The MEANING Multilingual Central Repository

J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, P. Vossen

January 27, 2004

http://www.lsi.upc.es/~nlp/meaning

Index

- The MEANING framework
- First Multilingual Central Repository
- First Uploading Process
  - SUMO
  - Selectional Preferences
  - Top Concept ontology
- First Porting Process
- The "pasta" example
- Conclusions and Future Work

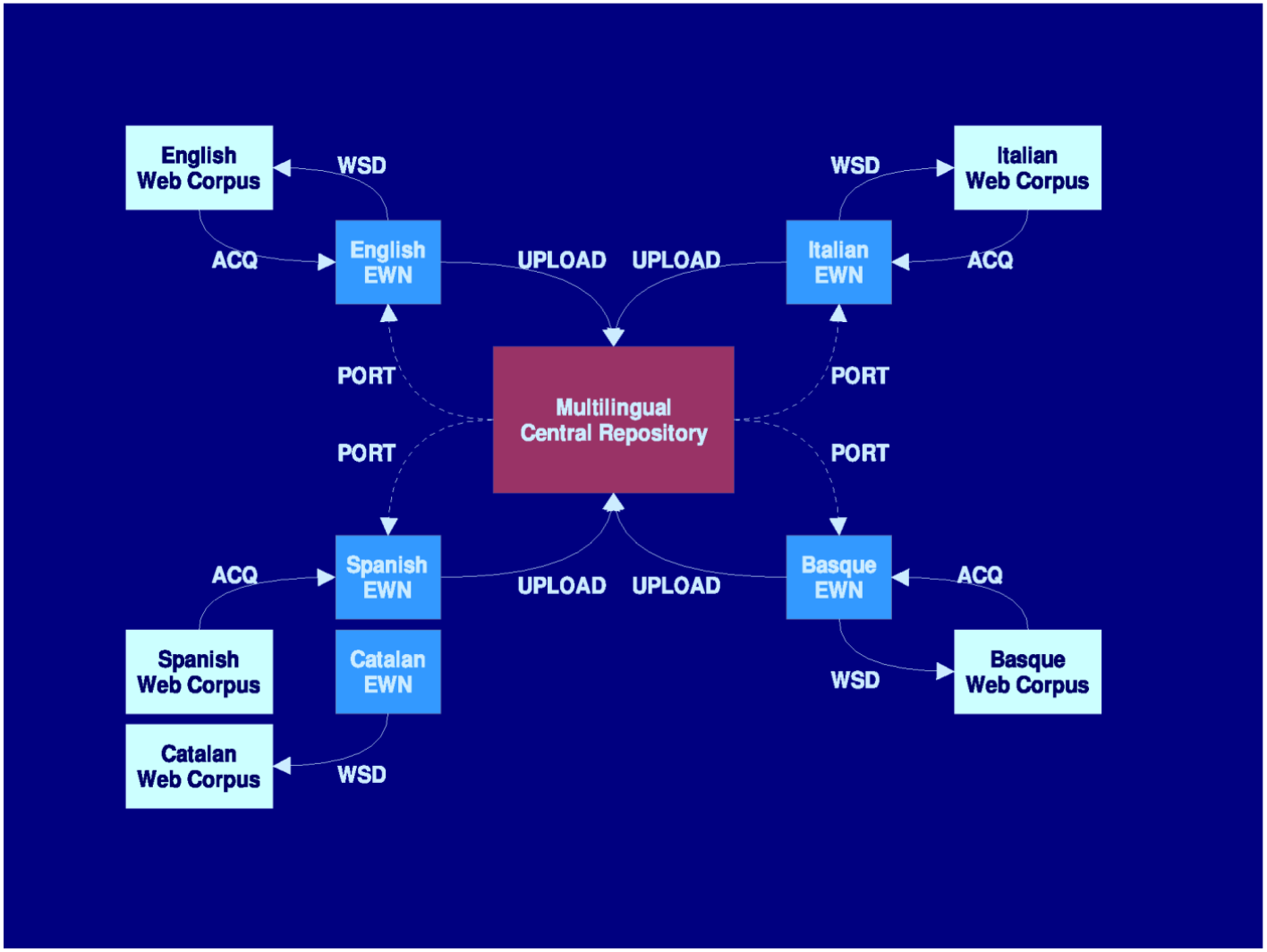
The MEANING Project

The MEANING project [Rigau *et al.* 02] [1]

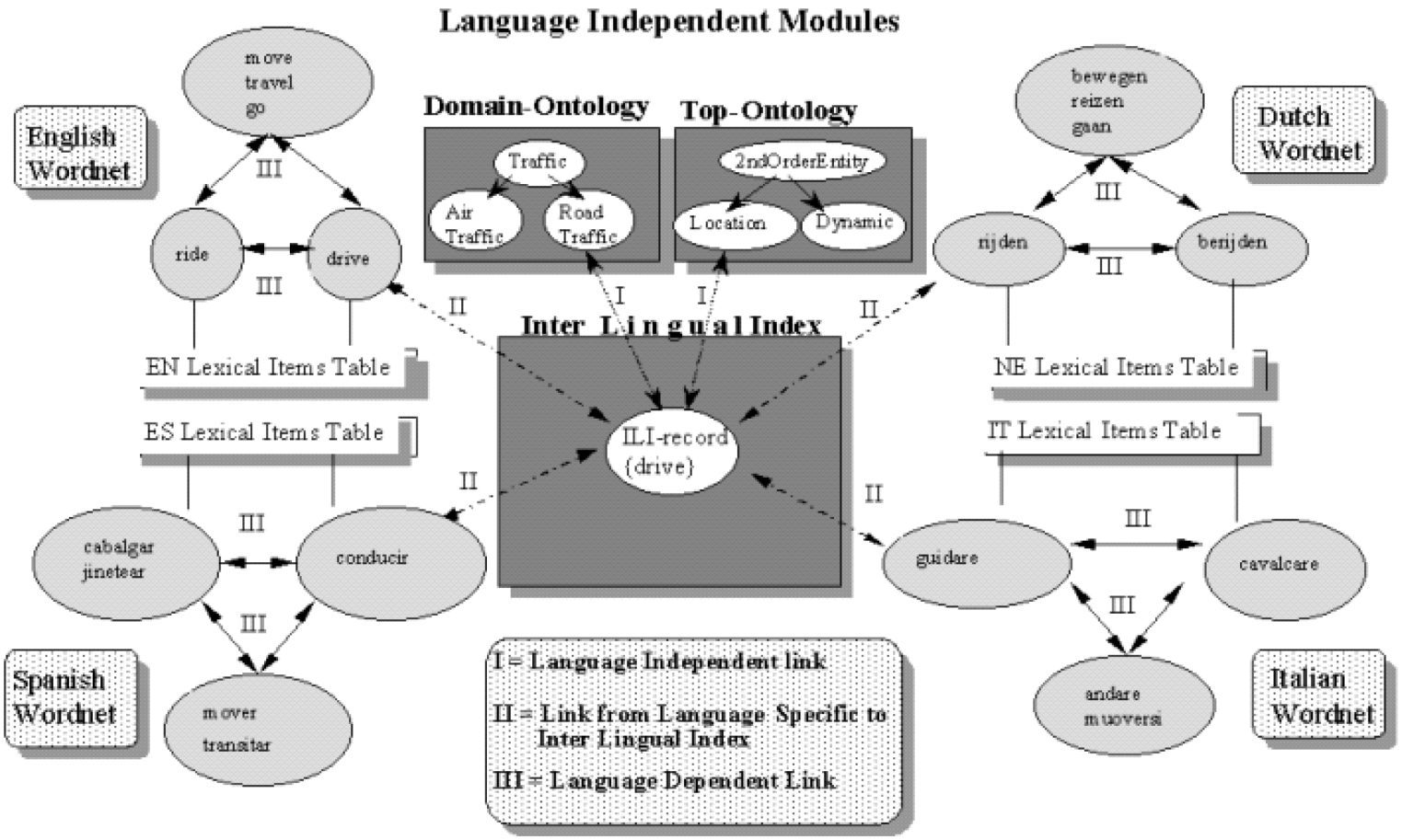
- Inter–dependency between WSD and knowledge acquisition
- Exploiting a multilingual architecture based on EuroWordNet [Vossen 98]
- Three consecutive cycles of large-scale WSD and acquisition

[1] http://www.lsi.upc.es/~nlp/meaning/meaning.html

MEANING Cycle



MCR Architecture



Content of the MCR

- ILI (WordNet 1.6)
  - Upgraded EuroWordNet Base Concepts
  - Upgraded EuroWordNet Top Concept ontology
  - MultiWordNet Domains
  - Suggested Upper Merged Ontology (SUMO)
- **Local wordnets**
  - English WordNet 1.5, 1.6, 1.7, 1.7.1
  - Basque, Catalan, Italian and Spanish wordnets
- **Large collections of semantic preferences**
  - Acquired from SemCor
  - Acquired from BNC
- **Instances**

Uploading

Main task: Transforming resources from wn1.5 to wn1.6.

- Project them using the mapping [Daudé *et al.* 99][2]
- Recovering Consistency (Base Concepts, Top Concept Ontology)
- Cross-Checking Resources

[2] http://www.lsi.upc.es/~nlp/tools/mapping.html

Uploading SUMO

- The Suggested Upper Merged Ontology (SUMO) [Niles & Pease 01] is an upper ontology proposed as starting point for the IEEE Standard Upper Ontology Working group.
- SUMO provides definitions for general purpose terms
- Currently only the SUMO labels and the SUMO ontology hyperonym relations are loaded into the MCR.
- We plan to cross–check the **Top Concept ontology** expansion and the **Domain ontology** with the SUMO ontology.

Uploading Selectional Preferences

- 390,549 weighted Selectional Preferences (SPs)
- A set [McCarthy 01] of weighted SPs was obtained by computing probability distributions over the WN1.6 noun hierarchy from parsing the BNC.
- The second set [Agirre & Martinez 02] was obtained from generalizations of grammatical relations extracted from Semcor.

Example: "pasta" (money sense) has the following preferences as object:
1.44 01576902-v {*raise#4*}, *0.45 01518840-v* {*take_in#5, collect#2*}
or *0.23 01565625-v* {*earn#2, garner#1*}

Uploading the Top Ontology

- TC aimed to enforce uniformity and compatibility of the different WNs
- EWN only performed a complete validation of its consistency among the Base Concepts.
- The properties assigned to the Base Concepts were not explicitly available from rest ILIs/synsets.
- Thus, we decide to perform an automatic expansion.

Uploading the Top Ontology

The **Top Concept ontology** has been uploaded in three steps:

1. Properties are assigned to WN1.6 synsets through the mapping.
2. Assign properties to the remaining WN1.6 Tops
3. The properties are propagated top–down through the WN hierarchy
4. The incompatibilities between properties block the propagation.

Expansion Problems

Problematic cases can be detected by cross-checking.

- **WordNet hierarchy**
  The classification of WN is not always consistent with the *Top Concept ontology*

- **Multiple inheritance**
  A synset inherits incompatible attributes from its ancestors

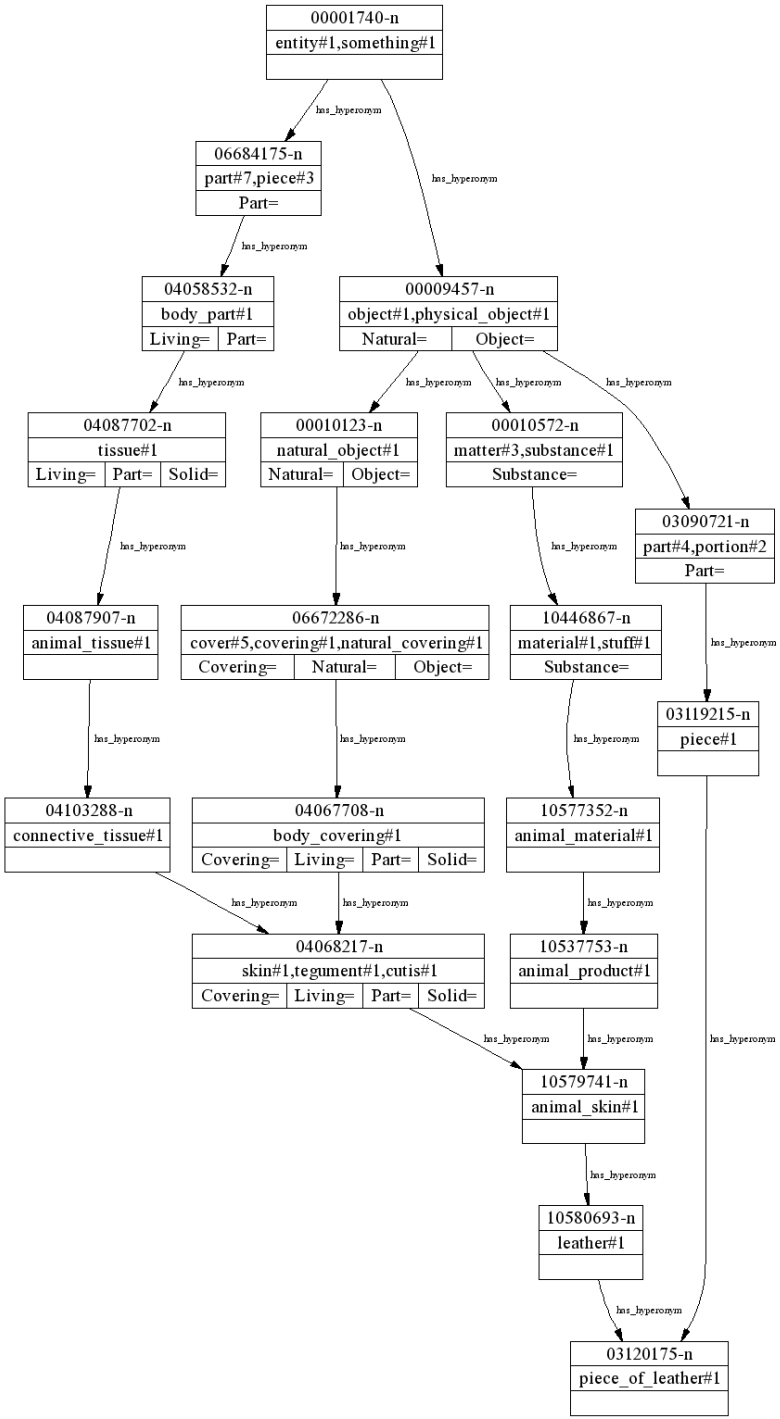
- **Cross-checking resources with different granularities**
  Human–Creature–Animal–Hominid

WordNet hierarchy

- **Animal vs Plant**

  00911639n phytoplankton_1 (SUMO.Plant+) and its direct descendant 00911809n planktonic_algae_1 (SUMO.Alga).

- **Liquid/Substance/Solid vs Object** body_part Object
  - Liquid *04195761n 105 liquid_body_substance_1 bodily_fluid_1 body_fluid_1*
  - Substances *4086329n 117 body_substance_1 the substance of the body*
  - Solid *06672286n covering_1 natural_covering_1 cover_5 any covering for the body or a body part*


00001740-n
entity#1,something#1
hns_hyperonym
06684175-n
part#7,piece#3
Part=
04058532-n
body_part#1
Living= Part=
00009457-n
object#1,physical_object#1
Natural= Object=
04087702-n
tissue#1
Living= Part= Solid=
00010123-n
natural_object#1
Natural= Object=
00010572-n
matter#3,substance#1
Substance=
03090721-n
part#4,portion#2
Part=
04087907-n
animal_tissue#1
06672286-n
cover#5,covering#1,natural_covering#1
Covering= Natural= Object=
10446867-n
material#1,stuff#1
Substance=
03119215-n
piece#1
04103288-n
connective_tissue#1
04067708-n
body_covering#1
Covering= Living= Part= Solid=
10577352-n
animal_material#1
04068217-n
skin#1,tegument#1,cutis#1
Covering= Living= Part= Solid=
10537753-n
animal_product#1
10579741-n
animal_skin#1
10580693-n
leather#1
03120175-n
piece_of_leather#1

Different Granularity

- **Human vs Animal**

  All the Hominids are considered animal by the semantic File)
  But Human by the top Concept ontology (SUMO Hominid+)

- **Human vs Creature**

  All the creatures (mainly the descendants of *imaginary_being_1* *imaginary_creature_1*) are classified as Human by the semantic File.

First Porting

- All the knowledge in MCR has been ported directly to the local WNs
- No extra semantic knowledge has been inferred in this process.
- All WordNets gain some knowledge
  - Spanish/Catalan/Basque EuWn gained Domains, SP
  - Italian MultiwordNet gained TC, EuWn-relations, SP
  - English WordNet gained EuWn-relations, (Domains, TC)

Porting Results I/II

	Spanish		English		Italian	
Relations	**UPLOAD**	**PORT0**	**UPLOAD**	**PORT0**	**UPLOAD**	**PORT0**
be_in_state	1,302	=	1,300	+2	364	+2
causes	240	=	224	+19	117	+15
near_antonym	7,444	=	7,449	+221	3,266	=
near_synonym	10,965	=	21,858	+19	4,887	+54
role	106	=	0	+106	0	+46
role_agent	516	=	0	+516	0	+227
role_instrument	291	=	0	+291	0	+151
role_location	83	=	0	+83	0	+39
role_patient	6	=	0	+6	0	+3
xpos_fuzzynym	37	=	0	+37	0	+23
xpos_near_synonym	319	=	0	+319	0	+181
Other relations	31,644	=	29,120	+2,627	9,541	+22
Total	53,272	=	59,951	+4,246	18,175	+763

PORT0 Main figures for Spanish, English and Italian

Porting Results II/II

	Spanish		English		Italian	
Relations	**UPLOAD**	**PORT0**	**UPLOAD**	**PORT0**	**UPLOAD**	**PORT0**
role_agent-semcor	0	+52,394	69,840	=	0	+41,910
role_agent-bnc	0	+67,109	95,065	=	0	+40,853
role_patient-semcor	0	+80,378	110,102	=	0	+41,910
role_patient-bnc	0	+79,443	115,102	=	0	+50,264
Role	0	+279,324	390,109	=	0	+174,937
Instances	0	+1,599	0	+2,198	791	=
Proper Nouns	1,806	=	17,842	=	2,161	=
Base Concepts	1,169	=	1,535	=	0	+935
Domains Links	0	+55,239	109,621	=	35,174	=
Domains Synsets	0	+48,053	96,067	=	30,607	=
Top Ontology Links	3,438	=	0	+4,148	0	+2,544
Top Ontology Synsets	1,290	=	0	+1,554	0	+946

PORT0 Main figures for Spanish, English and Italian

The "*pasta*" Example

Domain: chemistry-pure_science
Semantic File: 27-Substance
SUMO: Substance-SelfConnectedObject-Object-Physical-Entity

Top Concept ontology
Natural-Origin-1stOrderEntity
Substance-Form-1stOrderEntity

pasta#n#7 10541786-n
paste#1
gloss: any mixture of a soft and malleable consistency

The "*pasta*" Example

Domain: money-economy-soc.science
Semantic File: 21-MONEY
SUMO:
CurrencyMeasure-ConstantQuantity-PhysicalQuantity-Quantity-Abstract-Entity
Top Concept ontology
Artifact-Origin-1stOrderEntity
Function-1stOrderEntity
MoneyRepresentation-Representation-Function-1stOrderEntity

pasta#n#6 09640280-n
dough#2,bread#2,loot#2, ...
gloss: informal terms for money

The "*pasta*" Example

Domain: gastronomy-alimentation-applied_science
Semantic File: 13-FOOD **SUMO:** Food-...
Top Concept ontology Comestible-Function-1stOrderEntity
Substance-Form-1stOrderEntity

pasta#n#4 05886080-n
spread#5,paste#3
gloss: a tasty mixture to be spread on bread or crackers

pasta#n#1 05671312-n
pastry#1,pastry_dough#1
gloss: a dough of flour and water and shortening

pasta#n#3 05739733-n
pasta#1,alimentary_paste#1
gloss: shaped and dried dough made from flour and water & sometimes egg

pasta#n#5 05889686-n
dough#1
gloss: a dough of flour and water and shortenings

pasta#n#2 05671439-n *pie_crust#1,pie_shell#1*
gloss: pastry used to hold pie fillings

Conclusions

- MCR v0 integrates in a EWN framework (upgraded Base Concepts and Top Concept ontology and MWND) five local WNs (with four English WN versions) with hundreds of thousands of new semantic relations, instances and properties fully expanded.

- All WNs gain some kind of knowledge from other WNs (porting process).

- We intend the MCR to be a natural multilingual large-scale knowledge resource.

- A full range of new possibilities appears for improving both Acquisition and WSD tasks in the next two MEANING rounds.

Future Work

- Upload more resources (wn2.0, eXtended Wordnet)
- Maybe including language dependent data, such as syntactic information, subcategorization frames, diathesis alternations ...
- *porting process*, to investigate inference mechanisms to infer new explicit relations and knowledge (regular polysemy, nominalizations, etc).
- Investigate/check the correctness the semantic knowledge ported across languages.

Thanks for your attention

http://www.lsi.upc.es/~nlp/meaning

This research has been partially funded by the Spanish Research Department (HERMES TIC2000-0335-C03-02) and by the European Commission (MEANING IST-2001-34460).

Bibliography

References

[Agirre & Martinez 02] (Agirre & Martinez 02) E. Agirre and D. Martinez. Integrating selectional preferences in wordnet. In *Proceedings of the first International WordNet Conference in Mysore*, India, 21-25 January 2002.

[Daudé *et al.* 99] (Daudé *et al.* 99) J. Daudé, L. Padró, and G. Rigau. Mapping Multilingual Hierarchies Using Relaxation Labeling. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99)*, Maryland, US, 1999.

[McCarthy 01] (McCarthy 01) D. McCarthy. *Lexical Acqusition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Unpublished PhD thesis, University of Sussex, 2001.

[Niles & Pease 01] (Niles & Pease 01) I. Niles and A. Pease. Towards a standard upper ontology. In *In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 17–19. Chris Welty and Barry Smith, eds, 2001.

[Rigau *et al.* 02] (Rigau *et al.* 02) G. Rigau, B. Magnini, E. Agirre, P. Vossen, and J. Carroll. Meaning: A roadmap to knowledge technologies. In *Proceedings of COLLING Workshop 'A Roadmap for Computational Linguistics'*, Taipei, Taiwan, 2002.

[Vossen 98] (Vossen 98) P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers , 1998.