Starting Up the Multilingual Central Repository*

Jordi Atserias
TALP Research Center
Jordi Girona Salgado, 1-3.
08034 Barcelona
{batalla}@talp.upc.es

German Rigau
IXA Group
Euskal Herriko Unibertsitatea
Donostia.
{rigau}@si.ehu.es

Luís Villarejo
TALP Research Center
Jordi Girona Salgado, 1-3.
08034 Barcelona
{luisv}@talp.upc.es

Resumen: Este artículo presenta el diseño inicial del "Multilingual Central Repository". La primera versión del MCR integra, siguiendo en el marco de EuroWordNet, cinco wordnets locales (incluyendo tres versiones del WordNet de Princeton), la Top Ontology the EuroWordNet, los dominios de MultiWordNet y cientos de miles de nuevas relaciones semánticas y propiedades adquiridas automáticamente de corpus. De hecho, el MCR resultante constituye la más grande y rica base de conocimiento multilingüe nunca construida.
Palabras clave: Wordnet, EuroWordnet, MultiWordnet, Adquisición Multilingüe

Abstract: This paper describes the initial design of the Multilingual Central Repository. The first version of the MCR integrates into the same EuroWordNet framework, five local wordnets (including three versions of the English WordNet from Princeton), the EuroWordNet Top Ontology, MultiWordNet Domains, and hundreds of thousand of new semantic relations and properties automatically acquired from corpora. In fact, the resulting MCR is going to constitute the largest and richest multilingual lexical–knowledge ever build.
Keywords: Wordnet, EuroWordnet, MultiWordnet, Multilingual Acquisition

1 Introduction

Knowledge Technologies aim to provide meaning to the petabytes of information content our societies will generate in the near future. Information and knowledge management systems need to evolve accordingly, to enable the next generation of intelligent open domain Human Language Technologies (HLT) that will deal with the growing potential of the knowledge-rich and multilingual society.

In order to develop a trustable semantic web infrastructure and a multilingual ontology framework to support knowledge management, a wide range of techniques are required to progressively automate the knowledge lifecycle. In particular, this involves extracting high-level meaning from large collections of content data and its representation and management in a common knowledge base.

Even now, building large and rich knowledge bases takes a great deal of expensive manual effort; this has severely hampered Knowledge-Technologies and HLT application development. For example, dozens of person-years have been invest into the development of wordnets (Fellbaum, 1998) for various languages (Atserias et al., 1997; Benítez et al., 1998; Bentivogli, Pianta, and Girardi, 2002), but the data in these resources is still not sufficiently rich to support advanced concept-based HLT applications directly. Furthermore, resources produced by introspection usually fail to register what really occurs in texts. Applications will not scale up to working in the open domain without more detailed and rich general-purpose, which should perhaps include domain-specific linguistic knowledge.

The MEANING project (Rigau et al., 2002) [1] identifies two complementary and intermediate tasks which are crucial in order to enable the next generation of intelligent open domain HLT application systems: Word Sense Disambiguation (WSD) and large-scale enrichment of Lexical Knowledge Bases.

The advance in these two areas will allow for large-scale acquisition of shallow meaning from texts, in the form of relations among concepts. WSD provides the technology to convert relations between words into rela-

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

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

tions between concepts. Rich and large-scale Lexical Knowledge Bases and formalized Ontologies will be the repositories of all relations and other linguistic knowledge.

However, progress is difficult due to the following interdependence:

- In order to achieve accurate WSD, we need far more linguistic and semantic knowledge than is available in current lexical knowledge bases (e.g. current wordnets).
- In order to enrich existing Knowledge Bases we need to acquire information from corpora, which have been accurately tagged with word senses.

MEANING proposes an innovative bootstrapping process to deal with this inter–dependency between WSD and knowledge acquisition exploiting a multilingual architecture based on the EuroWordNet (Vossen, 1998).

MEANING plans to perform three consecutive cycles of large-scale WSD and acquisition processes in five European languages including Basque, Catalan, English, Italian and Spanish. As languages realize the meaning in different ways, some semantic relations that can be difficult to acquire in one language can be easy to capture in other languages.

The knowledge acquired for each language during the three consecutive cycles will be consistently upload and integrated into the respective local wordnets, and then ported and distributed across the rest of wordnets, balancing resources and technological advances across languages.

The Multilingual Central Repository (MCR) will grant the consistency and integrity of all the semantic knowledge produced by MEANING.

This paper describes the first version of the Multilingual Central Repository (MCR). After this introduction, section 2 presents the MCR structure, content and associated software tools. While section 3 describes the first uploading process, section 4 is devoted to the porting process. Finally, section 5 presents an example of the current content of the MCR and section 6 draws some conclusions and future work.

2 Multilingual Central Repository

The MCR acts as a multilingual interface for integrating and distributing all the knowledge acquired in MEANING.

2.1 MCR structure

The MCR follows the model proposed by the EuroWordNet project. EuroWordNet is a multilingual lexical database with wordnets for several European languages, which are structured as the Princeton WordNet (Fellbaum, 1998).

The Princeton WordNet contains information about nouns, verbs, adjectives and adverbs in English and is organized around the notion of a *synset*. A synset is a set of words with the same part-of-speech that can be interchanged in a certain context. For example, <*car, auto, automobile, machine, motorcar*> form a synset because they can be used to refer to the same concept. A synset is often further described by a gloss: "4-wheeled; usually propelled by an internal combustion engine". Finally, synsets can be related to each other by semantic relations, such as hyponymy (between specific and more general concepts), meronymy (between parts and wholes), cause, etc.

The EuroWordNet architecture includes the **Inter–Lingual–Index** (ILI), a **Domain Ontology** and a **Top Concept Ontology** (Vossen, 1998). The ILI consists of a list of so-called ILI-records which interconnets word meanings in the local wordnets, (possibly) to one or more Top Concepts and (possibly) to domains.

That is, the wordnets are linked to the ILI. 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 connected.

The ILI is enhanced, enriched and structured by two separate ontologies, which may be linked to ILI records:

- the Top Concept ontology, which is a hierarchy of language-independent concepts, reflecting important semantic distinctions, e.g. Object and Substance, Location, Dynamic and Static;
- a hierarchy of domain labels, which are knowledge structures grouping meanings in terms of topics or scripts, e.g. Transport, Sports, Medicine, Gastronomy;

Both the Ontological properties and the Domain Labels can be transferred via the equivalence relations of the ILI-records to the local word meanings. The Top Concepts Location and Dynamic are for example directly linked to the ILI-record *drive* and therefore indirectly also apply to all language-specific concepts related to this ILI-record. Via the local wordnet relations, the Top Concept can be further inherited by through other related language-specific concepts.

The main purpose of the **Top Ontology** is to provide a common framework for the most important concepts in all the wordnets. It consists of 63 basic semantic distinctions that classify a set of ILI-records connected to WordNet representing the most important concepts in the different wordnets. EuroWordNet distinguish at the first level 3 types of entities:

- **1stOrderEntity** Any concrete entity (publicly) perceivable by the senses and located at any point in time, in a three-dimensional space.
- **2ndOrderEntity** Any Static Situation (property, relation) or Dynamic Situation, which cannot be grasped, heard, seen, felt as an independent physical thing. They can be located in time and occur or take place rather than exist.
- **3rdOrderEntity** Any unobservable proposition which exists independently of time and space. They can be true or false rather than real. They can be asserted or denied, remembered or forgotten.

The **Domain Hierarchy** groups concepts in a different way, based on scripts rather than classification. The MCR uses the MultiWordNet Domains (Magnini and Cavaglià, 2000) which were partially derived from the Dewey Decimal Classification [2]. WordNet Domains is a hierarchy of 165 Domain Labels associated to WordNet 1.6 synsets.

Information brought by Domain Labels is complementary to what is already in WordNet. First of all a Domain Labels may include synsets of different syntactic categories: for instance MEDICINE groups together senses from nouns, such as *doctor* and *hospital*, and from verbs such as *to operate*. Second, a Domain Label may also contain senses from different WordNet subhierarchies (i.e. deriving from different *unique beginners* or from different *lexicographer files*. For example, the SPORT contains senses such as athlete, deriving from life form, game equipment, from physical object, sport from act, and playing field, from location.

[2] `http://www.oclc.org/dewey`

The knowledge acquired locally is uploaded and ported across the rest of languages via the EuroWordNet ILI, maintaining the compatibility among them.

In that way, the ILI structure (including the Top Ontology and the Domain Hierarchy) will act as a natural backbone to transfer the different knowledge acquired from each local wordnet to the rest of wordnets.

2.2 MCR content

The first version of the MCR includes only conceptual knowledge (semantic information), that is, local wordnets (only semantic information including lexical units connected to synsets like variants or phrasets) and +60 types of EuroWordNet semantic relations connecting synsets.

Only complete disambiguated semantic relations between synsets will be acquired, upload and ported across local wordnets. Also, if necessary, the relations acquired can be underspecified. An explicit hierarchy of relations must be provided. For instance:

INVOLVED ← INVOLVED-PATIENT ← INVOLVED-RESULT

Although these relations are not completely disambiguated, they will be uploaded and ported to be ready useful for other acquisition processes and languages. For instance, consider the following relation <gain> INVOLVED <money> captured as typical object. Although, this relation may be further refined into <gain> INVOLVED-PATIENT <money> in posterior cycles, other processes (like those that locate sense examples from large text collections) can take profit from a ported relation <ganar> INVOLVED <dinero>. For instance, (Leacock, Chodorow, and Miller, 1998), (Mihalcea and Moldovan, 1999) and (Agirre and Martinez, 2000) automatically generate arbitrarily large corpora for unsupervised WSD training, using the knowledge contained in WordNet to formulate search engine queries over large text collections or the Web.

The first version of the Mcr includes:

- Ili
  - WordNet 1.6
  - EuroWordNet Base Concepts
  - EuroWordNet Top Ontology
  - MultiWordNet Domains
- Local wordnets
  - English WordNet 1.5, 1.6, 1.7.1
  - Basque, Catalan, Italian and Spanish wordnets
- Large collections of semantic preferences
  - Acquired from SemCor
  - Acquired from BNC
- Instances
  - Named Entities

2.3 Mcr Access

The Mcr provides a web interface to the database based on Web EuroWordNet Interface (WEI)[3]. Three different APIs have been developed to provide flexible access to the Mcr: first, a SOAP API to allow any remote user to interact with the Mcr, an extension of wnQuery perl API to the Mcr and a C++ API for high performance software.

3 *Uploading Process*

To upload correctly all this knowledge into a single multilingual repository a very complex process must be performed. Once finished the first part of the upload the data released by the different partners (just checking errors and inconsistencies), a more complex second part must be performed. This second part consist of the correct integration of every piece of information into the Mcr. That is, linking correctly all this knowledge to the Ili. This second part involves a complex cross checking validation process and usualy a complex expansion/inference of large amounts of semantic properties and relations through the semantic structure.

Initially most of the knowledge to be uploaded into the Mcr has been derived from WordNet 1.6 (selectional preferences from SemCor and BNC) and the Italian WordNet and the MultiWordNet Domains, both developed at IRST are using WordNet 1.6 as Ili (Bentivogli, Pianta, and Girardi, 2002; Magnini and Cavaglià, 2000). Thus, Mcr uses Princeton WordNet 1.6 as Ili. This option also minimises side effects with other European initiatives (Balkanet, EuroTerm, etc.) and wordnet developments around Global WordNet Association. However, the Ili for Spanish, Catalan and Basque wordnets was WordNet 1.5 (Atserias et al., 1997; Benítez et al., 1998), as well as the EuroWordNet Top Ontology and the associated Base Concepts.

3.1 Uploading local wordnets based on WordNet1.5

Although the technology to provide compatibility across wordnets exits (J. Daudé, 1999; J. Daudé, 2000; Daudé, Padró, and Rigau, 2001)[4], new research is needed for uploading and porting the various types of knowledge across languages, and new ways to test and valididate the ported knowledge in the target languages.

Uploading local wordnets based on WordNet1.5 to the Mcr is a complex process, because between differents wordnet versions, synsets can be splited (1:N), joined (N:1), added (0:1) or deleted (1:0) throught mapping. Thus, even if we perform manual checking of these connections, for those remaining cases of spliting or joining synsets the information inside the synsets should be modified accordingly.

The whole process of the porting wordnets using Ili based on WordNet1.5 to the new Ili based on WordNet1.6 consist of:

1. For all splited synsets, all information of synset 1.5, including variants, is copied to each of the equivalent synsets in 1.6
2. For all joined synsets, all information of synsets 1.5, including variants, is copied to the equivalent synset in 1.6
3. Manual revision to validate the splitted and joined synsets.

3.2 Conceptual coverage

Table 1 shows the overlapping for nouns, verbs and adjectives between each wordnet pair.

At a synset level, noun overlapping is quite high and homogeneous between wordnet pairs. The maximum overlapping occurs between English and Spanish (29,502) and the lowest between Italian and Catalan (14,462).

[3] `http://nipadio.lsi.upc.es/wei.html`

[4] `http://www.lsi.upc.es/ñlp/meaning/meaning.html`

NOUN	en16	spwn	itwn	cawn	bawn
en16	66,025	29,502	22,634	26,197	22,722
spwn	-	31,241	16,355	24,582	19,020
itwn	-	-	25,402	14,462	15,000
cawn	-	-	-	17,936	16,763
bawn	-	-	-	-	24,461
VERB	**en16**	**spwn**	**itwn**	**cawn**	**bawn**
en16	12,127	7,464	4,281	4,952	3,138
spwn	-	7,563	3,071	3,789	2,809
itwn	-	-	4,312	2,358	1,844
cawn	-	-	-	5,051	2,333
bawn	-	-	-	-	3,237
ADJ	**en16**	**spwn**	**itwn**	**cawn**	**bawn**
en16	17,915	11,087	2,658	4,028	0
spwn	-	11,135	1,700	3,932	0
itwn	-	-	2,686	611	0
cawn	-	-	-	4,076	0
bawn	-	-	-	-	0

Tabla 1: Overlapping between wordnet pairs

For verbs, at a synset level, the overlapping is also quite high but less uniform between wordnet pairs. The maximum overlapping occurs also between English and Spanish (7,464) and the lowest between Italian and Basque (1,844).

At a synset level, adjective overlapping is not high because some wordnets provide poor coverage on adjectives. While Spanish provides good overlapping with English (the maximum overlapping with 11,087 synsets), Basque wordnet do not provide adjectives at all.

3.3 Uploading Base Concepts

The original set of Base Concepts from EuroWordNet based on WordNet 1.5 totalized 1,030 ILI-records. The Base Concepts from WordNet 1.5 has been mapped to WordNet 1.6. After a manual revision and expasion to all wordnet 1.6 top beginners, the resulting Base Concepts for WordNet 1.6 totalized 1,601 ILI-records.

3.4 Uploading the Top Ontology

	#verbal synsets	#nominal synsets	#relations
Semcor SUBJ	2,490	5,398	69,840
Semcor DOBJ	3,423	6,964	110,102
BNC SUBJ	6,151	2,588	95,065
BNC DOBJ	6,125	4,185	115,542

Tabla 2: Selectional Preferences

The purpose of the EuroWordNet Top Ontology was to enforce more uniformity and compatibility of the different wordnets. Using the Top Concepts (TCs), the Base Concepts (BCs) can be divided into coherent clusters.

By inheriting the Top Ontology assignments via the hyponymy relations it is possible to populate the complete ILI with top-concepts from the Base Concepts.

There are two things to be noted with respect to the inherited Top Ontology assigments. First of all, redundant assigments are added in so far they have not been inherited from higher levels. If an assigment list includes *Animal* but not *Natural*, then *Natural* is added because it is implied by *Animal* according to the Top Ontology hierarchy. The second point is that the hyperonym classification of WordNet is not always consistent with the Top Ontology assignement. This can be a matter of choice, because in EuroWordNet it was not agreed the WordNet classification or it may be incidental because top-concepts, assigned to the higher levels, are no longer valid at deeper levels of the hierarchy. In WordNet1.5 there are Examples of the former case, 3rdOrderEntities that have been classified below psychological_feature that goes to state together will all statitive nominals. A second possibility of inconsistences may arise at lower levels of the WordNet hierarchy. In Eurowordnet it was not performed the verification of the inherited top-concepts at all levels. We plan to cross–check the Top Ontology expansion using the SUMO ontology (Niles and Pease, 2001).

3.5 Uploading Selectional Preferences

The first version of the MCR has been also enriched by a large amount of new relations. A total of 390,549 weighted Selectional Preferences (SPs) (see Table 2) obtained from two different corpora and using different approaches has been uploaded into the MCR. The first set of weighted SPs was obtained (McCarthy, 2001) by means of probability distributions over the noun hierarchy of WordNet1.6 using the parsed trees generated by RASP(Carroll, Minnen, and Briscoe, 1998) from the BNC. The second set was obtained (Agirre and Martinez, 2001; Agirre and Martinez, 2002) from generalizations of the gramatical relations extracted using MiniPar(Lin, 1998) from Semcor.

The SPs have been included in MCR as

noun–verb relations (ROLE) [5]. Although we can distinguish subjects and objects in the database, all of them have been included as a more general ROLE relation, and in fact, most of them overlap.

4 *Porting Process*

Having all this types of different knowledge and properties completely expanded through the whole MCR a new set of inference mechanism can be devised in order to further infer new relations and knowledge. For instance, new relations can be generated when detecting particular *semantic patterns* occurring for some synsets having certain ontological properties, for a particular Domains, etc. That is, new relations can be generated when combining different methods and knowledge. For instance, creating new explicit relations (regular polisemy, nominalizations, etc.) when several relations derived in the integration process have particular confidence scores greater than certain thresholds, ocurring between certain ontological types, etc.

However, without having infered extra knowledge in this porting process all the knowledge integrated into the MCR has been ported (distributed) to the local wordnets.

All wordnets have gained some kind of new knowledge comming from other wordnets by means of the first porting process. A direct result of the upload/integration/porting effort is that all information associated to the ILIs has been automatically ported to the other wordnets. Thus, MultiwordNet Domains are now available to the rest of local wordnets, the EuroWordNet Top Ontology is also available for Italian MultiWordNet and for English WordNet 1.6. Moreover, local relations have been also ported to the rest of wordnets. Thus, Italian and English Wordnet has been enriched with all the new set of relations coming from EuroWordNet. In turn, Basque, Catalan, Italian and Spanish wordnets has been extensively enriched with large amounts of Selectional Preferences acquired from English (e.g Spanish wordnet has gained 279,324 of 390,109 selectional preferences).

5 *The* vaso *example*

When uploading coherently all this knowledge into the MCR a full range of new possibilities appear for improving both Acquisition and WSD problems. We will ilustrate these new capabilities by a simple example of the Spanish noun *vaso* which in the Spanish WordNet has three possible senses. As we will see, we added consistently a large set of explicit knowledge about each sense of *vaso* that can be used to differentiate and characterize better their particular meanings.

5.1 The *vaso* Container Sense

VASO_1 02755829-n 06-NOUN.ARTIFACT FACTOTUM	
English **Italian** **Basque** **Catalan**	drinking_glass glass bicchiere edontzi baso edalontzi got vas
Gloss	a glass container for holding liquids while drinking
TopOntology	
1stOrderEntity-Form-Object 1stOrderEntity-Origin-Artifact 1stOrderEntity-Function-Container 1stOrderEntity-Function-Instrument	

Tabla 3: Vaso_1

This sense of *vaso* is connected to the same ILI as the English synset *<drinking_glass glass>*. This ILI-record, belonging to the Semantic File ARTIFACT has no specific WordNet Domain (FACTOTUM). However, the EuroWordNet Top Ontology provides futher clues about its meaning: it has the following properties: Form-Object, Origin-Artifact, Function-Container and Function-Instrument. Further, comming from the Selectional Preferences acquired from SemCor, we know that the typical things that somebody does with this kind of *vaso* are for instance the corresponding equivalent translations to Spanish for *<polish, shine, smooth, smoothen>* or *<beautify, embellish, prettify>*. Finally, we must add that this also holds for the rest of languages connected (see Table 3).

5.2 The *vaso* Body-part sense

This sense of *vaso* is the equivalent translation of *<vessel, vas>*. This ILI-record, belonging to the Semantic File BODY has assigned a different WordNet Domain (ANATOMY). The EuroWordNet Top Ontology in this case, has the following proper-

[5]INVOLVED and ROLE relationships are defined symmetric

VASO_2 04195626-n 08-NOUN.BODY ANATOMY	
English **Italian** **Basque** **Catalan**	vessel vas vaso dotto canale hodi baso vas
Gloss	a tube in which a body fluid circulates
TopOntology	
1stOrderEntity-Origin-Natural-Living 1stOrderEntity-Form-Substance-Solid 1stOrderEntity-Composition-Part 1stOrderEntity-Function-Container	

Tabla 4: Vaso_2

ties: Form-Substance-Solid, Origin-Natural-Living, Composition-Part and Function-Container. From the Selectional Preferences acquired from SemCor, we know that the typical events applied to this kind of *vaso* are for instance the corresponding equivalent translations to Spanish for <*inject, shoot*> or <*administer, dispense*>. We must add that this knowledge can be also ported to the rest of languages connected (see Table 4).

5.3 The *vaso* Quantity sense

VASO_3 09914390-n 23-NOUN.QUANTITY NUMBER	
English **Italian** **Basque** **Catalan**	glassful glass bicchierata bicchiere basocada got vas
Gloss	the quantity a glass will hold
TopOntology	
1stOrderEntity-Composition-Part 2ndOrderEntity-SituationType-Static 2ndOrderEntity-SituationComponent-Quantity	

Tabla 5: Vaso_3

The last sense of *vaso* is the equivalent translation of <*glassful, glass*>. This ILI-record, belongs to the Semantic File QUANTITY and has has assigned a different WordNet Domain (FACTOTUM-NUMBER). The EuroWordNet Top Ontology in this case, has the following properties: Composition-Part SituationType-Static and SituationComponent-Quantity. From the Selectional Preferences acquired from SemCor, we know that the typical events applied to this kind of *vaso* are for instance the corresponding equivalent translations to Spanish for <*drink, imbibe*> or <*consume, have, ingest take, take_in*>. As before, we must add that this knowledge can be also ported to the rest of languages connected (see Table 5).

6 *Conclusions and Future Work*

The first version of the MCR integrates now into the same EuroWordNet framework (using an upgraded release of Base Concepts and Top Ontology and MultiWordNet Domains) five local wordnets (with three English WordNet versions) with hundreds of thousand of new semantic relations, instances and properties fully expanded. All wordnets gained some kind of new knowledge comming from other wordnets by means of the first porting process. In fact, the resulting MCR is the largest and richest multilingual lexical knowledge base ever build.

In this way, this version of the MCR produced by MEANING is going to constitute the natural multilingual large-scale linguistic resource for a number of semantic processes that need large amounts of linguistic knowledge to be effective tools (e.g. Semantic Web Ontologies). The fact that word senses will be linked to concepts in MCR will allow for the appropriate representation and storage of the acquired knowledge.

Future versions of the MCR may include language dependant information (including syntactic information, subcategorization frames, etc). The consortium must study the current standards for representing this information, i.e. the EAGLES recommendations, Lexical Conceptual Structures, complex semantic relations (Lin and Pantel, 2001), etc.

We need to investigate new inference facilities to enhance the uploading process as suggested before.

Now, after full expansion (**Realization**) of the EuroWordNet Top Ontology properties, we plan a full expansion through the nominal part of the hierarchy of the selectional preferences acquired from SemCor and BNC (and possibly other implicit semantic knowledge currently available in WordNet such as meronymy information).

We also plan further investigation to perform also full bottom–up expansion (**Generalization**), rather than merely expanding top–down the knowledge and properties represented into the MCR. In this case, different knowledge and properties can collapse on particular Base Concepts, Semantic Files, Domains and/or ontological nodes.

With respect the *porting process*, we plan to investigate also a new set of inference mechanism in order to further infer new explicit relations and knowledge (regular polisemy, nominalizations, etc).

References

Agirre, E. and D. Martinez. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the COLING workshop on Semantic Annotation and Intelligent Annotation*, Luxembourg.

Agirre, E. and D. Martinez. 2001. Learning class-to-class selectional preferences. In *Proceedings of CoNLL01*, Toulouse.

Agirre, E. and D. Martinez. 2002. Integrating selectional preferences in wordnet. In *Proceedings of the first International WordNet Conference in Mysore*, India.

Atserias, J., S. Climent, X. Farreres, G. Rigau, and H. Rodríguez. 1997. Combining multiple methods for the automatic construction of multilingual wordnets. In *Procceeding of RANLP'97*, pages 143–149.

Benítez, L., S. Cervell, G. Escudero, M. López, G. Rigau, and M. Taulé. 1998. Methods and tools for building the catalan wordnet. In *Proceedings of the ELRA Workshop on Language Resources for European Minority Languages.*

Bentivogli, L., E. Pianta, and C. Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

Carroll, J., G. Minnen, and E. Briscoe. 1998. Can subcategorisation probabilities help a statistical parser? In *Proceedings of the Sixth ACL/SIGDAT Workshop on Very Large Corpora*, pages 118–126.

Daudé, J., L. Padró, and G. Rigau. 2001. A complete wn1.5 to wn1.6 mapping. In *Proceedings of NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations"*, Pittsburg, PA, United States.

Fellbaum, C., editor. 1998. *WordNet. An Electronic Lexical Database.* MIT Press.

J. Daudé, L. Padró, G. Rigau. 1999. Mapping Multilingual Hierarchies Using Relaxation Labeling. In *Joint SIGDAT Conference on EMNLP/VLC'99*, Maryland.

J. Daudé, L. Padró, G. Rigau. 2000. Mapping WordNets Using Structural Information. In *Proceedings of 38th annual meeting of the Association for Computational Linguistics (ACL'2000).*

Leacock, C., M. Chodorow, and G. Miller. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–166.

Lin, D. 1998. Extracting collocations from text corpora. In *Proceedigns of First Workshop on Computational Terminology*, Montreal, Canada.

Lin, D. and P. Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Magnini, B. and G. Cavaglià. 2000. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*, Athens.

McCarthy, D. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences.* Ph.D. thesis, University of Sussex.

Mihalcea, R. and I. Moldovan. 1999. An Automatic Method for Generating Sense Tagged Corpora. In *Proceedings of the 16th National Conference on Artificial Intelligence.* AAAI Press.

Niles, I. and A. Pease. 2001. 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.

Rigau, G., B. Magnini, E. Agirre, P. Vossen, and J. Carroll. 2002. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of COLLING Workshop*, Taipei, Taiwan.

Vossen, P., editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers .