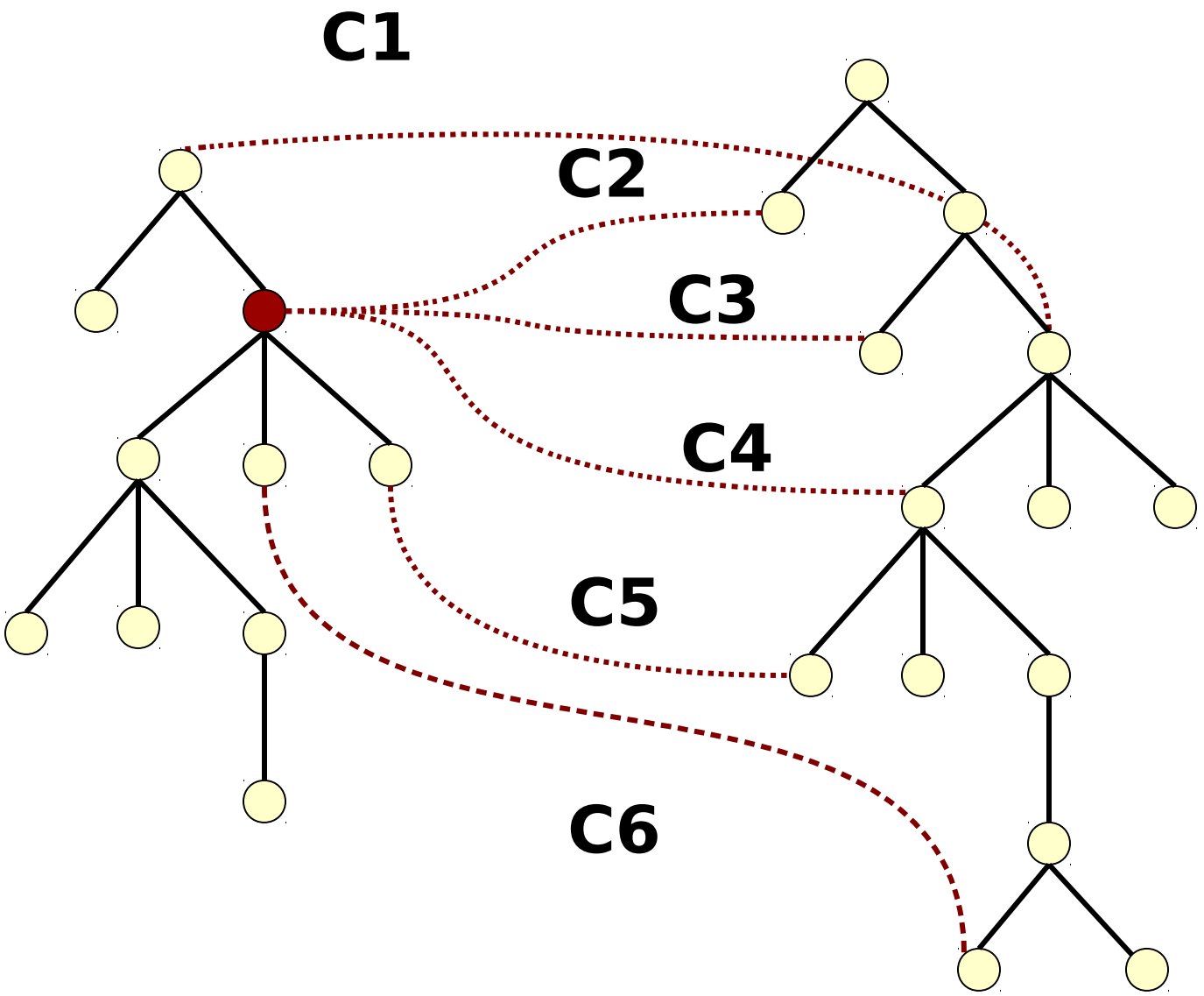
WordNet & EuroWordNet

MCR Content

- ILI
  - WordNet1.6
  - EuroWordNet Base Concepts => WN1.5
  - EuroWordNet Top Ontology => WN1.5
  - Multiwordnet Domains => WN1.6
  - SUMO => WN1.6
- Local wordnets
  - Wordnets of five European Languages
    - Basque, Catalan, English, Italian, Spanish
    - Five WordNet versions (1.5, 1.6, 1.7, 1.7.1, 2.0, 2.1, 3.0)
  - eXtended WordNet => WN1.7
- Large collections of Semantic Preferences
  - Acquired from SemCor => WN1.6
- Instances
  - Named Instances => WN1.6

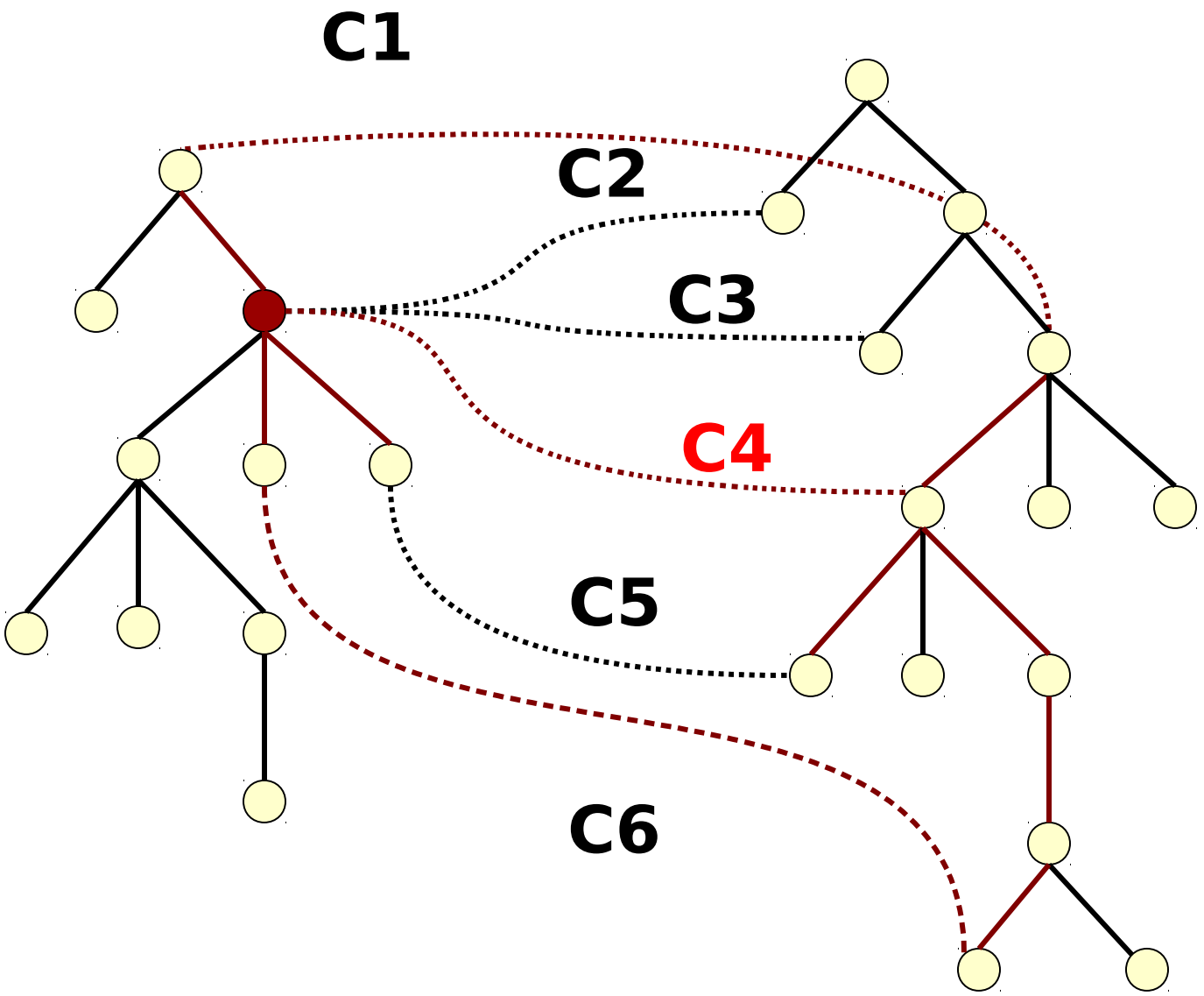
WordNet & EuroWordNet

Mapping technology



WordNet & EuroWordNet

Mapping technology



WordNet & EuroWordNet

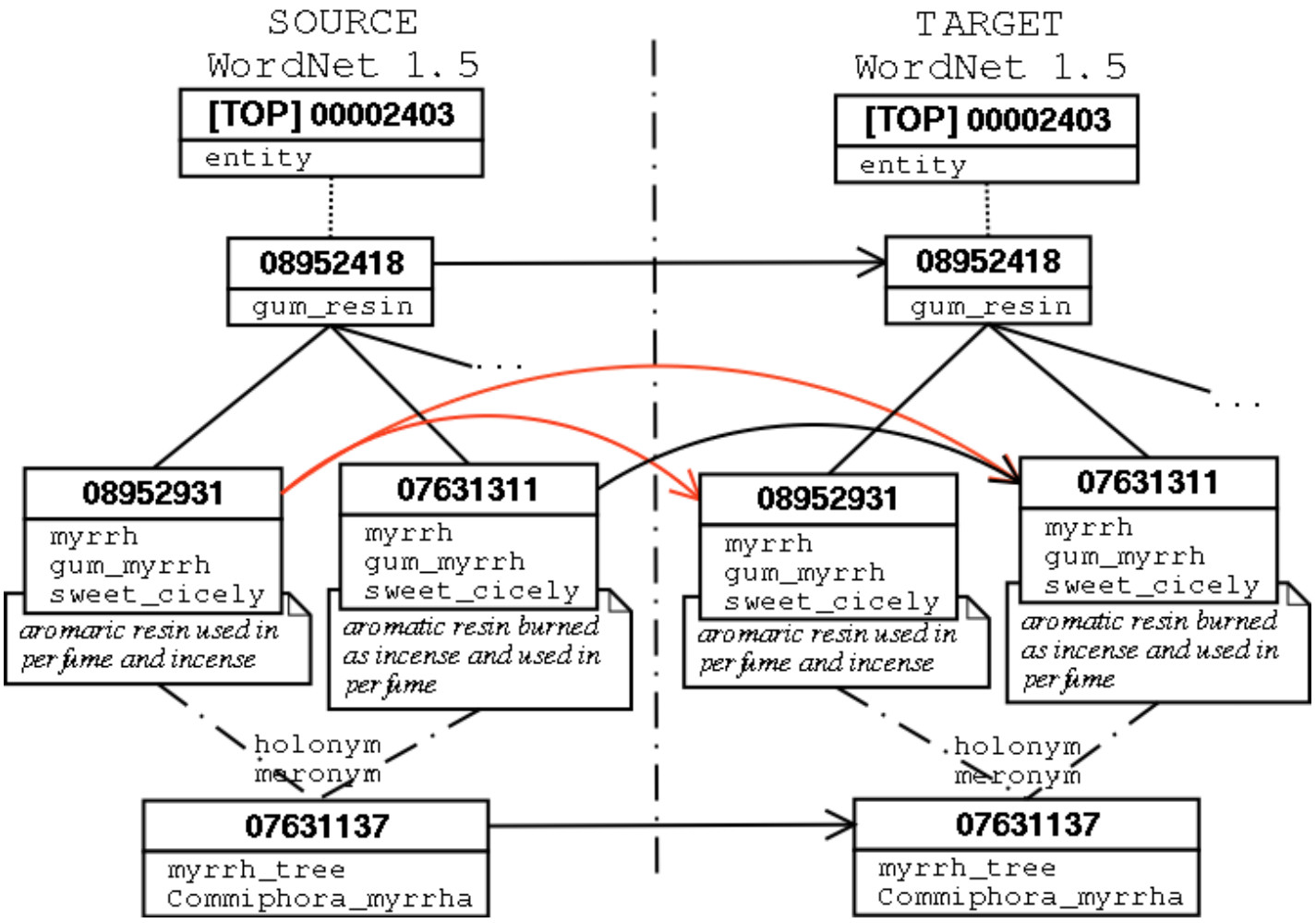
Mapping technology

- Mapping technology for connecting already existing semantic networks (i.e. wordnets)

- Relaxation Labelling Algorithm (Daudé et al. 2003)
- Iterative algorithm for function optimisation based on local information
- Local constraints with global effects!
  - Structural Constraints (hierarchical and non hierarchical)
  - Non structural constraints (synonym words, gloss, etc.)

- Given a set of constraints, provides de best possible mapping!

WordNet & EuroWordNet
Mapping technology



queso | Look up

Word | Nouns | Spanish_1.6

has_mero_madeof | English_1.6

Gloss Score Rels Full

English_1.6 Spanish_1.6 Catalan_1.6 Basque_1.6 Italian_1.6 English_1.7 English_1.7.1 English_2.0 Catalan_1.5 Spanish_1.5 English_1.5 English_2.1 English_3.0

Multilingual Central Repository

05881045n

gastronomy
base concept
food
Food
Artifact
Comestible
Solid
Substance

05881045n	32 **cheese_1**	a solid food prepared from the pressed curd of milk
05881045n	70 **queso_1**	
05881045n	70 **formatge_1**	
05881045n	6 **gazta_1**	
05881045n	34 cacio_1 formaggio_1	a solid food prepared from the pressed curd of milk
07850329n	38 **cheese_1**	

23 has_hyponym 1 has_mero_part 1 has_mero_madeof 6 gloss 1 has_hyperonym 64 rgloss
56 has_hyponym 1 has_mero_part 1 has_mero_madeof 1 has_hyperonym
56 has_hyponym 1 has_mero_part 1 has_mero_madeof 1 has_hyperonym
25 has_hyponym 1 has_mero_part 1 has_mero_madeof 1 has_hyperonym
24 has_hyponym 1 has_mero_part 1 has_mero_madeof 1 has_hyperonym
26 has_hyponym 1 has_mero_part 1 has_mero_madeof 3 gloss 2 has_hyperonym 81 rgloss

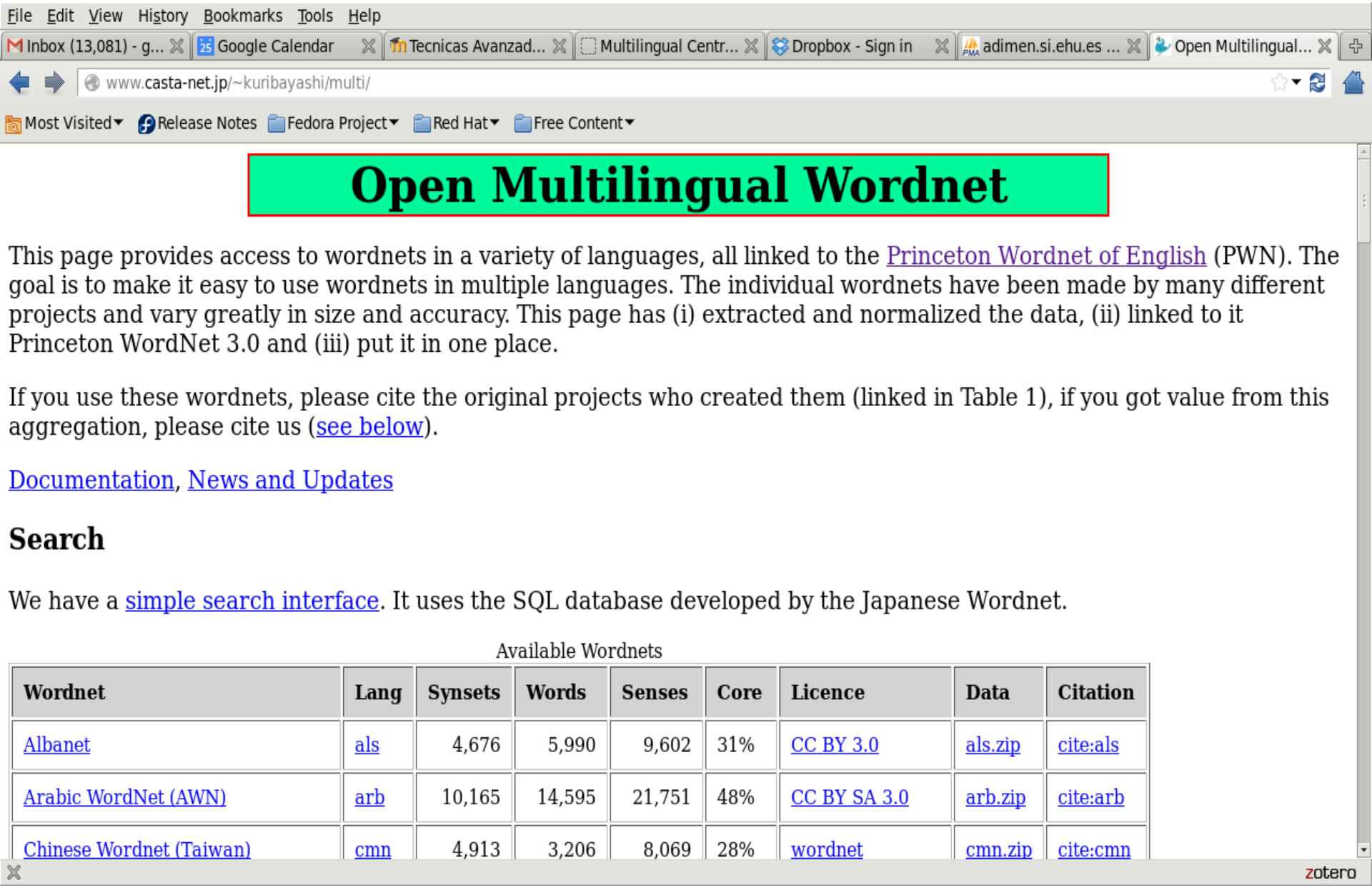
05880646n mn
99
gastronomy
food
Food
Comestible
Substance

05880646n	0 **curd_2**	coagulated milk; used to made cheese: *Little Miss Muffet sat on a tuffet eating some curds and whey;*
05880646n	0 **cuajada_1**	
05880646n	0 **quall_1**	
	quallada_1	
05880646n	0 **gatzatu_1**	
	mami_5	
05880646n	0 cagliata_1	coagulated milk; used to make cheese: *Little Miss Muffet sat on a tuffet eating some curds and whey;*
07849912n	0 **curd_2**	

- *6 gloss 1 has_hyperonym 1 rgloss 1 has_holo_madeof*
  1 has_hyperonym 1 has_holo_madeof
  1 has_hyperonym 1 has_holo_madeof
  1 has_hyperonym 1 has_holo_madeof
  1 has_hyperonym 1 has_holo_madeof
  3 gloss 1 has_hyperonym 7 rgloss 1 has_holo_madeof

Variants	Nombres	Verbos	Adjetivos	Adverbios	Synsets	%WN
EngWN3.0	147.360	25.051	30.004	5.580	118.431	100 %
SpaWN3.0	39.142	10.824	6.967	1.051	38.702	33 %
CatWN3.0	51.605	11.577	7.679	2	46.033	39 %
EusWN3.0	40.939	9.470	148	0	30.615	26 %
GalWN3.0	18.949	1.416	6.773	0	19.312	16 %
Glosas						
EngWN3.0	82.379	13.767	18.156	3.621	117.923	100 %
SpaWN3.0	12.533	3.325	1.917	670	18.445	16 %
CatWN3.0	6.294	44	840	1	7.179	6 %
EusWN3.0	2.690	2	0	0	2.692	2 %
GalWN3.0	4.997	2	3.111	0	8.111	7 %
Ejemplos						
EngWN3.0	10.433	11.583	15.615	3.674	41.305	100 %
SpaWN3.0	465	30	195	193	606	2 %
CatWN3.0	2.105	46	368	0	2.201	5 %
EusWN3.0	2.376	0	0	0	2.075	5 %
GalWN3.0	270	2	4.291	0	2.416	6 %

Cuadro 1: Número actual de *variants*, *synsets*, definiciones y ejemplos de cada wordnet.


File Edit View History Bookmarks Tools Help
Inbox (13,081) - g...
Google Calendar
Tecnicas Avanzad...
Multilingual Centr...
Dropbox - Sign in
adimen.si.ehu.es ...
Open Multilingual...
www.casta-net.jp/~kuribayashi/multi/
Most Visited
Release Notes
Fedora Project
Red Hat
Free Content
Open Multilingual Wordnet
This page provides access to wordnets in a variety of languages, all linked to the Princeton Wordnet of English (PWN). The goal is to make it easy to use wordnets in multiple languages. The individual wordnets have been made by many different projects and vary greatly in size and accuracy. This page has (i) extracted and normalized the data, (ii) linked to it Princeton WordNet 3.0 and (iii) put it in one place.
If you use these wordnets, please cite the original projects who created them (linked in Table 1), if you got value from this aggregation, please cite us (see below).
Documentation, News and Updates
Search
We have a simple search interface. It uses the SQL database developed by the Japanese Wordnet.
Available Wordnets
Wordnet | Lang | Synsets | Words | Senses | Core | Licence | Data | Citation
Albanet | als | 4,676 | 5,990 | 9,602 | 31% | CC BY 3.0 | als.zip | cite:als
Arabic WordNet (AWN) | arb | 10,165 | 14,595 | 21,751 | 48% | CC BY SA 3.0 | arb.zip | cite:arb
Chinese Wordnet (Taiwan) | cmn | 4,913 | 3,206 | 8,069 | 28% | wordnet | cmn.zip | cite:cmn
zotero


File Edit View History Bookmarks Tools Help
Inbox (13,083)...
Google Calendar
Tecnicas Avan...
Multilingual C...
Dropbox - Sign...
adimen.si.ehu....
Open Multiling...
BabelNet: a W...
lcl.uniroma1.it/babelnet/search.jsp?word=star&lang=EN
Most Visited
Release Notes
Fedora Project
Red Hat
Free Content
BabelNet 1.1.1
A very large multilingual ontology
search · explore · publications · download
Type a term: star
English
search
(examples: plane, apple, star, Italian, bus driver, calcio, drive#n, bus#n#en, horse#en, mela#it)
Noun
Title: star1 • ID: bn:00073964n • Type: Concept
explore
Senses:
star1
Estrella, Star, Étoile, Stern, Stella, Estrella
Estel fix, Estels, Estrela, Estrelles, Forma d'estrella, Estrella (astronomia), Estel·lar, Estel, Stellum, End of the Sun, Stars, Wishstar, Star fission, Etoile, Objet stellaire, Zentralstern, Stella (astronomia), Stelle, ★, Estrellas, ☆, Estrellas masivas, Estrella masiva, ESTRELLAS
estrella, estrelles, étoile, étoiles, sterne, stella, stelle, estrella,
zotero

WordNet & EuroWordNet
MCR example

vaso_1 02755829n 06-NOUN.ARTIFACT FACTOTUM
GLOSS: a glass container for holding liquids while drinking

TO: 1stOrderEntity-Form-Object
TO: 1stOrderEntity-Origin-Artifact
TO: 1stOrderEntity-Function-Container
TO: 1stOrderEntity-Function-Instrument
SUMO: Artifact

EN: drinking_glass glass
IT: bicchiere
BA: edontzi baso edalontzi
CA: got vas

DOBJ SemCor
00849393v 0.0074 polish shine smooth ...
00201878v 0.0013 beautify embellish prettify
00826635v 0.0010 get_hold_of take
00140937v 0.0001 ameliorate amend ...
00083947v 0.0000 alter change

WordNet & EuroWordNet
MCR example

vaso_2 04195626n 08-NOUN.BODY ANATOMY
GLOSS: a tube in which a body fluid circulates

TO: 1stOrderEntity-Form-Substance-Solid
TO: 1stOrderEntity-Origin-Natural-Living
TO: 1stOrderEntity-Composition-Part
TO: 1stOrderEntity-Function-Container
SUMO: BodyVessel

EN: vessel vas
IT: vaso canale
BA: hodi baso
CA: vas

DOBJ SemCor
01781222v 0.0334 be occur
00058757v 0.0072 inject shoot
01357963v 0.0068 flow travel_along
00055849v 0.0045 administer dispense ...

SUBJ SemCor
01831830v 0.0133 stop terminate
01357963v 0.0127 flow travel_along
01830886v 0.0043 discontinue
01779664v 0.0008 cease end finish ...

WordNet & EuroWordNet

MCR example

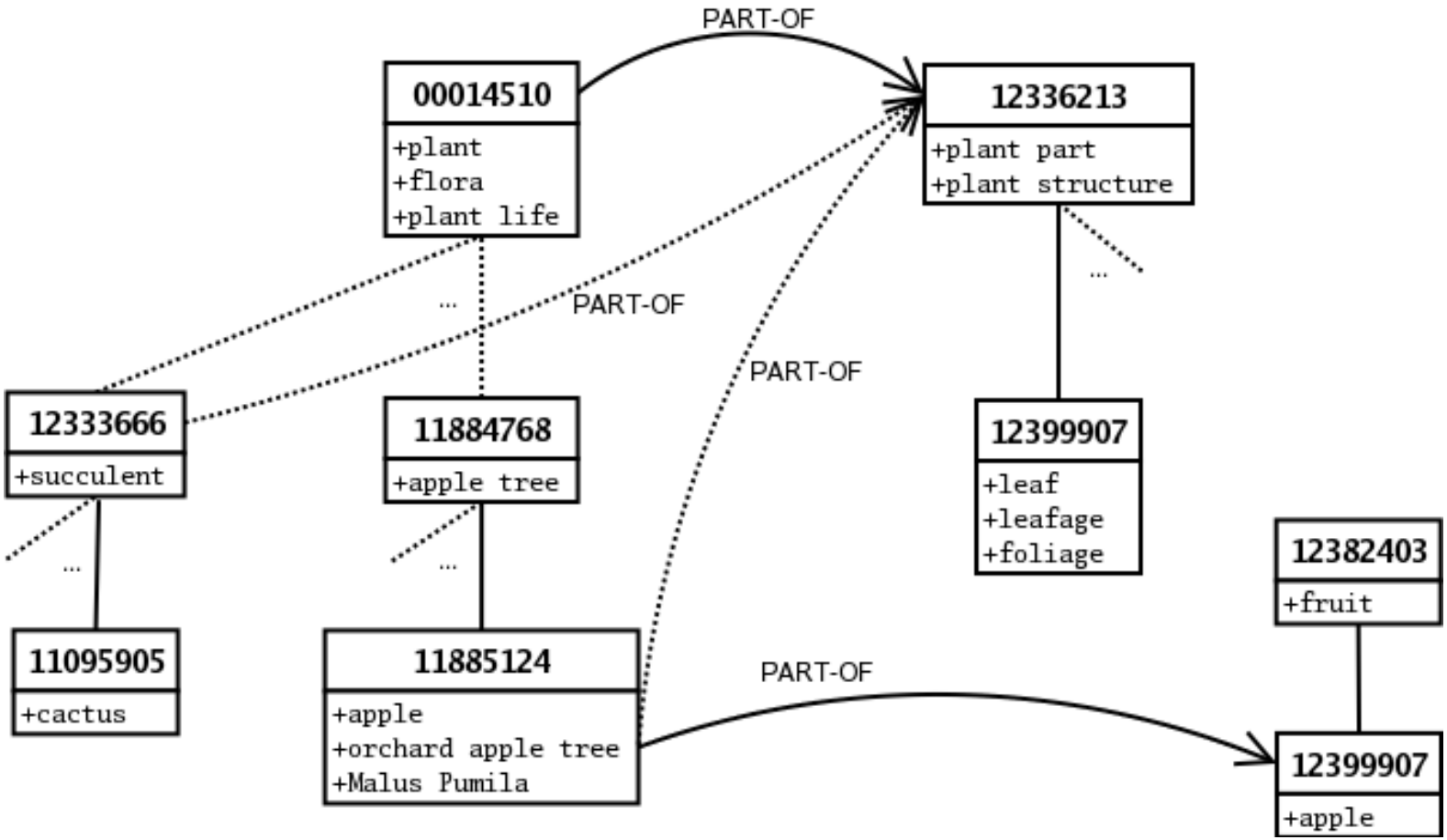
vaso_3 09914390n 23-NOUN.QUANTITY NUMBER
GLOSS: the quantity a glass will hold

TO: 1stOrderEntity-Composition-Part
TO: 2ndOrderEntity-SituationType-Static
TO: 2ndOrderEntity-SituationComponent-Quantity
SUMO: ConstantQuantity

EN: glassful glass
IT: bicchierata bicchiere
BA: basokada
CA: got vas

DOBJ SemCor
00795711v 0.0026 drink imbibe
01530096v 0.0009 accept have take
00786286v 0.0009 consume have ingest take take_in
01513874v 0.0001 acquire get

WordNet & EuroWordNet
Reasoning



- Does an orchard apple tree have leaves?
- Does an orchad apple tree have fruits?
- Does a cactus have leaves?

Suggested Upper Merged Ontology (SUMO)

The second annual SUMO Reasoning Prize took place at The CADE ATP System Competition at CADE 2009. Congratulations to 2009 first place winner Krystof Hoder for SInE-LTB 0.4.

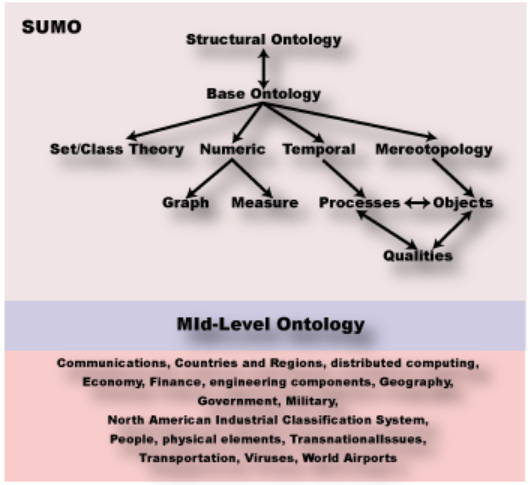
The Suggested Upper Merged Ontology (SUMO) and its domain ontologies form the largest formal public ontology in existence today. They are being used for research and applications in search, linguistics and reasoning. SUMO is the only formal ontology that has been mapped to all of the WordNet lexicon. SUMO is written in the SUO-KIF language. SUMO is free and owned by the IEEE. The ontologies that extend SUMO are available under GNU General Public License. Adam Pease is the Technical Editor of SUMO.

Find an English word and its corresponding formal term in SUMO:

English Word: table Noun Show

Features

- Mappings to all of WordNet
- Language generation templates for Hindi, Chinese, Italian, German, Czech and English
- Tool support for browsing and editing
- Largest free, formal ontology available, with 20,000 terms and 70,000 axioms when all domain ontologies are combined. These consist of SUMO itself, the MId-Level Ontology (MILO), and ontologies of Communications, Countries and Regions, distributed computing, Economy, Finance, engineering components, Geography, Government, Military (general, devices, processes, people), North American Industrial Classification System, People, physical elements, TransnationalIssues, Transportation, Viruses, World Airports A-K, World Airports L-Z, WMD. See also a large amount of instance content from DBPedia about people and the YAGO, project which includes millions of facts from Wikipedia merged with SUMO, and an initial merge of the Mondial geographical data with SUMO. The Open Biomedical Ontologies are being mapped to SUMO. Additional ontologies of terrorism are available on request.
- Richly axiomatized, not just a taxonomy. All terms are formally defined. Meanings are not dependent on a particular inference implementation. An inference and ontology management system however is provided. An additional system that supports visual editing, and does a better job of displaying the ontologies, especially in non-Western languages is the KSMSA system.



Building wordnets

German Rigau i Claramunt

german.rigau@ehu.es

IXA group

Departamento de Lenguajes y Sistemas Informáticos

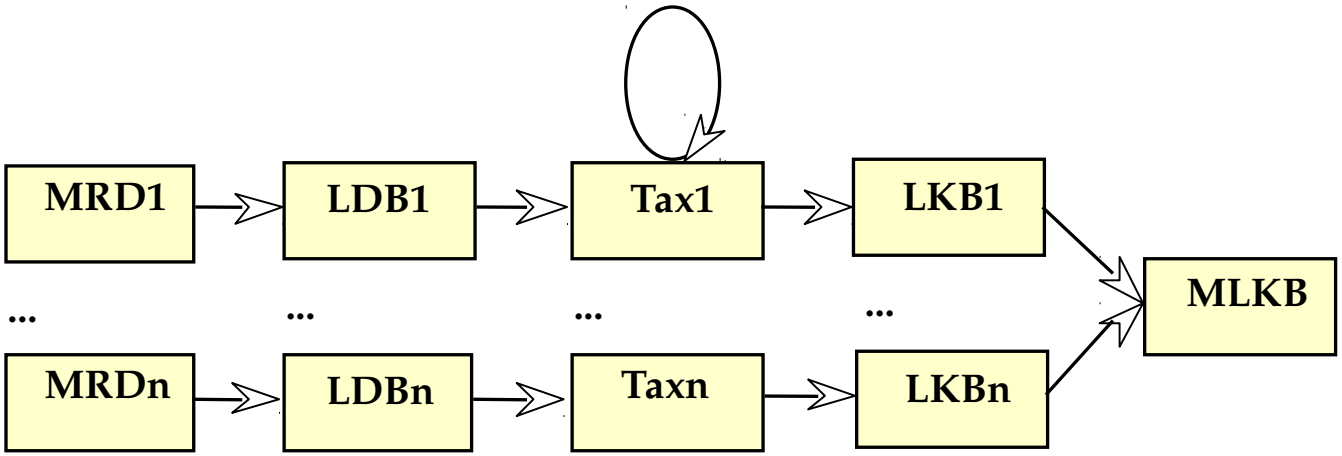
UPV/EHU

Outline

- **Merge approach**
  - Taxonomy construction: monolingual MRDs
  - Mapping taxonomies: bilingual MRDs
- Expand approach
  - Translation of synsets: bilingual MRDs
- Interface for manual revision
- Conclusions

Merge approach

Main Methodology



Merge approach

Main Methodology

- Taxonomy construction: (Rigau et al. 98, 97)
  - monolingual MRDs
  - **Step 1**: Selection of the main top beginners for a semantic primitive
  - **Step 2**: Exploiting genus, construction of taxonomies for each semantic primitive
- Mapping taxonomies: (Daudé et al. 99, 00, 01, 03)
  - bilingual MRDs
  - **Step 3**: Creation of translation links

Merge approach: Taxonomy Construction

Methodology

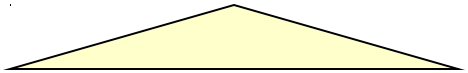
- Problems following a pure descriptive approach
  - Circularity
  - Errors and inconsistencies
  - Definitions with omitted genus
- Top dictionary senses do not usually represent useful knowledge for the LKB
  - Too general
  - Too specific

Merge approach: Taxonomy Construction

Methodology

Prescriptive approach

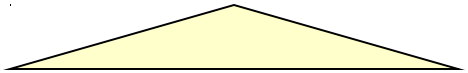
Manual construction of the **Top Structure**



Merge approach: Taxonomy Construction
Methodology

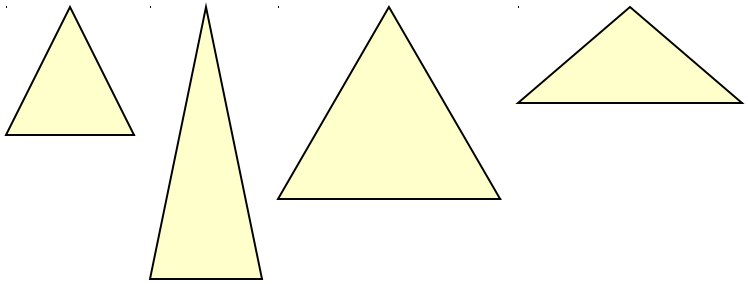
Prescriptive approach

Manual construction of the **Top Structure**



Descriptive approach

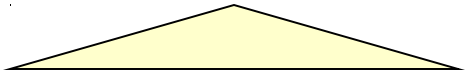
Acquiring implicit information from MRDs



Merge approach: Taxonomy Construction
Methodology

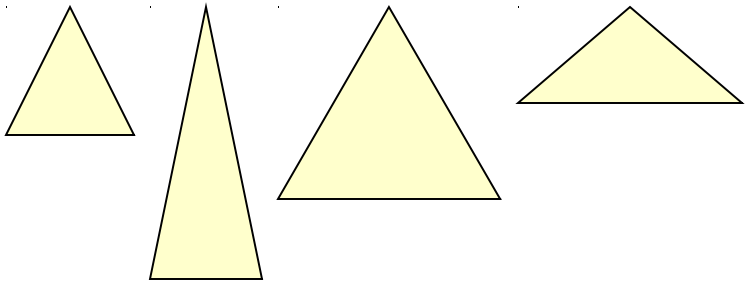
Prescriptive approach

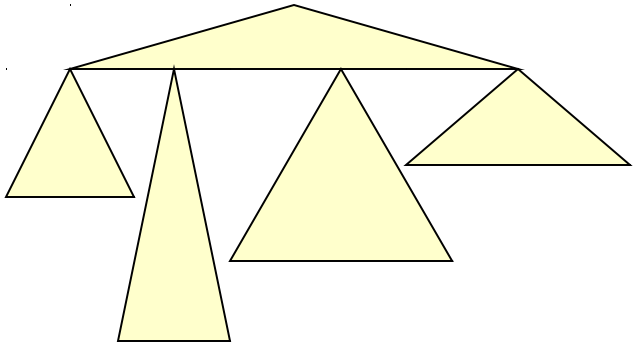
Manual construction of the **Top Structure**



Descriptive approach

Acquiring implicit information from MRDs





Merge approach: Taxonomy Construction

Step 1: Selection of the main top beginners

Word sense:	**zumo_1_1**
Attached-to:	**c_art_subst** type.
Definition:	**líquido** que se extrae de las flores, hierbas, frutos, etc.

Merge approach: Taxonomy Construction

Step 1: Selection of the main top beginners

A) Attaching DGILE senses to semantic primitives

1) First labelling:

Conceptual Distance (Rigau 94)

2) Second labelling:

Salient Words (Yarowsky 92)

B) Filtering Process

Merge approach: Taxonomy Construction

Step 1: Selection of the main top beginners

A.1) First labelling:

Conceptual Distance (Agirre et al. 94)

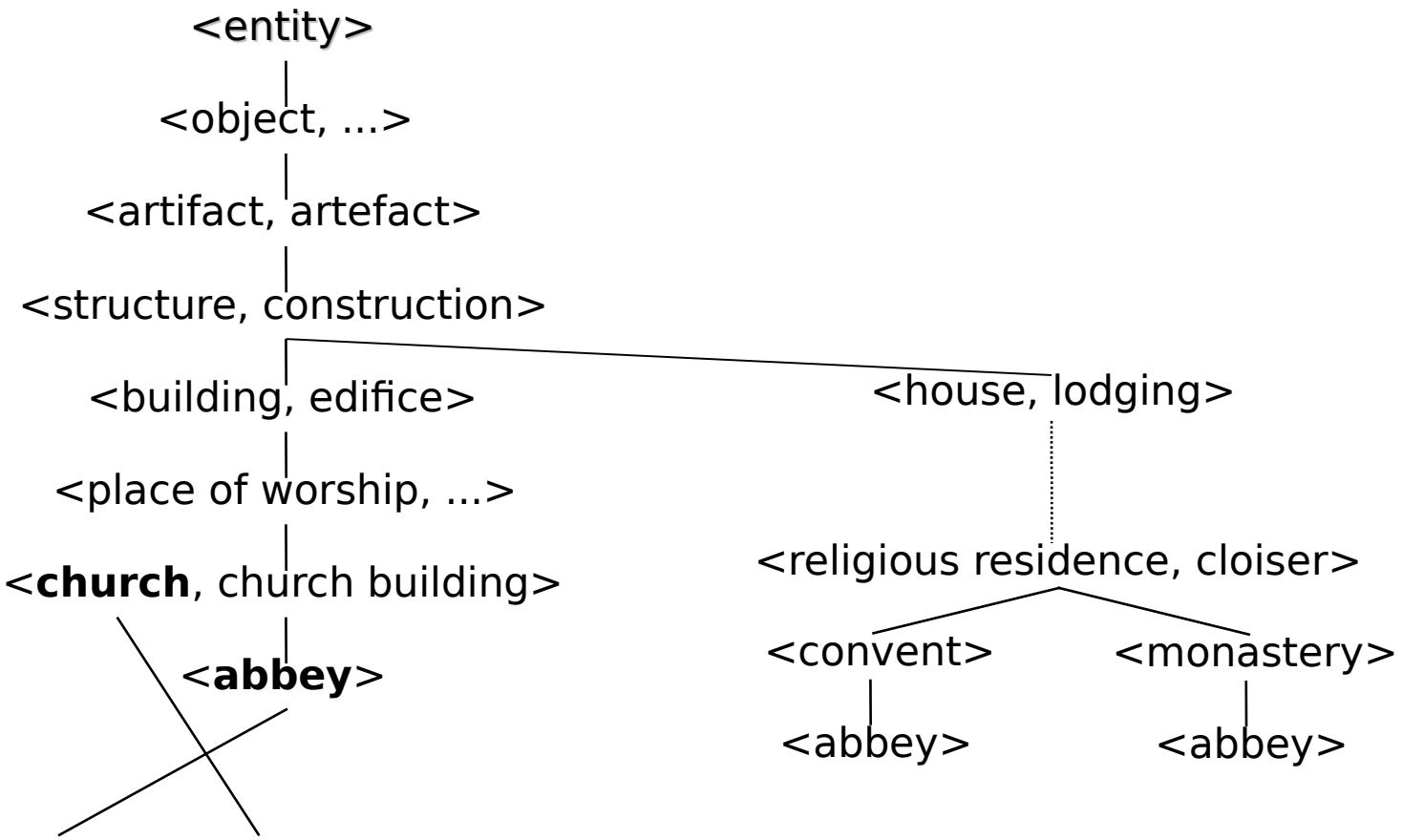
- length of the shortest path
- specificity of the concepts

$$\mathrm{dist}(w_1, w_2) = \min_{\substack{c_{1_i} \in w_1 \\ c_{2_i} \in w_2}} \sum_{c_k \in \mathrm{path}(c_{1_i}, c_{2_i})} \frac{1}{\mathrm{depth}(c_k)}$$

- using WordNet
- Bilingual dictionary

Merge approach: Taxonomy Construction

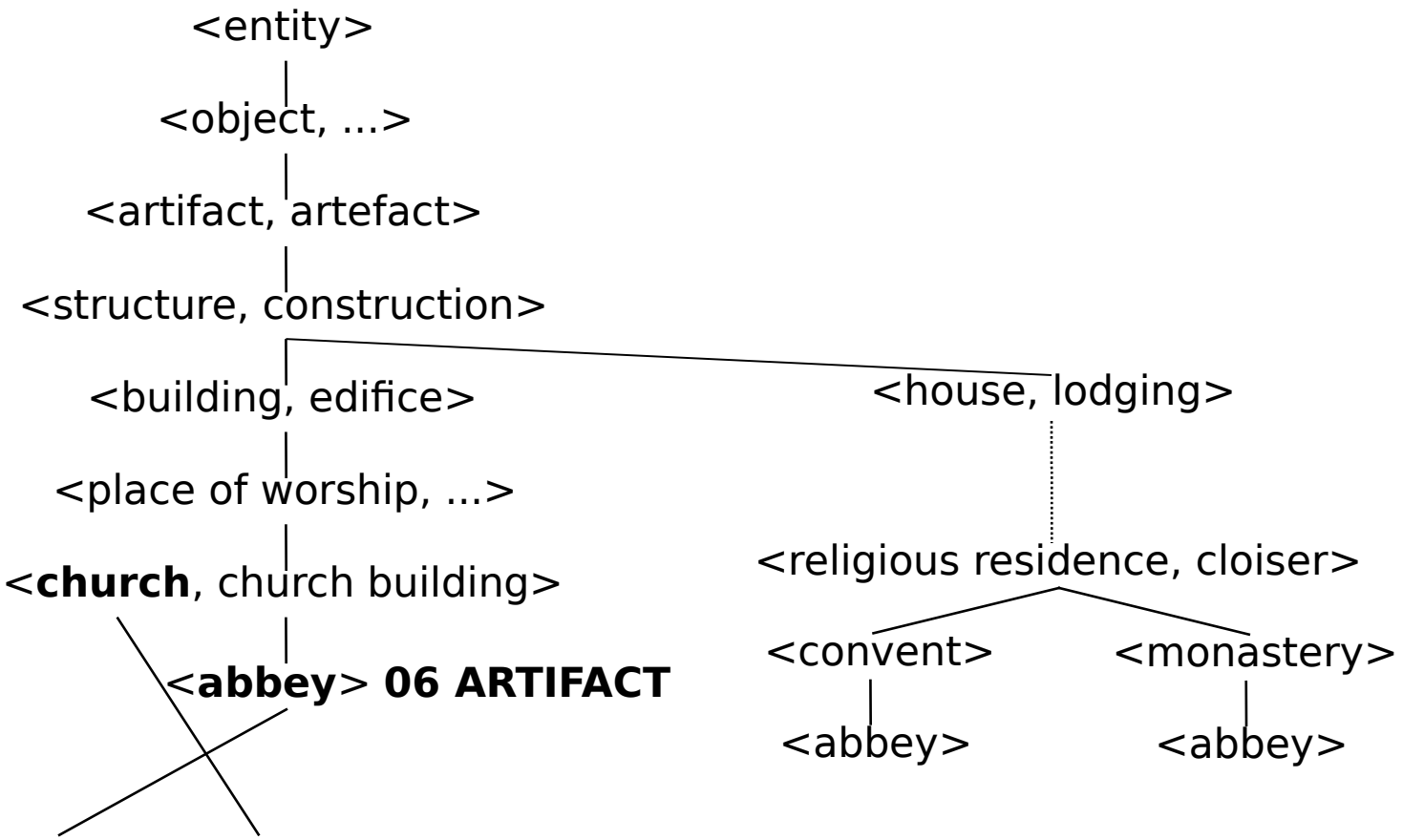
Step 1: Selection of the main top beginners



abadía_1_2 **Iglesia** o **monasterio** regido por un abad o abadesa (*abbey, a church or a monastery ruled by an abbot or an abbess*)

Merge approach: Taxonomy Construction

Step 1: Selection of the main top beginners



abadía_1_2 **Iglesia** o **monasterio** regido por un abad o abadesa (*abbey, a church or a monastery ruled by an abbot or an abbess*)

Merge approach: Taxonomy Construction

Step 1: Selection of the main top beginners

A.1) First labelling (Results)

- 29,205 labelled definitions (31% coverage)
- 61% accuracy at a sense level
- 64% accuracy at a file level

Merge approach: Taxonomy Construction

Step 1: Selection of the main top beginners

A.2) Second labelling:

Salient Words (Yarowsky 92)

$$AR(w,SC) = Pr(w|SC)\log_2 \frac{Pr(w|SC)}{Pr(w)}$$

Importance

- local frequency
- appears more significantly more often in the corpus of a semantic category than at other points in the whole corpus

Merge approach: Taxonomy Construction

Step 1: Selection of the main top beginners

- A.2) Second labelling (Results):

biberón_1_1 **ARTIFACT** 4.8399 **Frasco** de cristal ...
(*glass flask ...)*
biberón_1_2 **FOOD** 7.4443 **Leche** que contiene este fras
(*milk contained in that flask ...)*

- 86,759 labelled definitions (93%)
- 80% accuracy at a file level

Merge approach: Taxonomy Construction

Step 1: Selection of the main top beginners

B) Filtering process (FOODs)

- removes all genus terms

FILTER 1: not FOODs by the bilingual mapping

FILTER 2: appear more often as genus in other Semantic Primitive

FILTER 3: with a low frequency

Merge approach: Taxonomy Construction

Step 1: Selection of the main top beginners

B) Filtering process (FOOD Results)

	FILTER 1		FILTER 2	
LABEL2	#GT	Accuracy	#GT	Accuracy
LABEL2+F3>9	31	**94%**	31	**100%**
LABEL2+F3>8	35	**95%**	35	**100%**
LABEL2+F3>7	37	**91%**	37	**95%**
LABEL2+F3>6	43	**92%**	41	**94%**
LABEL2+F3>5	49	**92%**	47	**92%**
LABEL2+F3>4	55	**91%**	56	**91%**
LABEL2+F3>3	64	85%	65	87%
LABEL2+F3>2	85	82%	82	83%
LABEL2+F3>1	125	78%	123	82%

Merge approach: Taxonomy Construction

Step 2: Exploiting Genus

Word sense: **vino_1_1**
Hypernym: **zumo_1_1**.
Definition: zumo de uvas fermentado.

Word sense: **rueda_2_1**
Hypernym: **vino_1_1**.
Definition: vino procedente de la región de Rueda (Valladolid).

Merge approach: Taxonomy Construction

Step 2: Exploiting Genus

- Genus Sense Identification
  - 97% accuracy for nouns
- Genus Sense Disambiguation
  - Unrestricted WSD (coverage 100%)
  - Knowledge-based WSD (not supervised)
  - Eight Heuristics (McRoy 92)
    - Combining several lexical resources
    - Combining several methods

Merge approach: Taxonomy Construction

Step 2: Exploiting Genus

Results:

	Polysemous		Overall	
	Prec.	Cov.	Prec.	Cov.
Heuristic 1: Monosemous Genus Term	-	-	100%	16%
Heuristic 2: Entry Sense Ordering	70%	100%	75%	100%
Heuristic 3: Explicit Semantic Domain	100%	1%	100%	2%
Heuristic 4: Word Matching	72%	61%	79%	56%
Heuristic 5: Simple Concordance	57%	100%	65%	95%
Heuristic 6: Cooccurrence Vectors	60%	100%	66%	97%
Heuristic 7: Semantic Vectors	58%	99%	63%	94%
Heuristic 8: Conceptual Distance	49%	95%	57%	89%
Sum	**79%**	100%	**83%**	100%

Merge approach: Taxonomy Construction

Step 2: Exploiting Genus

Knowledge provided by each heuristic:

	Overall	
	Prec.	Cov.
- Heuristic 1: Monosemous Genus Term	79%	100%
- Heuristic 2: Entry Sense Ordering	72%	100%
- Heuristic 3: Explicit Semantic Domain	82%	98%
- Heuristic 4: Word Matching	81%	100%
- Heuristic 5: Simple Concordance	81%	100%
- Heuristic 6: Cooccurrence Vectors	81%	100%
- Heuristic 7: Semantic Vectors	81%	100%
- Heuristic 8: Conceptual Distance	77%	100%
Sum	**83%**	100%

Merge approach: Taxonomy Construction

Step 2: Exploiting Genus

FOOD	[Castellón 93]	F2+F3>9	F2+F3>4
Genus terms	63	33	68
Dictionary senses	392	952	1,242
Levels	6	5	6
Senses in level 1	2	18	48
Senses in level 2	67	490	604
Senses in level 3	88	379	452
Senses in level 4	67	44	65
Senses in level 5	87	21	60
Senses in level 6	6	0	13

F2+F3>9: 35,099 definitions

F2+F3>4: 40,754 definitions

No filters: 111,624 definitions

Merge approach: Taxonomy Construction

Step 2: Exploiting Genus

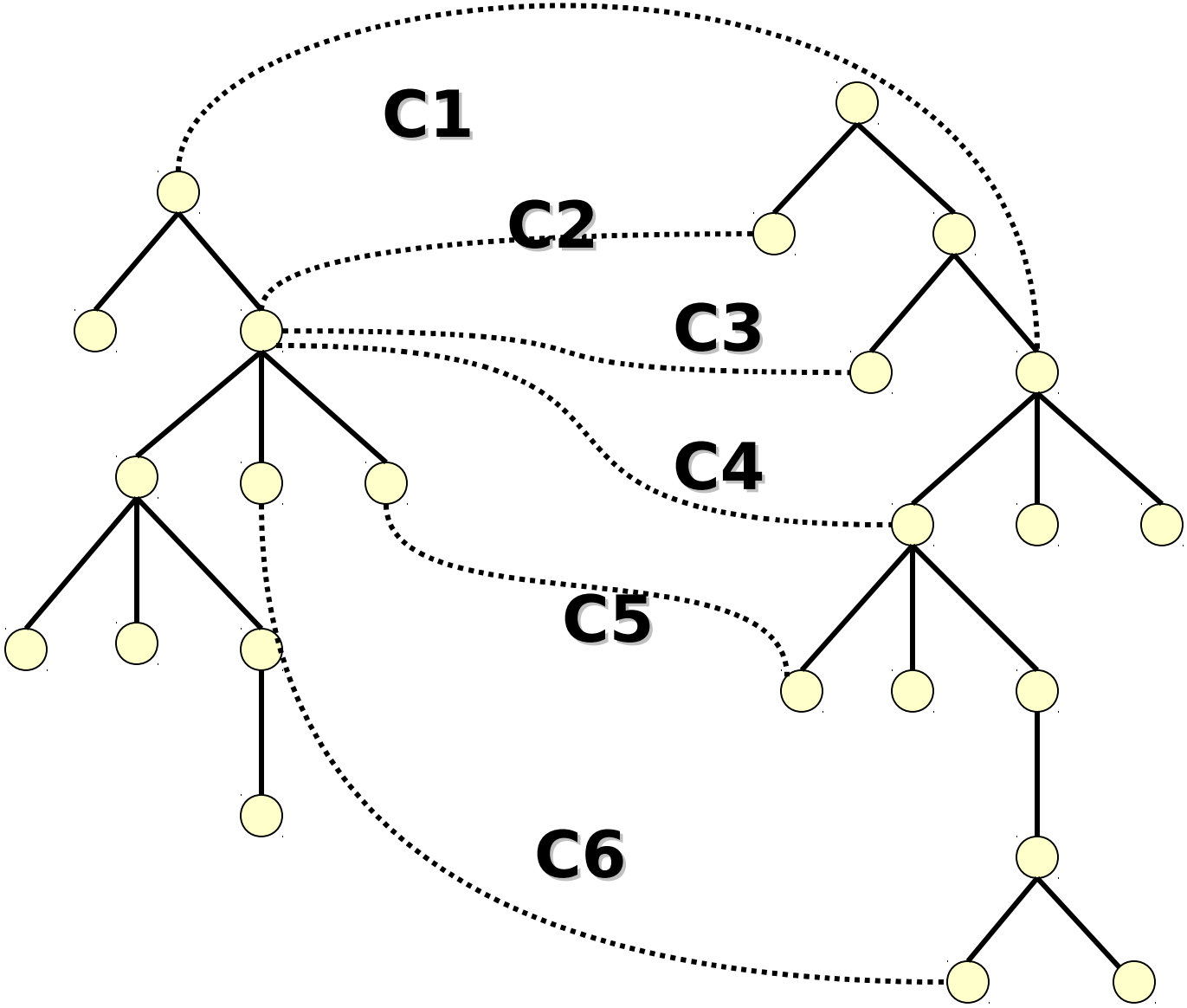
...

zumo_1_1	vino_1_1	quianti_1_1
zumo_1_1	vino_1_1	raya_1_8
zumo_1_1	vino_1_1	requena_1_1
zumo_1_1	vino_1_1	reserva_1_12
zumo_1_1	vino_1_1	ribeiro_1_1
zumo_1_1	vino_1_1	rioja_1_1
zumo_1_1	vino_1_1	roete_1_1
zumo_1_1	vino_1_1	rosado_1_3
zumo_1_1	vino_1_1	rueda_2_1
zumo_1_1	vino_1_1	sherry_1_1
zumo_1_1	vino_1_1	tarragona_1_1
zumo_1_1	vino_1_1	tintilla_1_1
zumo_1_1	vino_1_1	tintorro_1_1
zumo_1_1	vino_1_1	toro_3_1

...

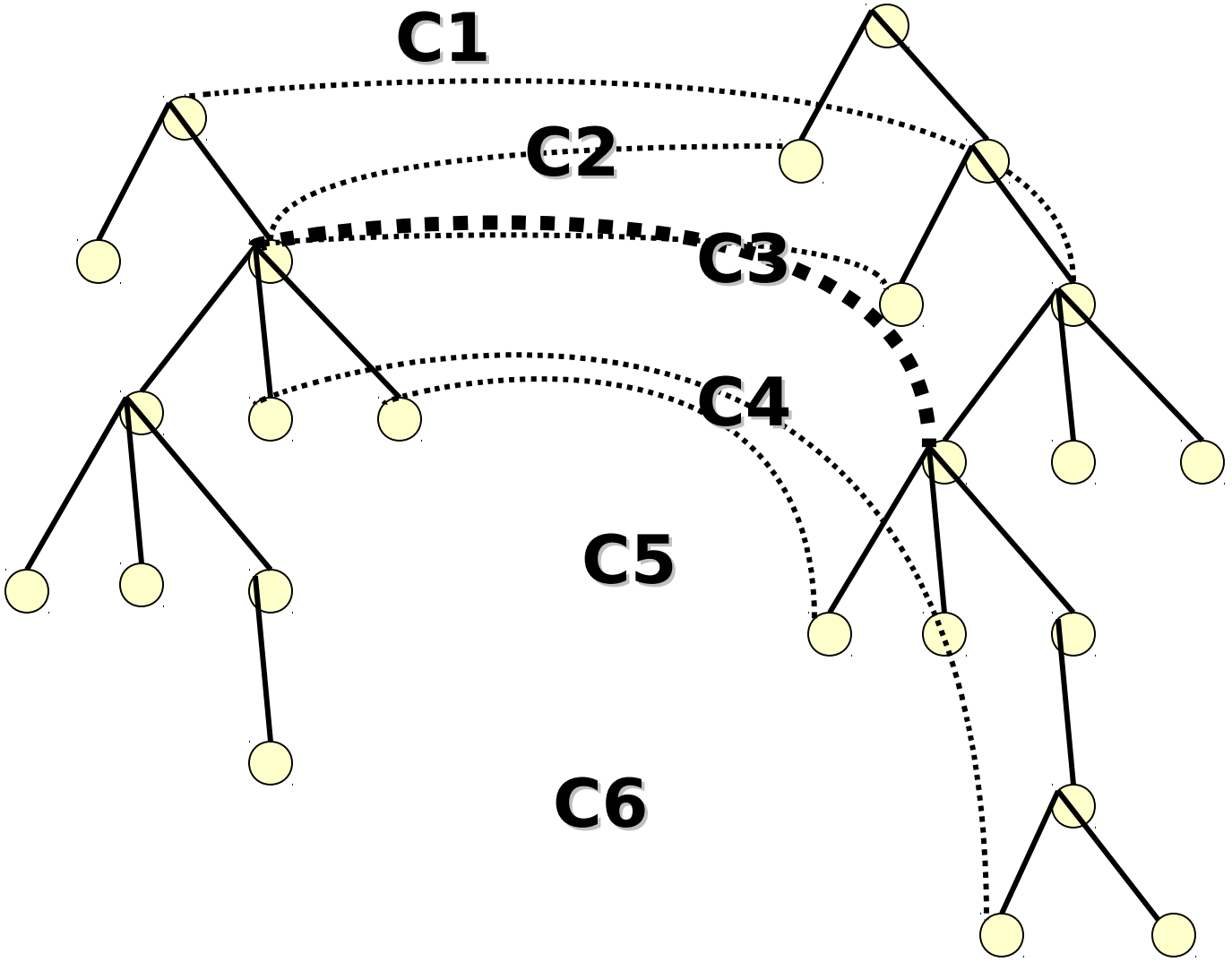
Merge approach: Mapping Taxonomies

Step 3: Creation of translation links_



Merge approach: Mapping Taxonomies

Step 3: Creation of translation links_



Merge approach: Mapping Taxonomies

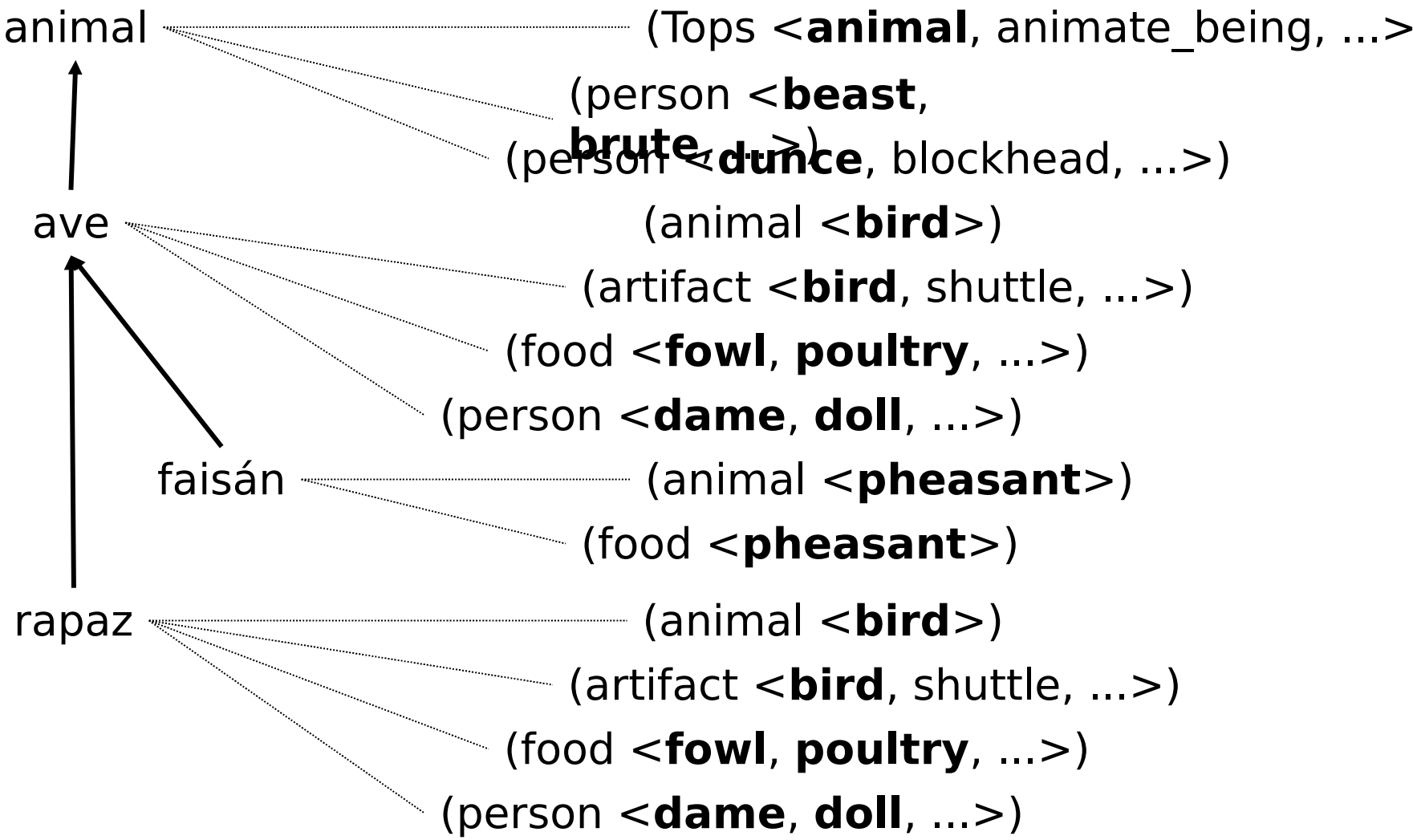
Step 3: Creation of translation links_

- Connecting already existing Hierarchies
  - Relaxation labelling Algorithm
  - Constraints

- Between
  - Spanish taxonomy automatically derived from an MRD (Rigau et al. 98)
  - WordNet
  - using a bilingual MRD

Merge approach: Mapping Taxonomies

Step 3: Creation of translation links_



Merge approach: Mapping Taxonomies

Step 3: Relaxation Labelling algorithm_

- Iterative algorithm for function optimisation based on local information
- it can deal with any kind of constraints
  - variables (senses of the taxonomy)
  - labels (synsets)
- Finds a weight assignment for each possible label for each variable
  - weights for the labels of the same variable add up to one
  - weight assignation satisfies -to the maximum possible extent- the set of constraints

Merge approach: Mapping Taxonomies

Step 3: Relaxation Labelling algorithm_

1) Start with a random weight assignment

2) Compute the support value for each label of each variable (according to the constraints)

3) Increase the weights of the labels more compatible with context and decrease those and decrease those of the less compatible labels.

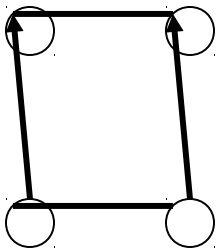
4) If a stopping/convergence is satisfied, stop, otherwise go to step 2.

Merge approach: Mapping Taxonomies

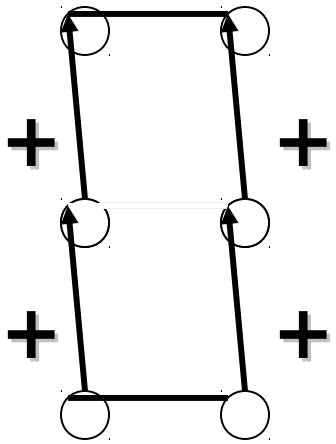
Step 3: Constraints_

- Rely on the taxonomy structure
- Coded with three characters
  - X: Spanish Taxonomy, I (immediate), A (ancestor)
  - Y: English Taxonomy,
  - X: Relation, E (hypernym), O (hyponym), B (both)
- Examples:

IIE



AAB



Merge approach: Mapping Taxonomies

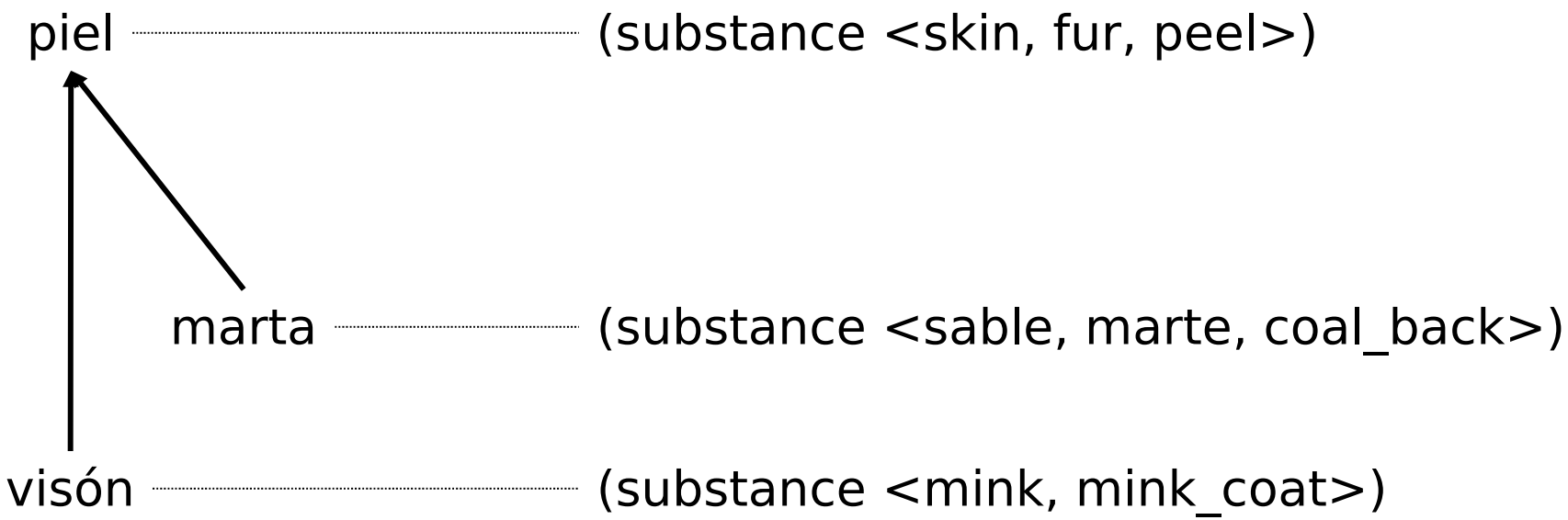
Step 3: Results_

Poly	TOK, FOK	TOK, FNOK	total
animal	279 (90%)	30 (91%)	209 (90%)
food	166 (94%)	3 (100%)	169 (94%)
cognition	198 (67%)	27 (90%)	225 (69%)
communication	533 (77%)	40 (97%)	573 (78%)

all	TOK, FOK	TOK, FNOK	total
animal	424 (93%)	62 (95%)	486 (90%)
food	166 (94%)	83 (100%)	249 (96%)
cognition	200 (67%)	245 (90%)	445 (82%)
communication	536 (77%)	234 (97%)	760 (81%)

Merge approach: Mapping Taxonomies

Step 3: Example_



Outline_

- Merge approach
  - Taxonomy construction: monolingual MRDs
  - Mapping taxonomies: bilingual MRDs
- **Expand approach**
  - Translation of synsets: bilingual MRDs
- Interface for manual revision
- Conclusions

Expand approach

Translation of synsets_

- Take one WordNet as starting point
- Translate synsets:
  - English: <car, automobile>
  - Basque: <auto, berebil>
- We obtain a structurally similar WordNet in another language, but some of the synsets will be missing
- Use bilingual dictionary

  maintien n.m. (attitude) bearing; (conservation) maintenance

  1. Keep bilingual senses (Agirre & Rigau 95)

  maintien1: (attitude) bearing **maintien2**: (conservation) maintenance

  2. Produce all translation pairs (Atserias et al. 97)

  maintien - bearing

  maintien - maintenance

Expand approach

Translation of synsets_

- Used to produce the first version of the nominal part of the Spanish WordNet
- Based on WN 1.5
- Both directions in bilingual dictionary merged
  - Spanish/English: 19,443 translation pairs
  - English/Spanish: 16,324 translation pairs
  - Harmonized bilingual: 28,131 translation pairs
  - Overlap with WordNet: 12,665 nouns (14%)
- Two methods:
  - class methods: consider only pairings
  - conceptual distance methods: consider similarity of synsets

Expand approach

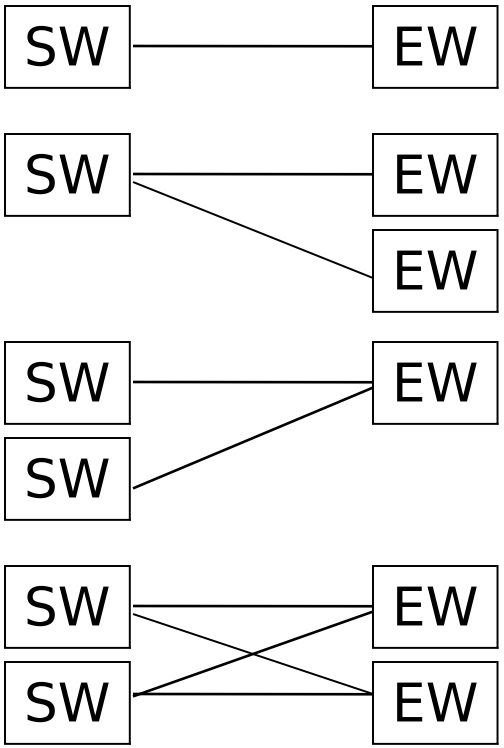
Methods_

- Ten class methods
  - Four monosemic criteria
  - Four polysemic criteria
  - Two hybrid criteria
- Three conceptual distance methods
  - CD1: using pairwise word coocurrences
  - CD2: using headword and genus
  - CD3: using bilingual Spanish entries with multiple translations

Expand approach

Class methods_

- Four possible configurations for pairs which either share an English word or an Spanish word: connected graph.

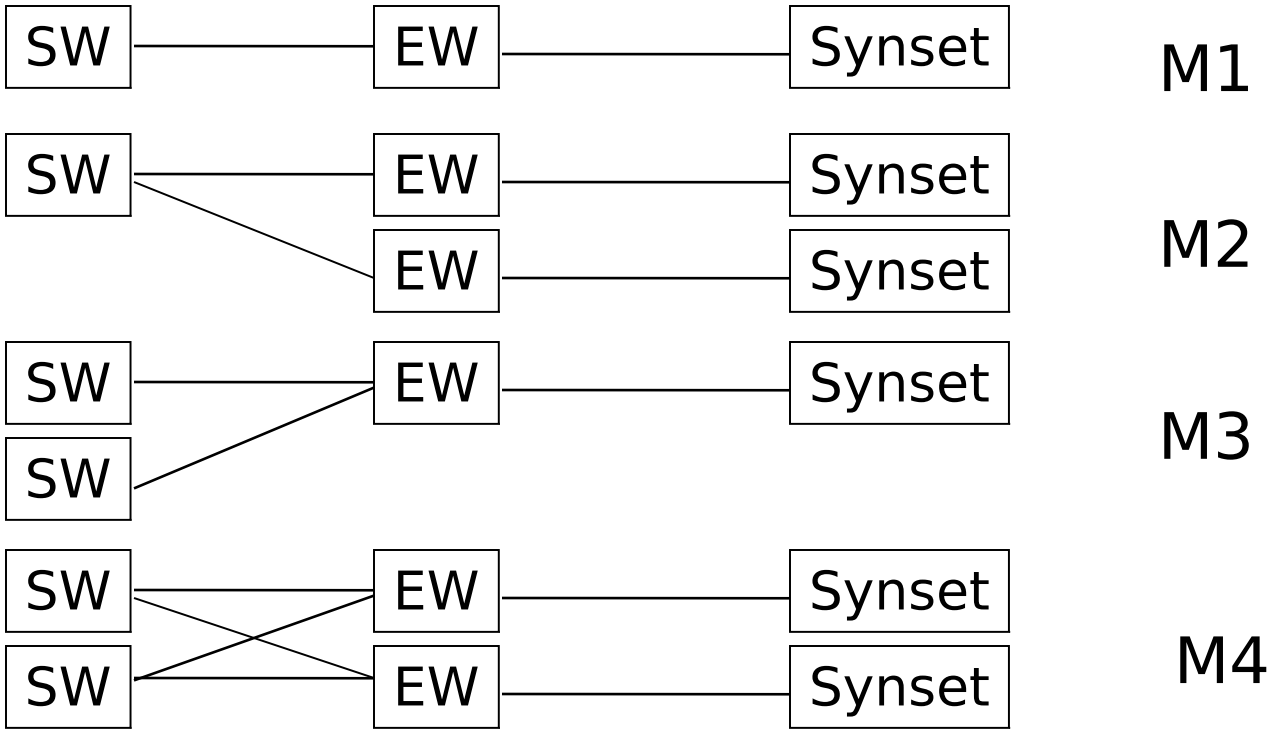


Expand approach

Class methods_

4 monosemous class methods:

- All English words involved are monosemous in WN

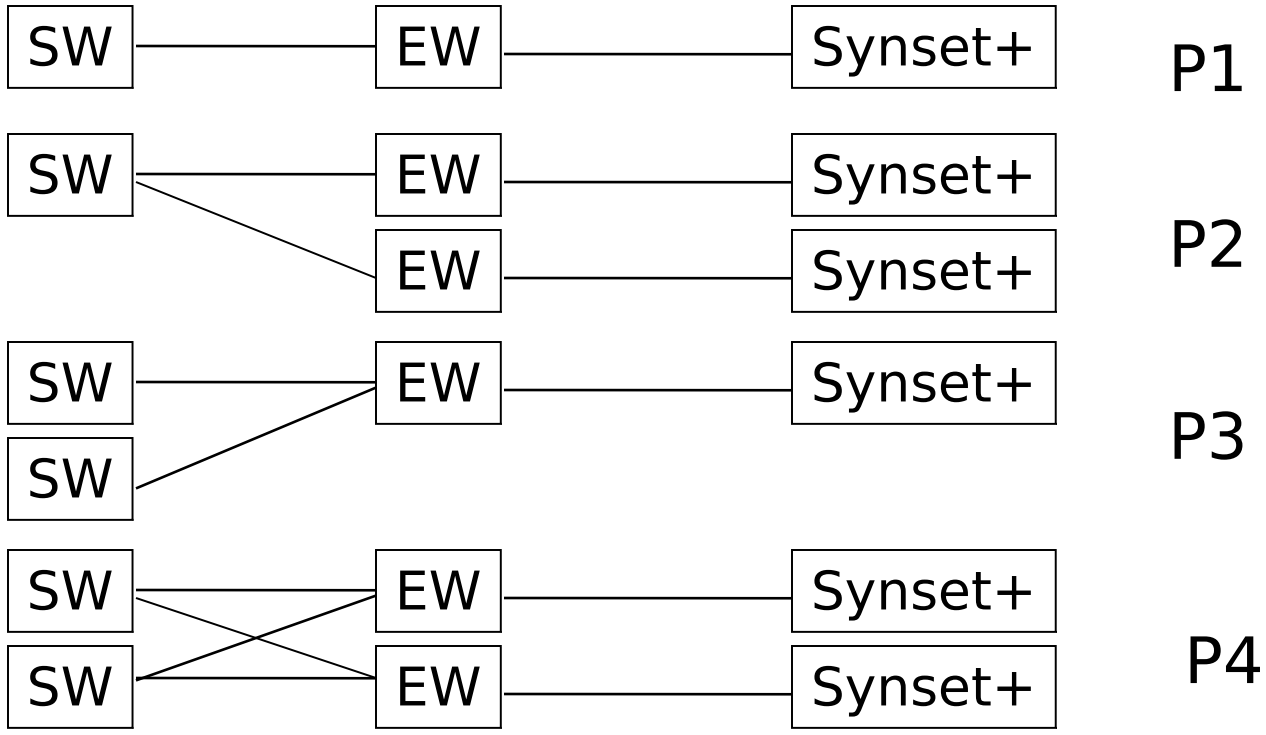


Expand approach

Class methods_

4 polysemous class methods:

- At least 1 English word involved is polysemous

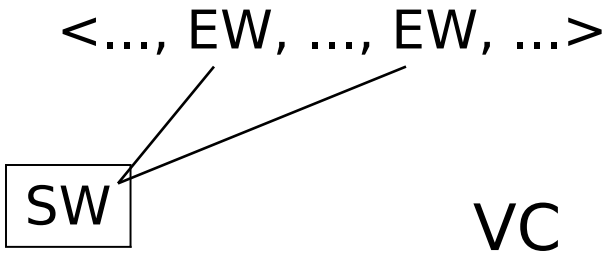


Expand approach

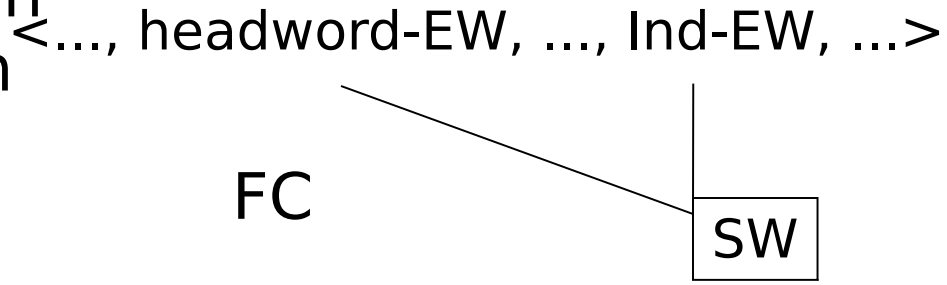
Class methods_

2 other class methods

- Variant criterion:
  two synonyms share
  a single SW



- Field criterion:
  use field indicators in
  bilingual entry when
  available



Expand approach
Class methods_

Ten class methods (results)

Criterion	#links	#synsets	#words	%ok
mono1	3697	3583	3697	92
mono2	935	929	661	89
mono3	1863	1158	1863	89
mono4	2688	1328	2063	85
poly1	5121	4887	1992	80
poly2	1450	1426	449	75
poly3	11687	6611	3165	58
poly4	40298	9400	3754	61
Variant	3164	2195	2261	85
Field	510	379	421	78

Expand approach

Conceptual distance methods_

Conceptual Distance Methods (Agirre et al. 94)

- length of the shortest path
- specificity of the concepts

$$\mathrm{dist}(w_1, w_2) = \min_{\substack{c_{1_i} \in w_1 \\ c_{2_i} \in w_2}} \sum_{c_k \in \mathrm{path}(c_{1_i}, c_{2_i})} \frac{1}{\mathrm{depth}(c_k)}$$

- Using WordNet
- Bilingual dictionary

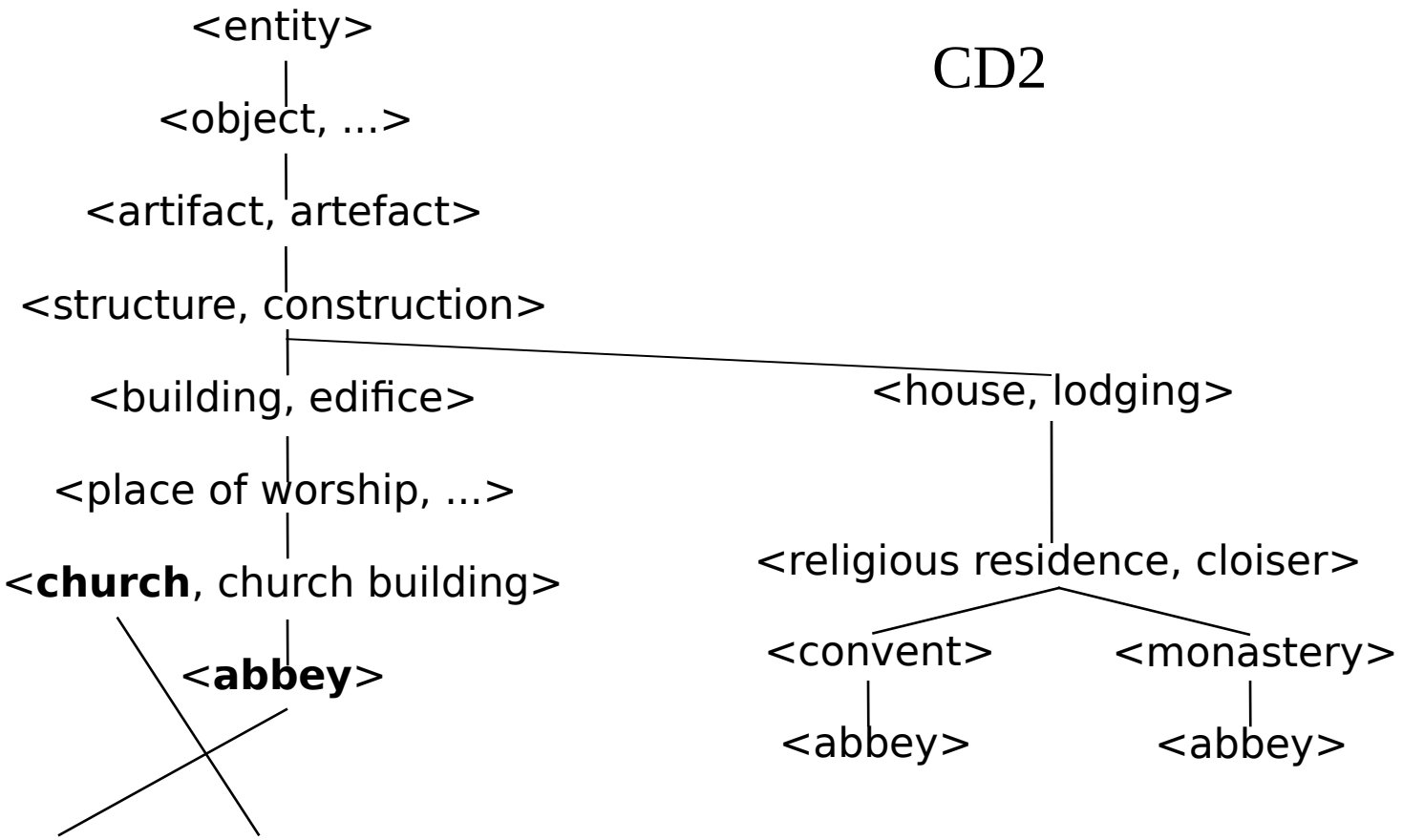
Expand approach

Conceptual distance methods_

- Three conceptual distance methods
  - CD1: using pairwise word coocurrences from monolingual dict.
  - CD2: using headword and genus from monolingual def.
  - CD3: using bilingual Spanish entries with multiple translations

Expand approach

Conceptual distance methods_



abadía_1_2 **Iglesia** o **monasterio** regido por un abad o abadesa
(*abbey, a church or a monastery ruled by an abbot or an abbess*)

Expand approach

Conceptual distance methods_

Three conceptual distance methods

Criter.	#links	#synsets	#words	%ok
CD - 1	23,828	11,269	7,283	56
CD - 2	24,739	12,709	10,300	61
CD - 3	4,567	3,089	2,313	75

Expand approach
Quality_

- Keep SW-synset pairs produced by methods with precision above 85%
  - mono1
  - mono2
  - mono3
  - mono4
  - variant
- But, if two different methods propose the same SW-synset pair, it could get better confidence
  - try pairwise combinations of methods

Expand approach

Combination of methods_

Combinations of methods: higher precision in some cases

		method2					
method1		cd2	cd3	p1	p2	p3	p4
cd1	size	15736	1849	2076	556	3146	15105
	%ok	79	85	86	86	72	64
cd2	size	0	2401	2536	592	3777	13246
	%ok	0	86	88	86	75	67
cd3	size	0	0	205	180	215	3114
	%ok	0	0	95	95	100	77
p1	size	0	0	0	0	77	178
	%ok	0	0	0	0	100	88
p2	size	0	0	0	0	28	78
	%ok	0	0	0	0	77	96

Expand approach
Results_

- SpWN v 0.1
- BasqueWN v 0.1:
  - 2 bilingual dictionaries
  - apply first 8 class methods only

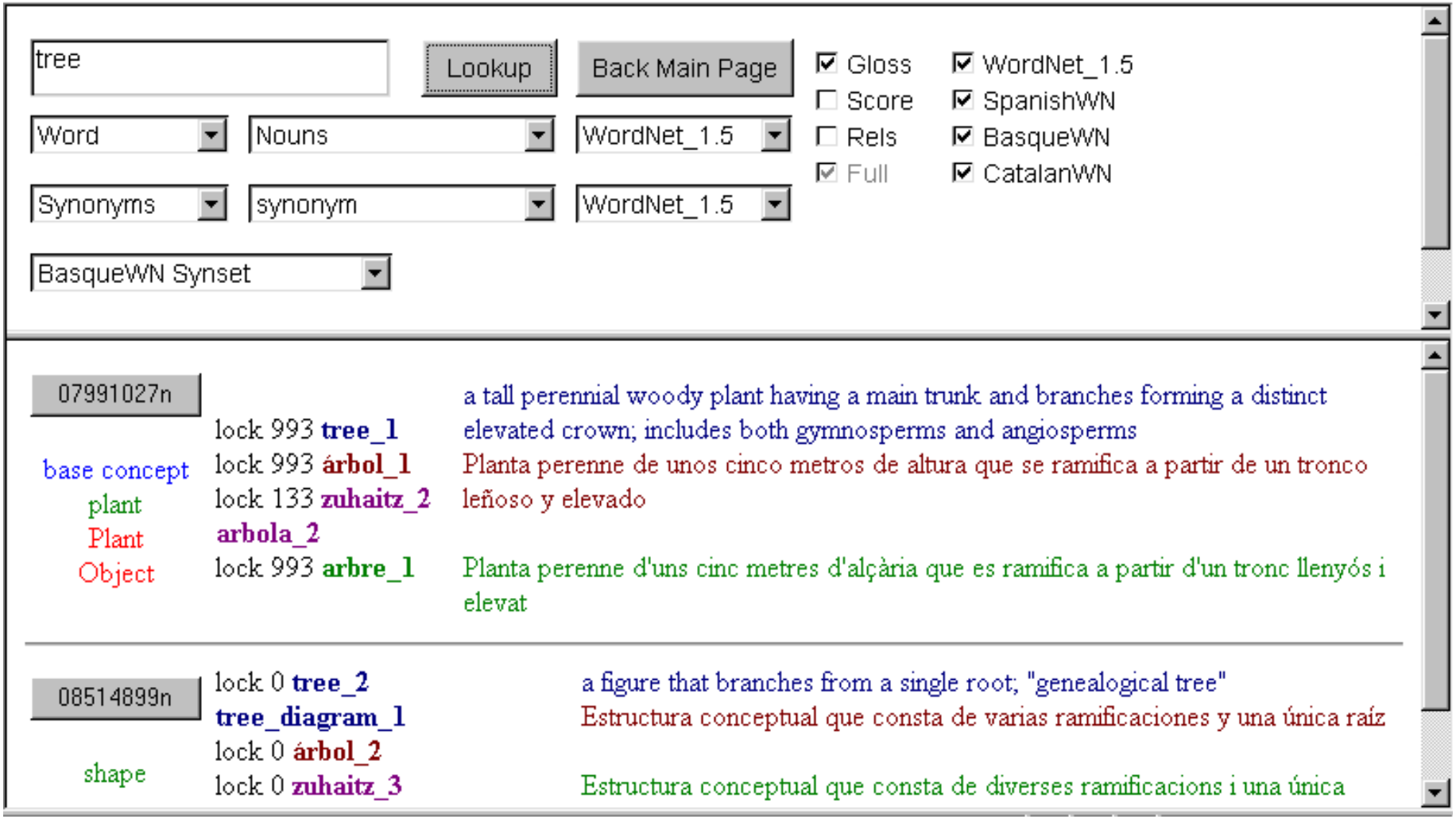
WNs	#links	#synsets	#word	#CS	#poly links
SpWN v0.0	10,982	7,131	8,396	87.4	1,777
Combination	7,244	5,852	3,939	85.6	2,075
SpWN v0.1	15,535	10,786	9,986	86.4	3,373
BasqueWN v0.1	41,107	23,486	22,166	>80.0	-

Outline_

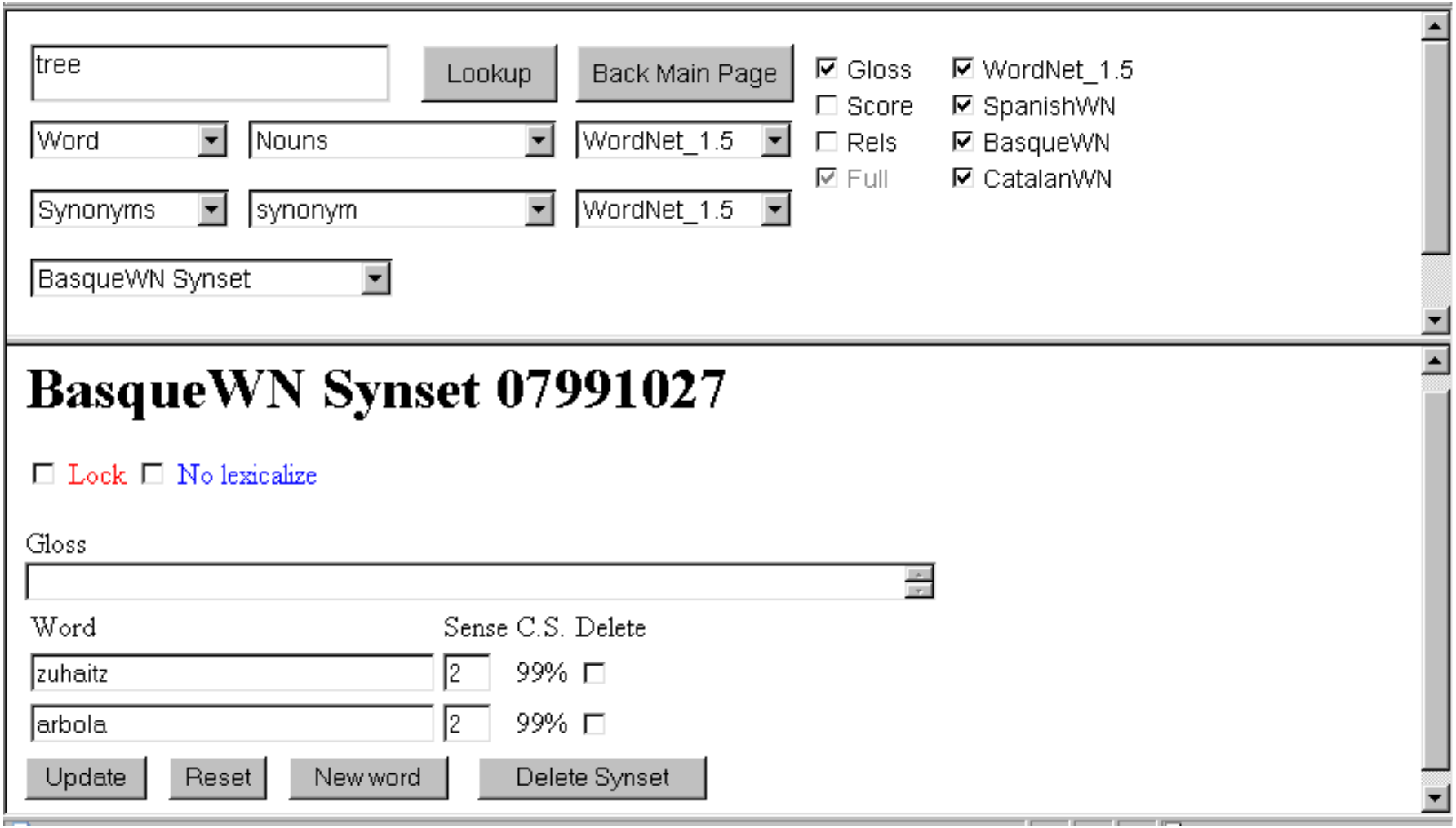
- Merge approach
  - Taxonomy construction: monolingual MRDs
  - Mapping taxonomies: bilingual MRDs
- Expand approach
  - Translation of synsets: bilingual MRDs
- **Interface for manual revision**
- Conclusions

Building wordnets

Web EuroWordNet Interface_



Web EuroWordNet Interface_



Outline_

- Merge approach
  - Taxonomy construction: monolingual MRDs
  - Mapping taxonomies: bilingual MRDs
- Expand approach
  - Translation of synsets: bilingual MRDs
- Interface for manual revision
- **Conclusions**

Conclusions_

- methods to automatically produce **preliminary versions**
- methods mainly for nouns
- need to manually revise
- merge approach
  - method to produce native hierarchies and word senses
  - trust lexicographer’s hierarchies
  - need to map to ILI in independent process
- expand approach
  - method to translate English WN’s synsets
  - trusts WN’s hierarchies, sense distinctions
  - mapping to ILI for free

Building wordnets
Conclusions_

- merge approach
  - manual work:
    - revising and re-organizing the automatic hierarchies (hard)
    - revising automatic mapping (very hard)
  - allows for integration of data from monolingual dictionary
    - definition text itself
    - lexico-semantic relations from definitions
- expand approach
  - manual work:
    - revise proposed translations (fast)
    - review the rest of the synsets (many)
    - include glosses

Conclusions

- Interface to speed up manual work
- Downloadable soon:
  - WN 1.5 in data-base format
  - Interface
- WordNets can be checked at:
  - http://adimen.si.ehu.es

Wordnet & EuroWordNet

German Rigau i Claramunt

german.rigau@ehu.es

IXA group

Departamento de Lenguajes y Sistemas Informáticos

UPV/EHU