Building wordnets

German Rigau i Claramunt

german.rigau@ehu.es

IXA group

Departamento de Lenguajes y Sistemas Informáticos

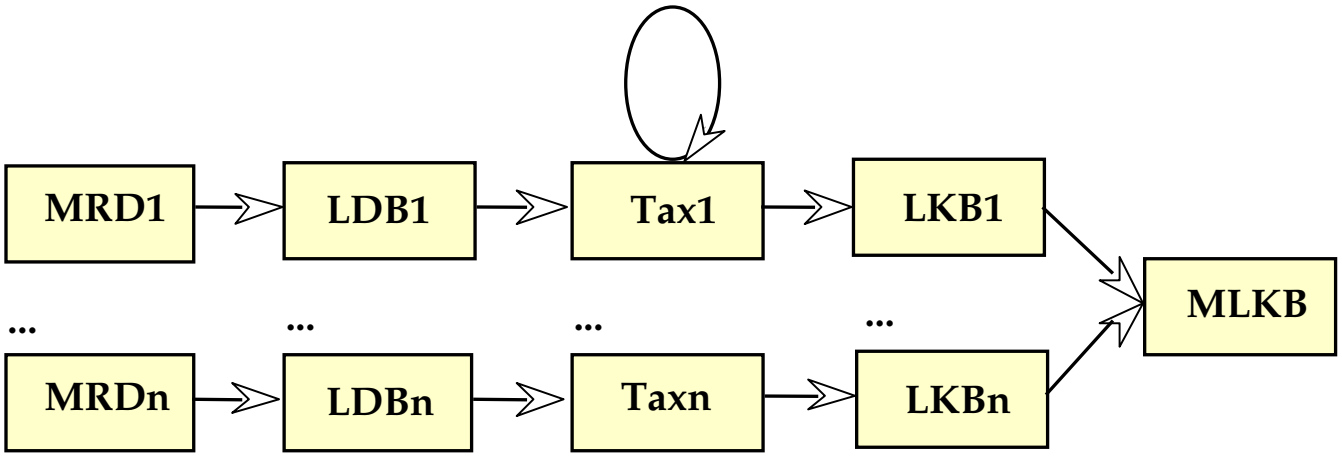
UPV/EHU

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

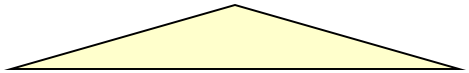
- Problems following a pure descriptive approach
  - Circularity
  - Errors and inconsistencies
  - Definitions with omitted genus
- Top dictionary senses do not usually represent useful knowledge for the LKB
  - Too general
  - Too specific

Merge approach: Taxonomy Construction

Methodology

Prescriptive approach

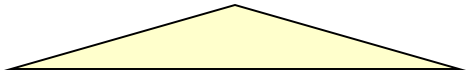
Manual construction of the **Top Structure**



Merge approach: Taxonomy Construction
Methodology

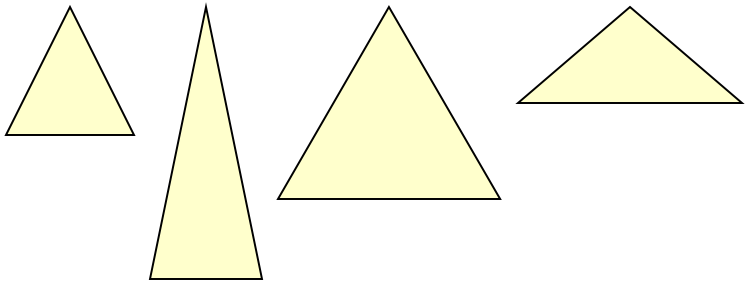
Prescriptive approach

Manual construction of the **Top Structure**



Descriptive approach

Acquiring implicit information from MRDs

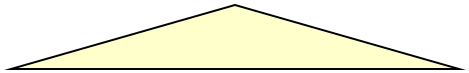


Merge approach: Taxonomy Construction

Methodology

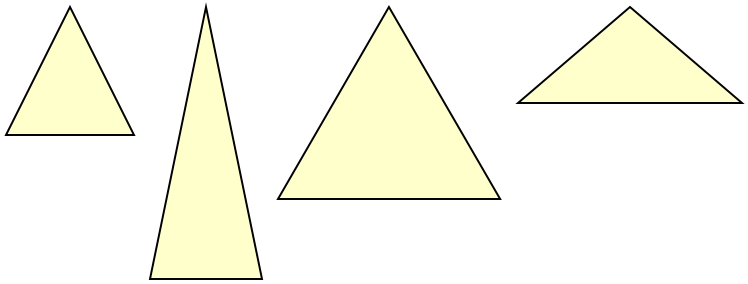
Prescriptive approach

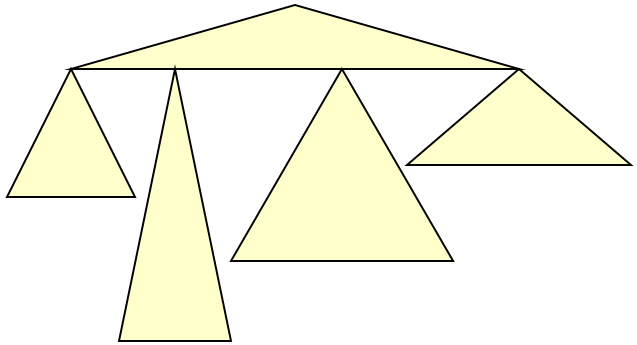
Manual construction of the **Top Structure**



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)

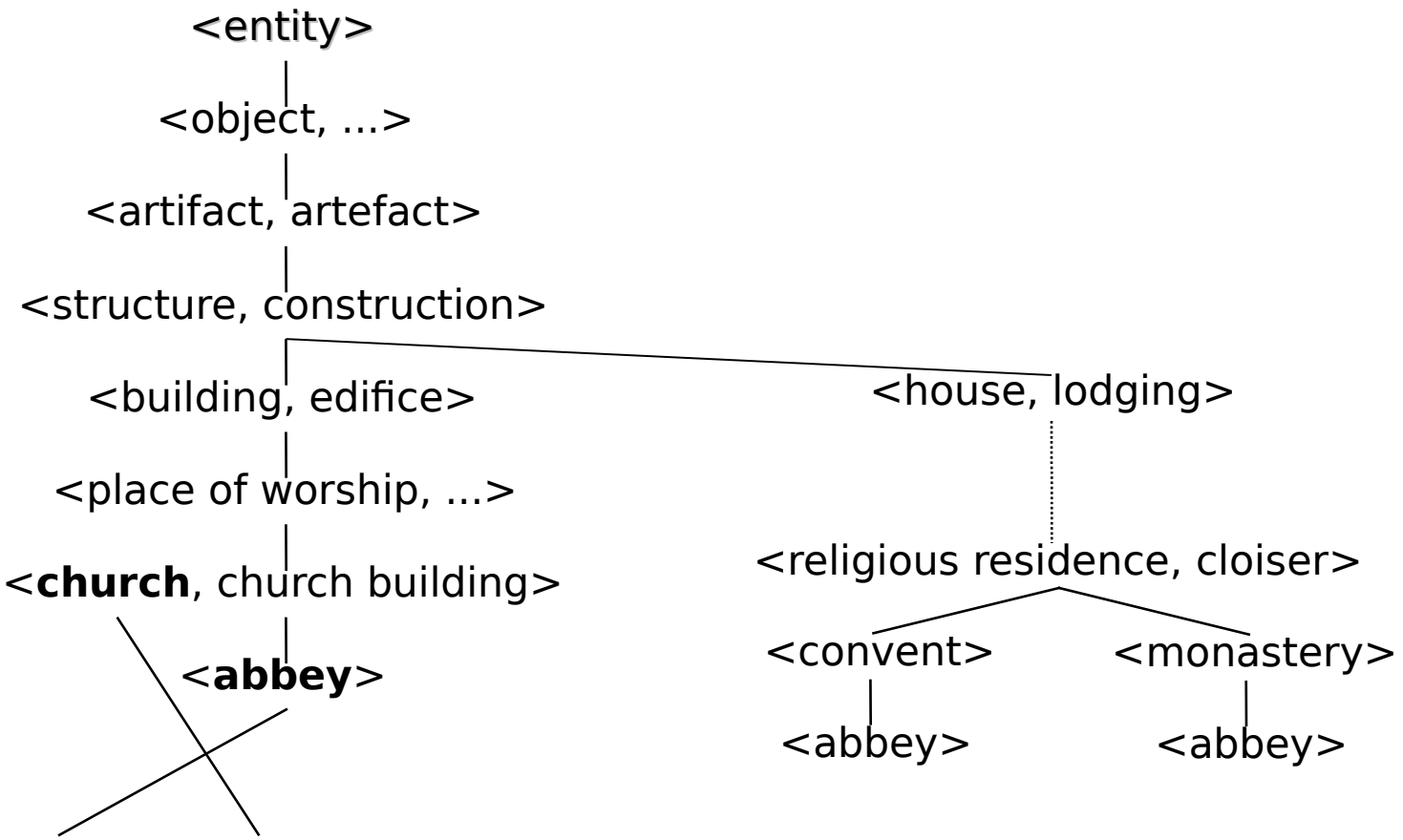
- length of the shortest path
- specificity of the concepts

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

- using WordNet
- Bilingual dictionary

Merge approach: Taxonomy Construction

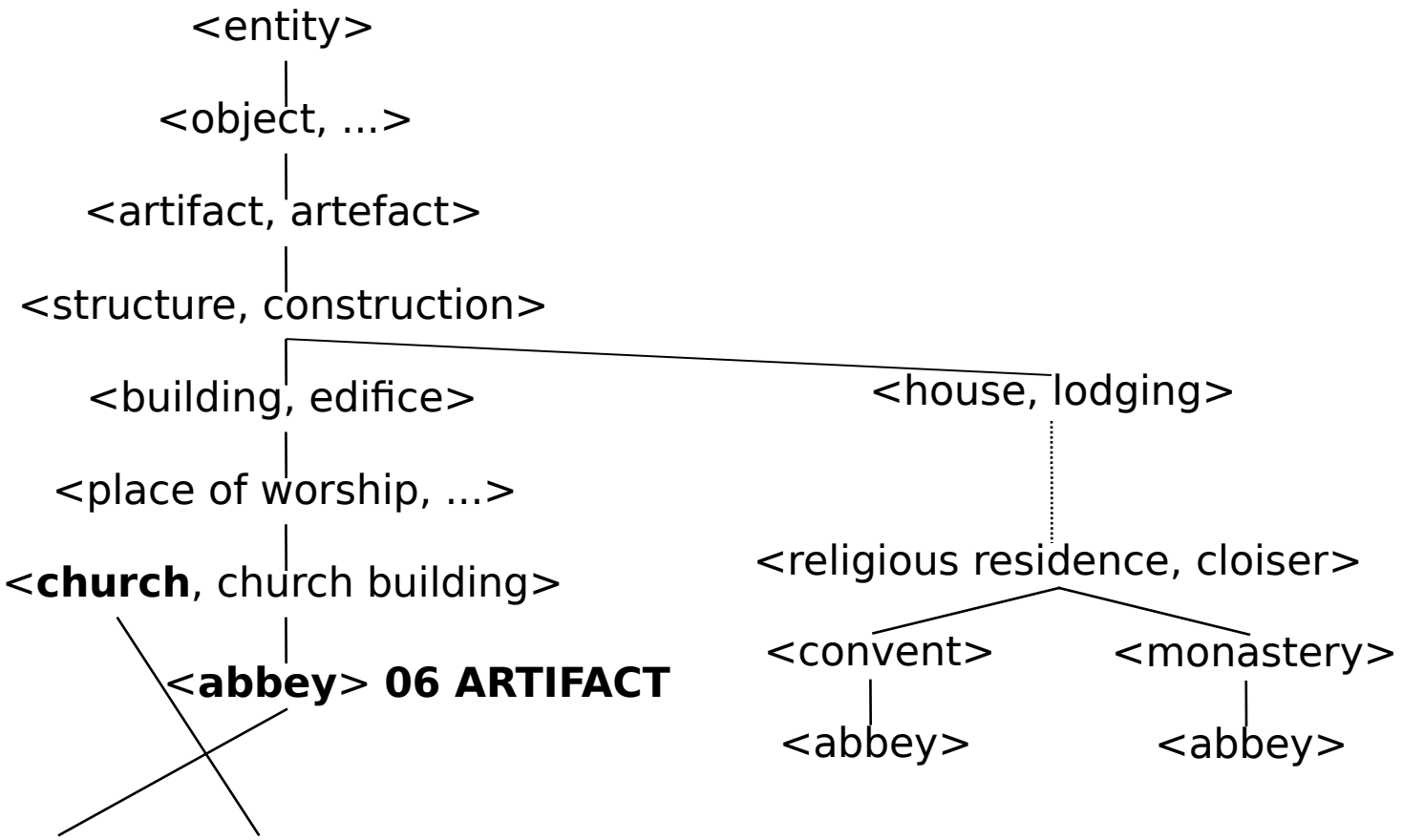
Step 1: Selection of the main top beginners



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

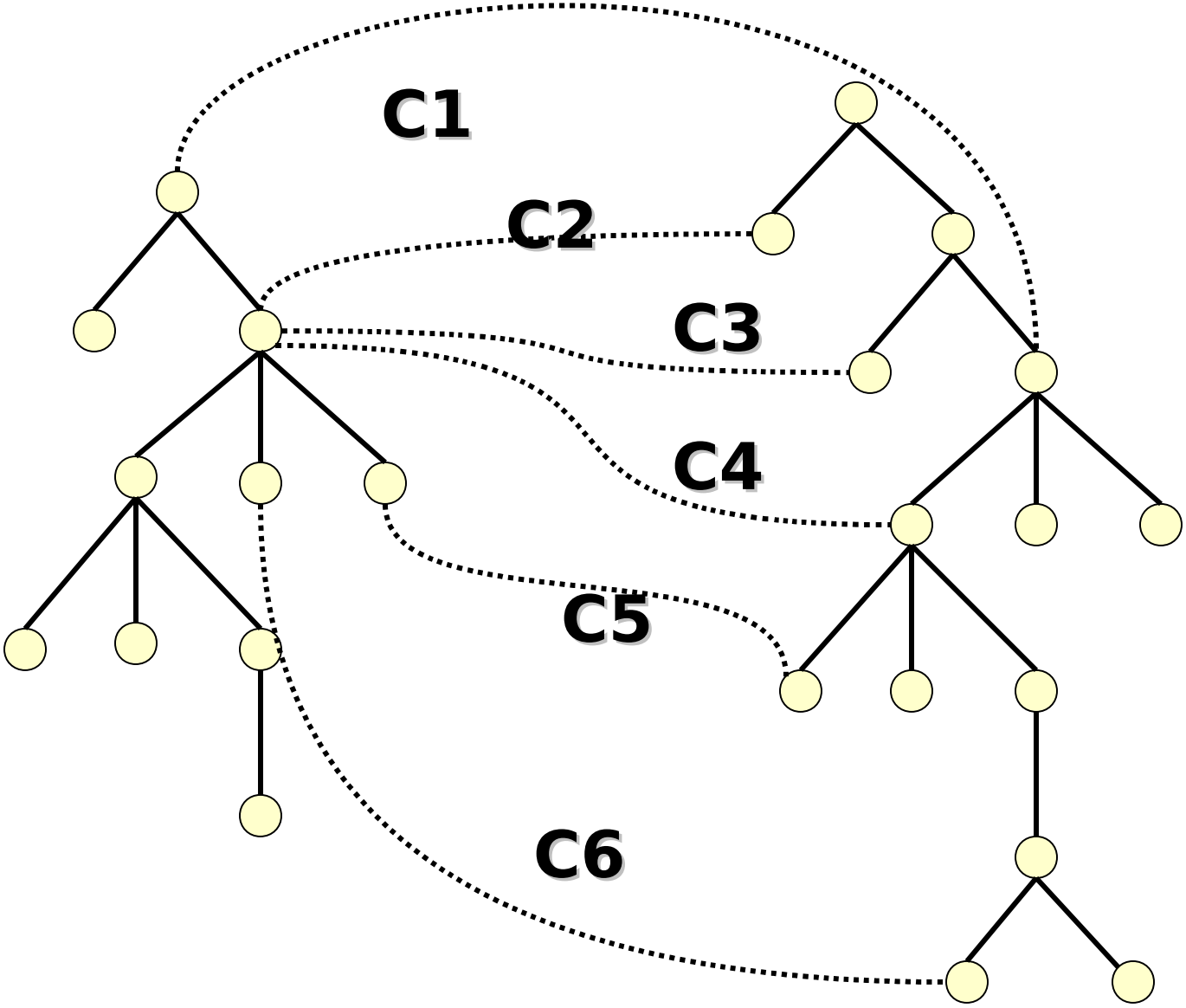
...

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

...

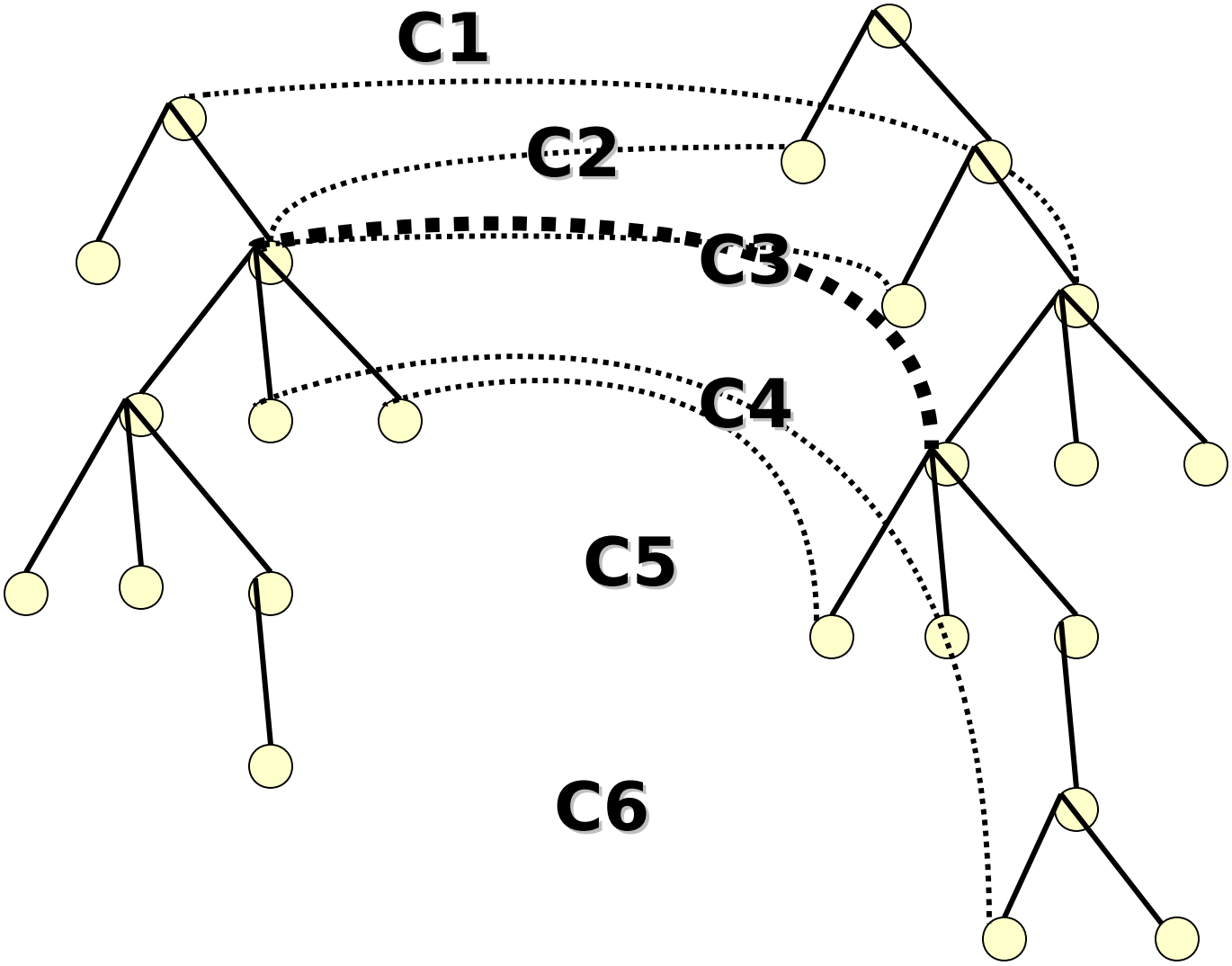
Merge approach: Mapping Taxonomies

Step 3: Creation of translation links_



Merge approach: Mapping Taxonomies

Step 3: Creation of translation links_



Merge approach: Mapping Taxonomies

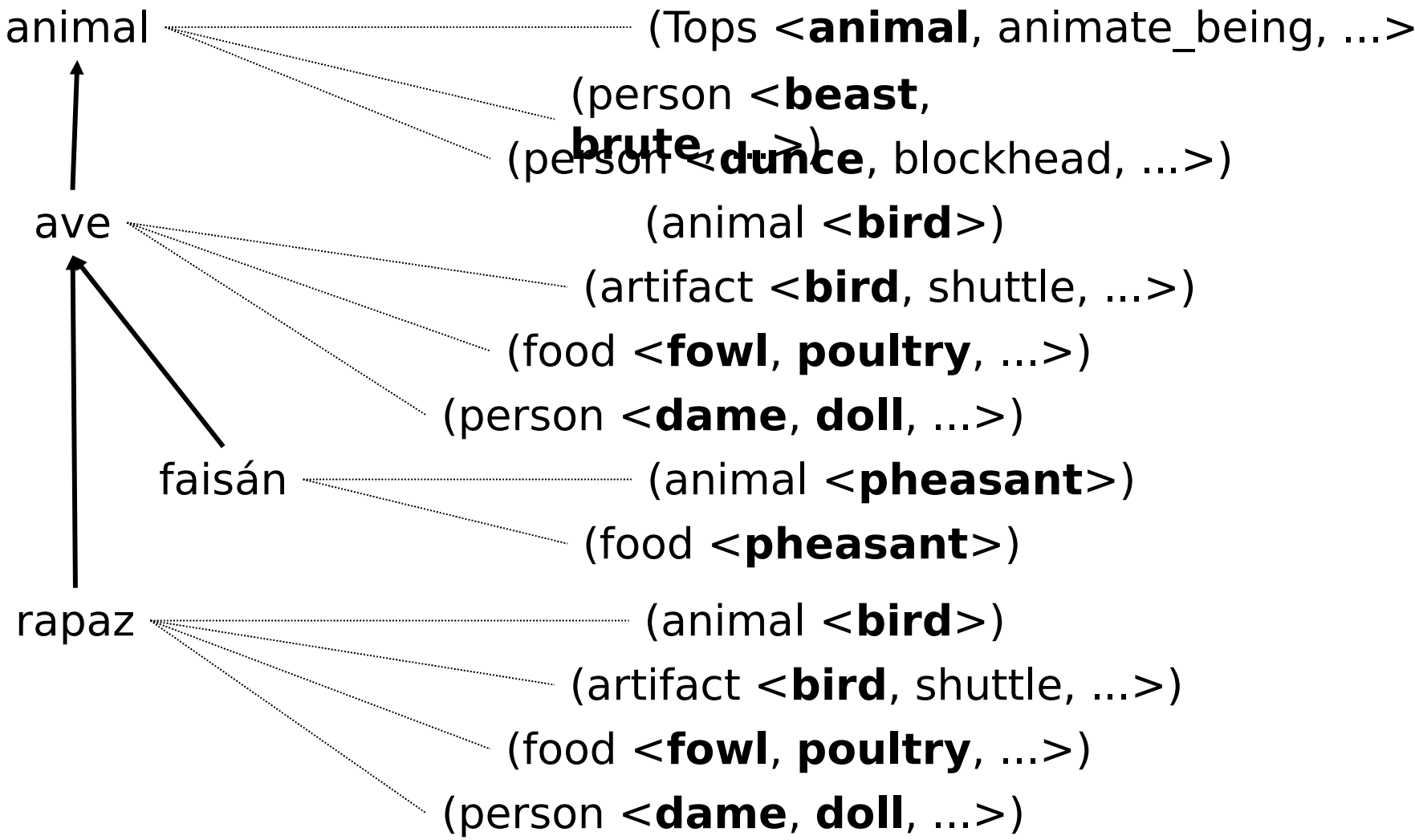
Step 3: Creation of translation links_

- 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.

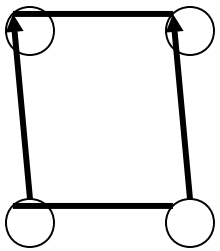
4) If a stopping/convergence is satisfied, stop, otherwise go to step 2.

Merge approach: Mapping Taxonomies

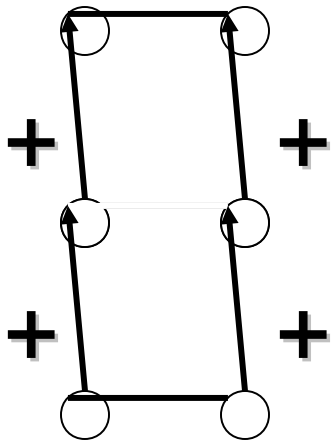
Step 3: Constraints_

- 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

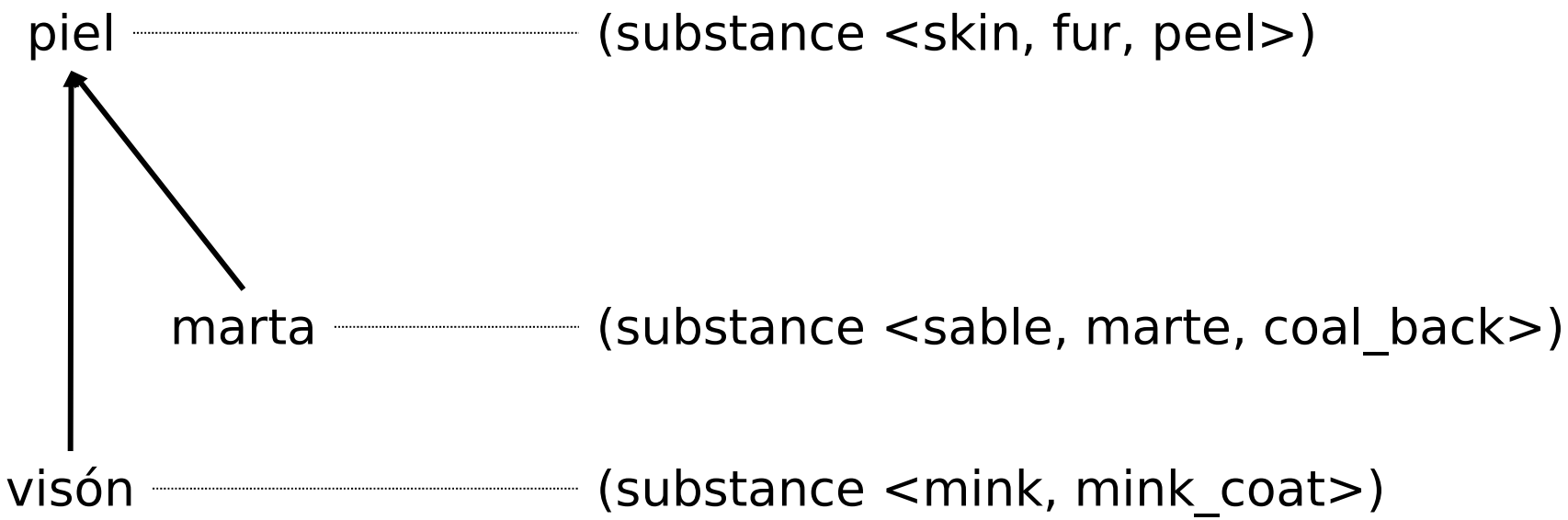
Step 3: Results_

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

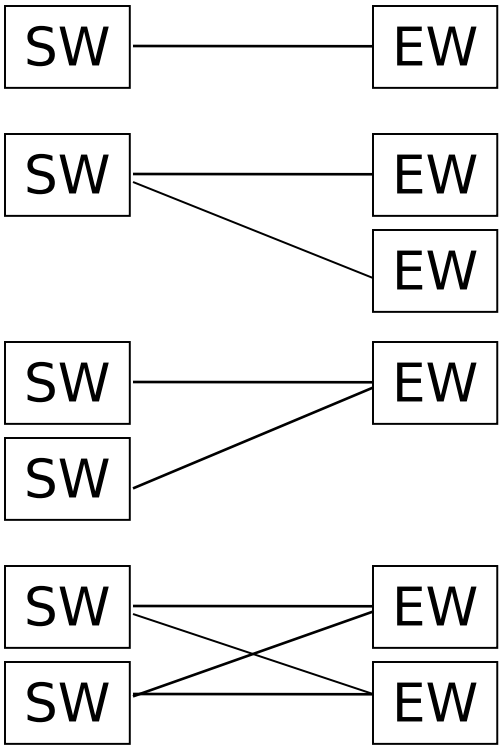
Methods_

- 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.

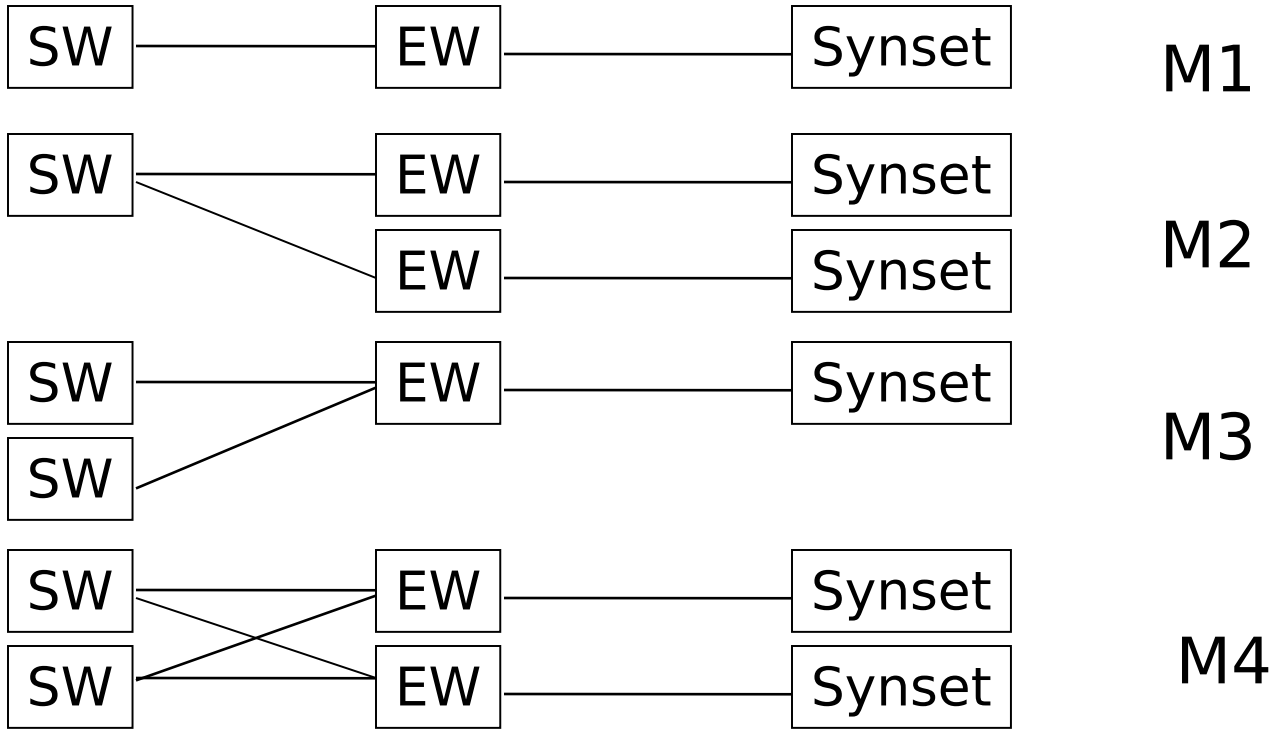


Expand approach

Class methods_

4 monosemous class methods:

- All English words involved are monosemous in WN

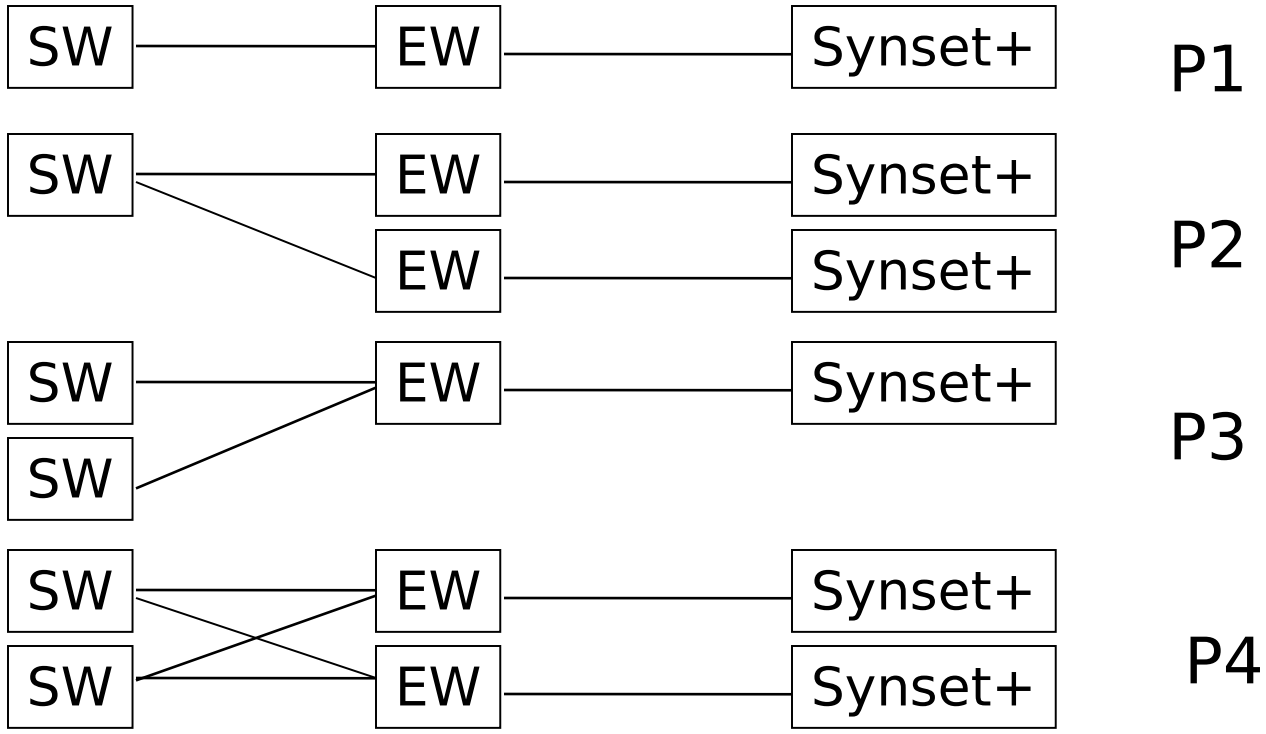


Expand approach

Class methods_

4 polysemous class methods:

- At least 1 English word involved is polysemous

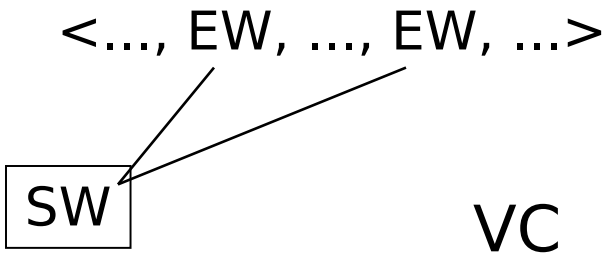


Expand approach

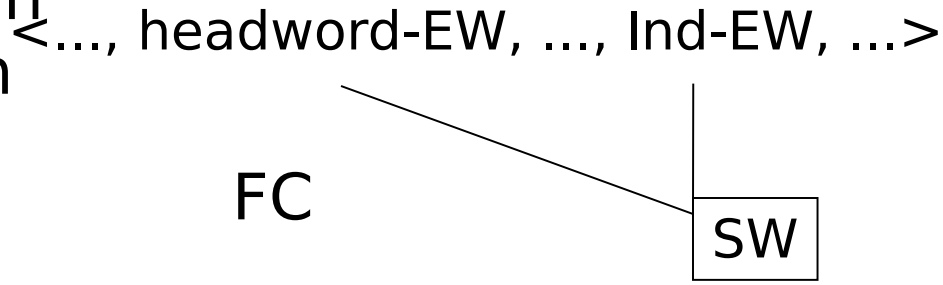
Class methods_

2 other class methods

- Variant criterion:
  two synonyms share
  a single SW



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



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

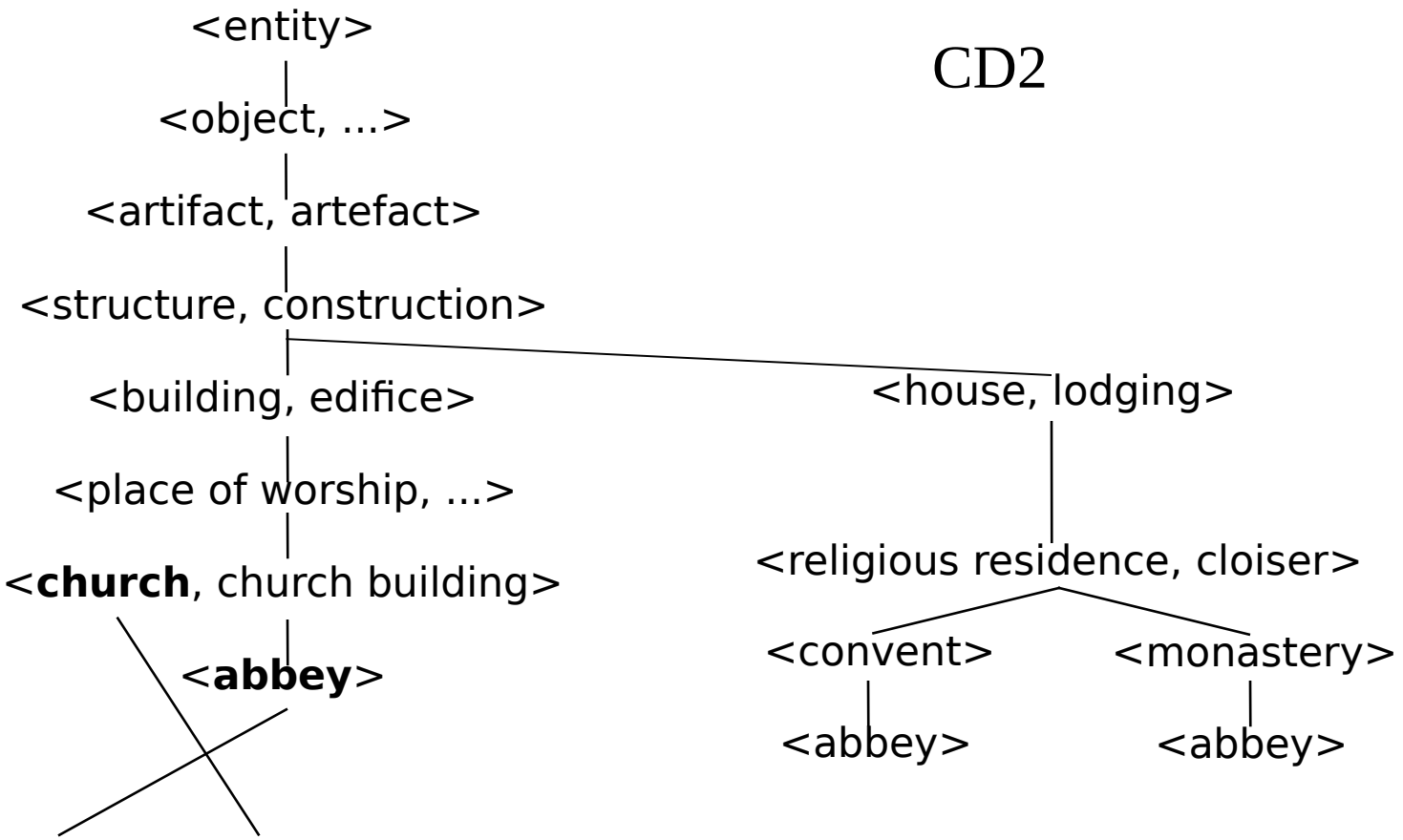
Expand approach

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

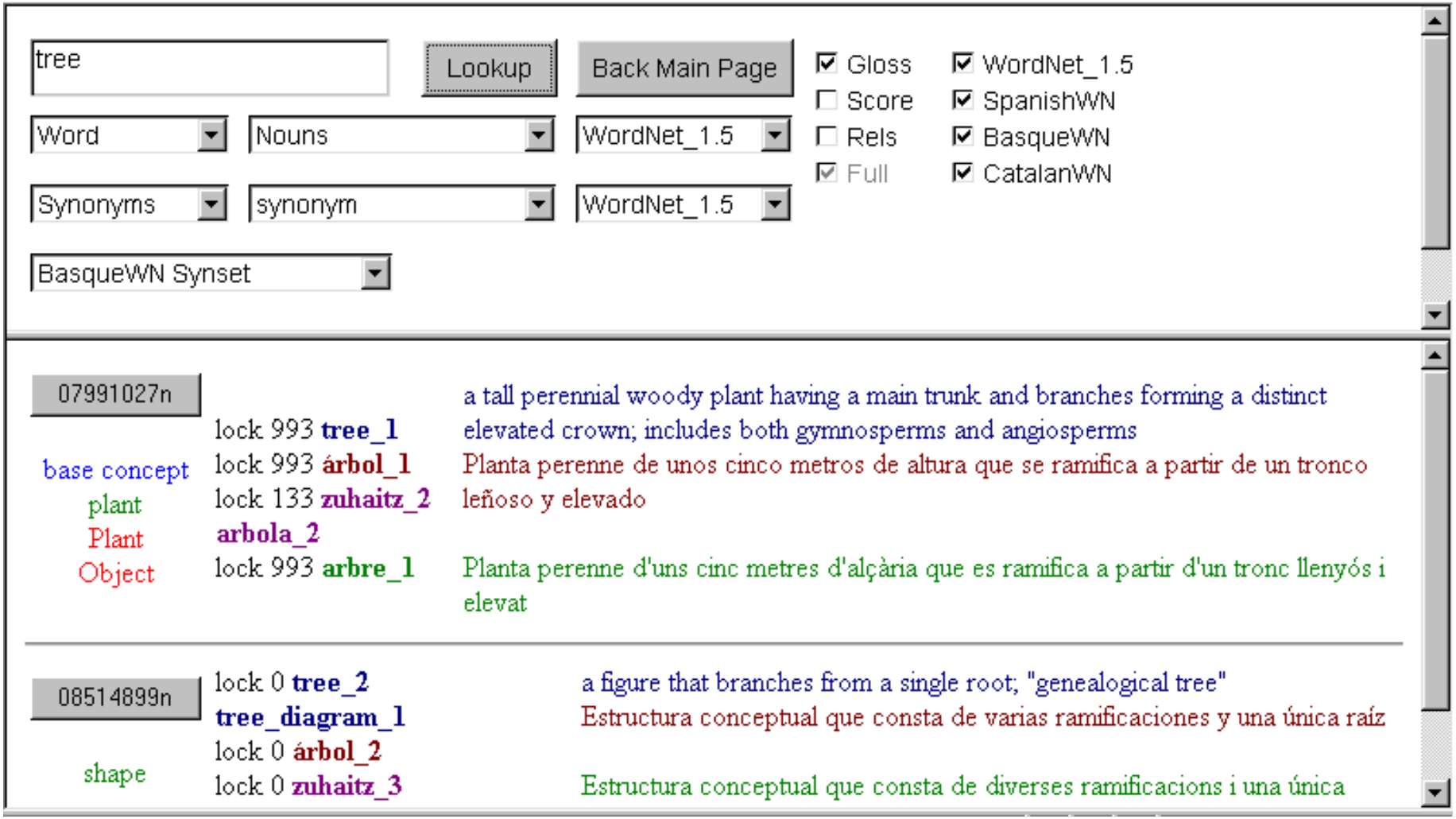
WNs	#links	#synsets	#word	#CS	#poly links
SpWN v0.0	10,982	7,131	8,396	87.4	1,777
Combination	7,244	5,852	3,939	85.6	2,075
SpWN v0.1	15,535	10,786	9,986	86.4	3,373
BasqueWN v0.1	41,107	23,486	22,166	>80.0	-

Outline_

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

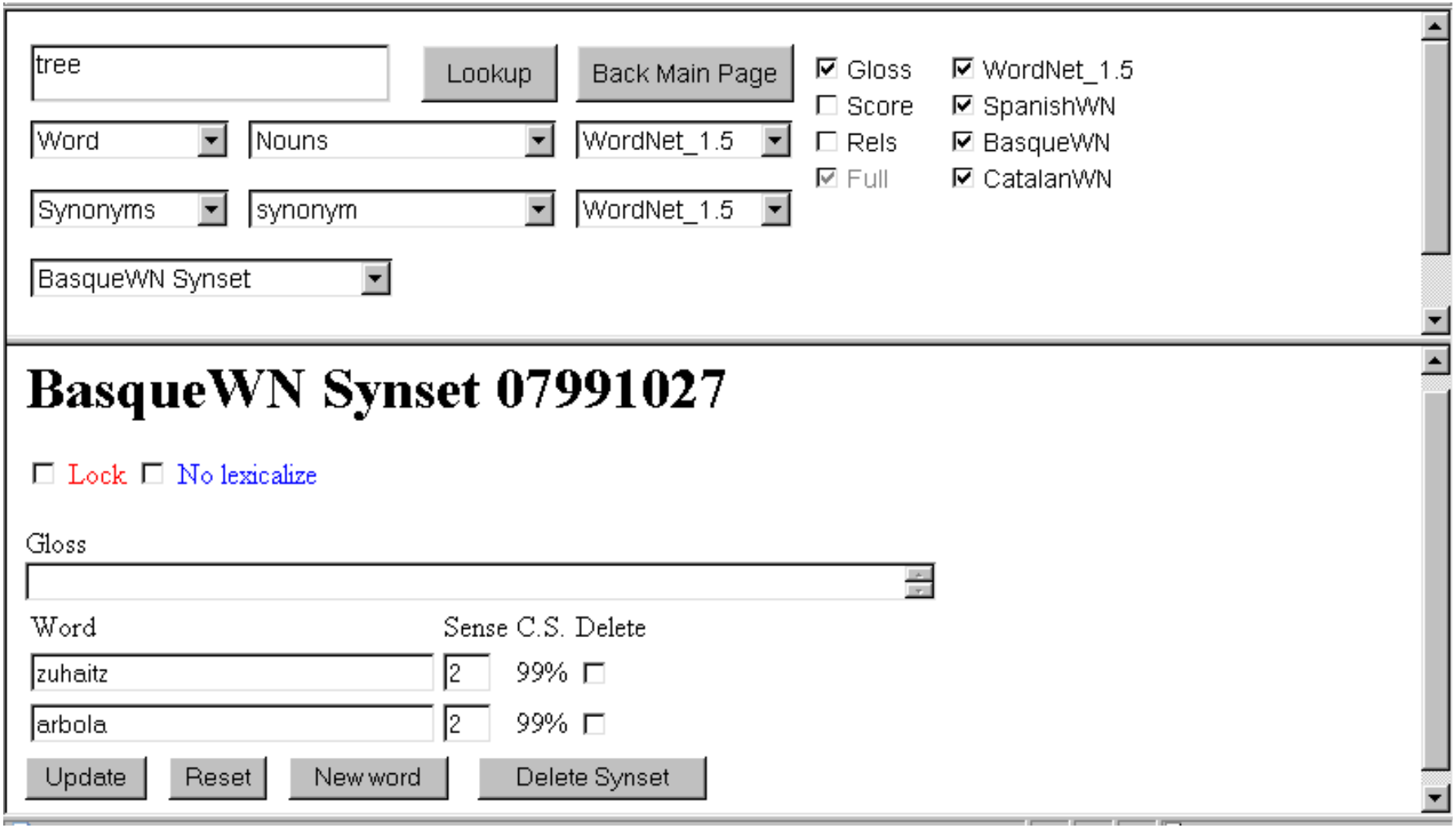
Building wordnets

Web EuroWordNet Interface_



Building wordnets

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

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

Building wordnets

German Rigau i Claramunt

german.rigau@ehu.es

IXA group

Departamento de Lenguajes y Sistemas Informáticos

UPV/EHU