Third year published papers

Document Number	Working Paper 9.4
Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Project URL	http://www.lsi.upc.es/~nlp/meaning/meaning.html
Availability	Public
Authors: German Rigau (UPV/EHU)	

meaning

INFORMATION SOCIETY TECHNOLOGIES

Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Security (Distribution level)	Public
Contractual date of delivery	February 2005
Actual date of delivery	May 9, 2005
Document Number	Working Paper 9.4
Type	Report
Status & version	v FINAL
Number of pages	24
WP contributing to the deliberable	WP9
WPTask responsible	German Rigau (UPV/EHU)
Authors	German Rigau (UPV/EHU)
Other contributors	
Reviewer	
EC Project Officer	Evangelia Markidou
Authors: German Rigau (UPV/EHU)	
Keywords:	
Abstract: This document provides a brief summary of the published work resulting from the third year of MEANING	

Contents

1 Executive Summary **3**
1.1 Conferences 3
1.2 Workshops 3
1.3 Journals 4

2 Papers related to WP3: Linguistic Processors and Infrastructure **4**

3 Papers related to WP4: (Knowledge) Integration **10**

4 Papers related to WP5: Acquisition **10**

5 Papers related to WP6: Word Sense Disambiguation **14**

6 Papers related to WP7: Evaluation and Assessment **18**

7 Papers related to WP8: User Validation **20**

1 Executive Summary

This document provides a brief summary of the published work resulting from the second year of MEANING. Last year, the consortium published 51 papers: 26 papers in International Conference Proceedings, 22 papers in International WorkShops and 3 papers in international journals. These papers covers different MEANING working parts: from WP2 to WP9.

Next, we provide the complete list of Conferences and Workshops attended by the MEANING partners. We included here also Journals and books.

The rest of sections of this Working Paper provide, one per Work Part, a detailed list of the published papers. Each paper comes together with a brief summary.

1.1 Conferences

- (10 papers) Fourth International Conference on Language Resources and Evaluation (LREC'04)
- (4 papers) 20th International Conference on Computational Linguistics (COLING'04)
- (4 papers) Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'05)
- (2 papers) International Conference on Empirical Methods in Natural Language Processing (EMNLP'04)
- (2 papers) España for Natural Language Processing (EsTAL'04)
- (1 paper) 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)
- (1 paper) 8th Conference on Computational Natural Language Learning (CoNLL'04
- (1 paper) Document Understanding Conference (DUC'04)
- (1 paper) 20th Conference of the Spanish Society for NLP (SEPLN'04)

1.2 Workshops

- (10 papers) Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)
- (2 papers) LREC Workshop on "XML-based Richly Annotated Corpora"
- (1 paper) LREC Workshop on "The Amazing Utility of Parallel and Comparable Corpora"
- (1 paper) ACL Workshop on "Multiword Expressions".

- (1 paper) ACL Student Research Workshop.
- (1 paper) COLING Workshop on "Recent Advances in Dependency Grammar".
- (1 paper) COLING Workshop on "Multilingual Linguistic Resources".
- (1 paper) 4th International SALTMIL (ISCA SIG) LREC workshop on "First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation".
- (1 paper) IJCNLP Workshop on "Named Entity Recognition".
- (1 paper) IBERAMIA Workshop on "Lexical Resources and The Web for Word Sense Disambiguation"
- (1 paper) Workshop on "International Proofing Tools and Language Technologies".
- (1 paper) MEANING workshop

1.3 Journals

- (1 paper) Machine Learning, special issue on Speech and Natural Language Processing
- (1 paper) Journal of Artificial Intelligence Research
- (1 paper) Computer Speech and Language

2 Papers related to WP3: Linguistic Processors and Infrastructure

1. [Arranz *et al.*, 2005] *Multiword Expressions and Word Sense Disambiguation*

   This paper studies the impact of multiword expressions on Word Sense Disambiguation (WSD). Several identification strategies of the multiwords in WordNet2.0 are tested in a real Senseval-3 task: the disambiguation of WordNet glosses. Although we have focused on Word Sense Disambiguation, the same techniques could be applied in more complex tasks, such as Information Retrieval or Question Answering.

2. [Giménez and Màrquez, 2004] *SVMTool: A general POS tagger generator based on Support Vector Machines*

   This paper presents the SVMTool, a simple, flexible, effective and efficient part of speech tagger based on Support Vector Machines. The SVMTool offers a fairly good balance among these properties which make it really practical for current NLP applications. It is very easy to use and easily configurable so as to perfectly fit the needs

of a number of different applications. Results are also very competitive, achieving an accuracy of 97.2% for English on the Wall Street Journal corpus. It has been also successfully applied to Spanish and Catalan exhibiting a similar performance. A first release of the SVMTool prototype (Perl/C++) is now freely available for public use.

3. [Carreras *et al.*, 2004] *Hierarchical Recognition of Propositional Arguments with Perceptrons*

   In this paper, we describe a system for the CoNLL-2004 Shared Task on Semantic Role Labeling. The system implements a two-layer learning architecture to recognize arguments in a sentence and predict the role they play in the propositions. The exploration strategy visits possible arguments bottom-up, navigating through the clause hierarchy. The learning components in the architecture are implemented as Perceptrons, and are trained simultaneously online, adapting their behavior to the global target of the system. The learning algorithm follows the global strategy introduced in (Collins02) and adapted in (Carreras04) for partial parsing tasks.

4. [Carreras *et al.*, 2005] *iltering-Ranking Perceptron Learning for Partial Parsing.*

   This work introduces a general phrase recognition system based on perceptrons, and a global online learning algorithm to train them together. The method applies to complex domains in which some structure has to be recognized. This global problem is broken down into two layers of local subproblems: a filtering layer, which reduces the search space by identifying plausible phrase candidates; and a ranking layer, which builds the optimal phrase structure by discriminating among competing phrases. A recognition-based feedback rule is presented which reflects to each local function its committed errors from a global point of view, and allows to train them together online as perceptrons. As a result, the learned functions automatically behave as filters and rankers, rather than binary classifiers, which we argue to be better for this type of problems. Extensive experimentation on partial parsing tasks gives state-of-the-art results and evinces the advantages of the global training method over optimizing each function locally and independently.

5. [Artola, 2004] *Laying Lexical Foundations for NLP: the Case of Basque at the Ixa Research Group*

   The purpose of this paper is to present the strategy and methodology followed at the Ixa NLP Group of the University of The Basque Country in laying the lexical foundations for language processing. Monolingual and bilingual dictionaries, text corpora, and linguists knowledge have been the main information sources from which lexical knowledge currently present in our NLP system has been acquired. The main lexical resource we use in research and applications is a lexical database, EDBL, that currently contains more than 80,000 entries richly coded with the lexical information needed in language processing tasks. A Basque wordnet has also been built (it has currently more than 50,000 word senses), although it is not yet fully integrated into the processing chain as EDBL is. Monolingual dictionaries have been exploited in

order to obtain knowledge that is currently being integrated into a lexical knowledge base (EEBL). This knowledge base is being connected to the lexical database and to the wordnet. Feedback obtained from users of the first language technology practical application produced by the research group, i.e. a spelling checker, has also been an important source of lexical knowledge that has permitted to improve, correct and update the lexical database. In the paper, doctorate research work on the lexicon finished or in progress at the group is outlined as well, as long as a brief description of the end-user applications produced so far.

6. [Alegria *et al.*, 2004a] *Representation and Treatment of Multiword Expressions in Basque*

   This paper describes the representation of Basque Multiword Lexical Units and the automatic processing of Multiword Expressions. After discussing and stating which kind of multiword expressions we consider to be processed at the current stage of the work, we present the representation schema of the corresponding lexical units in a general-purpose lexical database. Due to its expressive power, the schema can deal not only with fixed expressions but also with morphosyntactically flexible constructions. It also allows us to lemmatize word combinations as a unit and yet to parse the components individually if necessary. Moreover, we describe HABIL, a tool for the automatic processing of these expressions, and we give some evaluation results. This work must be placed in a general framework of written Basque processing tools, which currently ranges from the tokenization and segmentation of single words up to the syntactic tagging of general texts.

7. [de Ilarraza *et al.*, 2005] *Design and Development of a System for the Detection of Agreement Errors in Basque*

   This paper presents the design and development of a system for the detection and correction of syntactic errors in free texts. The system is composed of three main modules: a) a robust syntactic analyser, b) a compiler that will translate error processing rules, and c) a module that coordinates the results of the analyser, applying different combinations of the already compiled error rules. The use of the syntactic analyser (a) and the rule processor (b) is independent and not necessarily sequential. The specification language used for the description of the error detection/correction rules is abstract, general, declarative, and based on linguistic information.

8. [Aranzabe *et al.*, 2004] *Towards a Dependency Parser of Basque*

   We present the Dependency Parser, called M axuxta, for the linguistic processing of Basque, which can serve as a representative of agglutinative languages that are also characterized by the free order of its constituents. The Dependency syntactic model is applied to establish the dependency-based grammatical relations between the components within the clause. Such a deep analysis is used to improve the output of the shallow parsing where syntactic structure ambiguity is not fully and explicitly resolved. Previous to the completion of the grammar for the dependency parsing,

the design of the Dependency Structure-based Scheme had to be accomplished; we concentrated on issues that must be resolved by any practical system that uses such models. This scheme was used both to the manual tagging of the corpus and to develop the parser. The manually tagged corpus has been used to evaluate the accuracy of the parser. We have evaluated the application of the grammar to corpus, measuring the linking of the verb with its dependents, with satisfactory results.

9. [de Ilarraza *et al.*, 2004] *Abar-Hitz: An Annotation Tool for the Basque Dependency Treebank*

   This paper presents the process followed to design and build a graphical and language independent tool, Abar-Hitz, for the creation and management of the Basque Dependency Treebank. Abar-Hitz makes the annotation process faster and avoids possible mistakes linguists can make. It is composed of three areas: the corpus area, the tagging area and the tree visualizer area. Three linguists used Abar-Hitz to tag 25.000 word-forms from the Eus3LB corpus, making clear, as the evaluation results show, its utility.

10. [Aduriz *et al.*, 2004] *A Cascaded Syntactic Analyser for Basque*

    This article presents a robust syntactic analyser for Basque and the different modules it contains. Each module is structured in different analysis layers for which each layer takes the information provided by the previous layer as its input; thus creating a gradually deeper syntactic analysis in cascade. This analysis is carried out using the Constraint Grammar (CG) formalism. Moreover, the article describes the standardisation process of the parsing formats using XML.

11. [Palomar *et al.*, 2004] *3LB: Construcción de una base de árboles sintáctico-semánticos para el catalán, euskera y castellano*

    In this paper, we present the results of the 3LB project, which consist on the development of three corpora (one for Catalan, one for Spanish, and one for Basque) with syntactic and semantic annotation. We show the criteria followed for each annotation, the different tools developed for each tagging and the results of annotation evaluation.

12. [Artola *et al.*, 2004] *EULIA: a graphical web interface for creatin, browsing and editing linguistically annotated corpora*

    In this paper we present EULIA, a tool which has been designed for dealing with the linguistic annotated corpora generated by a set of different linguistic processing tools. The objective of EULIA is to provide a flexible and extensible environment for creating, consulting, visualizing, and modifying documents generated by existing linguistic tools. The documents used as input and output of the different tools contain TEI-conformant feature structures (FS) coded in XML. The tools integrated until now are a lexical database, a tokenizer, a wide-coverage morphosyntactic analyzer, a general purpose tagger/lemmatizer, and a shallow syntactic analyzer.

13. [Alegria *et al.*, 2004b] *Design and Development of a Named Entity Recognizer for an Agglutinative Language*

    This paper presents the conclusions reached from the development of a system for Named Entity recognition in written Basque. The system was designed in four steps: first, the development of a recognizer based on linguistic information represented on finite-state-transducers; second, the generation of semi-automatically annotated corpora from the result of these transducers; third, the achievement of the best possible recognizer by training different ML techniques on these corpora; and finally, the combination of the different recognizers obtained. Being Basque an agglutinative language, a linguistic preprocess previous to these steps was required.

14. [Ansa *et al.*, 2004] *Integrating NLP Tools for Basque in Text Editors*

    In this paper we present the integration of several NLP tools in text editors. These tools have been developed following a strategy of five phases that we have designed for the processing of Basque. We are nowadays involved in the fourth phase of the mentioned strategy and have already developed and integrated three significant NLP tools the spelling checker/corrector Xuxen, the Spanish/Basque Elhuyar Dictionary and the Synonym Dictionary . Our current goal is the grammar checker/corrector, called Xuxeng, and we hope its first version will be integrated in text editors in a short time. From our experience, we know all this technology is relevant to make easier the use of written Basque as well as to help in the standardisation process of our language.

15. [Bentivogli *et al.*, 2004b] *Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus*

    In this paper we illustrate and evaluate an approach to the creation of high quality linguistically annotated resources based on the exploitation of aligned parallel corpora. This approach is based on the assumption that if a text in one language has been annotated and its translation has not, annotations can be transferred from the source text to the target using word alignment as a bridge. The transfer approach has been tested in the creation of the MultiSemCor corpus, an English/Italian parallel corpus created on the basis of the English SemCor corpus. In MultiSemCor texts are aligned at the word level and semantically annotated with a shared inventory of senses. We present some experiments carried out to evaluate the different steps involved in the methodology. The results of the evaluation suggest that the cross-language annotation transfer methodology is a promising solution allowing for the exploitation of existing (mostly English) annotated resources to bootstrap the creation of annotated corpora in new (resourcepoor) languages with greatly reduced human effort.

16. [Pianta and Bentivogli, 2004b] *Knowledge Intensive Word Alignment with KNOWA*

    In this paper we present KNOWA, an English/Italian word aligner, developed at ITC-irst, which relies mostly on information contained in bilingual dictionaries. The

performances of KNOWA are compared with those of GIZA++, a state of the art statistics-based alignment algorithm. The two algorithms are evaluated on the EuroCor and MultiSemCor tasks, that is on two English/Italian publicly available parallel corpora. The results of the evaluation show that, given the nature and the size of the available English-Italian parallel corpora, a language-resource-based word aligner such as KNOWA can outperform a fully statistics-based algorithm such as GIZA++.

17. [Bentivogli *et al.*, 2004a] *Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing*

    The continuous expansion of the multilingual information society has led in recent years to a pressing demand for multilingual linguistic resources suitable to be used for different applications. In this paper we present the WordNet Domains Hierarchy (WDH), a language-independent resource composed of 164, hierarchically organized, domain labels (e.g. Architecture, Sport, Medicine). Although WDH has been successfully applied to various Natural Language Processing tasks, the first available version presented some problems, mostly related to the lack of a clear semantics of the domain labels. Other correlated issues were the coverage and the balancing of the domains. We illustrate a new version of WDH addressing these problems by an explicit and systematic reference to the Dewey Decimal Classification. The new version of WDH has a better defined semantics and is applicable to a wider range of tasks.

18. [Pianta and Bentivogli, 2004a] *Annotating Discontinuous Structures in XML: the Multiword Case*

    In this paper, we address the issue of how to annotate discontinuous elements in XML. We will take discontinuous multiwords as a case study to investigate different annotation possibilities, in the framework of the linguistic annotation of the MEANING Italian Corpus.

19. [Ranieri *et al.*, 2004] *Browsing Multilingual Information with the MultiSemCor Web Interface*

    Parallel and comparable corpora represent a crucial resource for different Natural Language Processing tasks like machine translation, lexical acquisition, and knowledge structuring but are also suitable to be consulted by humans for different purposes, such as linguistic teaching, corpus linguistics, translation studies, lexicography, multilingual information browsing. To enhance their exploitation by human users, specially designed interfaces need to be developed. In this paper we present the design and implementation of the MultiSemCor Web Interface. MultiSemCor is a parallel English/Italian corpus, which is being developed at ITC-irst starting from the English corpus SemCor. In MultiSemCor the texts are aligned at word level and semantically annotated with WordNet senses. The MultiSemCor Web Interface allows the users to exploit at best the potentiality of the corpus. We will describe the main functions of the interface, which provides two distinct browsing modalities:

a bi-text-oriented modality and a word-oriented modality, which amounts to a bilingual semantic concordancer. Moreover, the MultiSemCor Web Interface is integrated with the on-line MultiWordNet browser, which gives access to the reference lexicon for MultSemCor.

3 Papers related to WP4: (Knowledge) Integration

1. [Artola *et al.*, 2004] *Laying Lexical Foundations for NLP: the Case of Basque at the Ixa Research Group*

   See the abstract in section 2.

2. [Atserias *et al.*, 2004c] *Spanish WordNet 1.6: Porting the Spanish Wordnet across Princeton versions*

   This paper describes the new Spanish Wordnet aligned to Princeton WordNet1.6 and the analysis of the transformation from the previous version aligned to Princeton WordNet1.5. Although a mapping technology exists, to our knowledge it is the first time a whole local wordnet has been ported to a newer release of the Princeton WordNet.

3. [Atserias *et al.*, 2004a] *Towards the MEANING Top Ontology: Sources of Ontological Meaning*

   This paper describes the initial research steps towards the Top Ontology for the Multilingual Central Repository (MCR) built in the MEANING project. The current version of the MCR integrates five local wordnets plus four versions of Princeton s EnglishWordNet, three ontologies and hundreds of thousands of new semantic relations and properties automatically acquired from corpora. In order to maintain compatibility among all these heterogeneous knowledge resources, it is fundamental to have a robust and advanced ontological support. This paper studies the mapping of main Sources of Ontological Meaning onto the wordnets and, in particular, the current work in mapping the EuroWordNet Top Concept Ontology.

4 Papers related to WP5: Acquisition

1. [McCarthy *et al.*, 2004b] *Finding Predominant Senses in Untagged Text*

   ACl'2004 Best paper award

   In word sense disambiguation (WSD), the heuristic of choosing the most common sense is extremely powerful because the distribution of the senses of a word is often skewed. The problem with using the predominant, or first sense heuristic, aside from the fact that it does not take surrounding context into account, is that it assumes some quantity of hand-tagged data. Whilst there are a few hand-tagged corpora

available for some languages, one would expect the frequency distribution of the senses of words, particularly topical words, to depend on the genre and domain of the text under consideration. We present work on the use of a thesaurus acquired from raw textual corpora and the WordNet similarity package to find predominant noun senses automatically. The acquired predominant senses give a precision of 64task. This is a very promising result given that our method does not require any hand-tagged text, such as SemCor. Furthermore, we demonstrate that our method discovers appropriate predominant senses for words from two domain-specific corpora.

2. [McCarthy *et al.*, 2004a] *Automatic Identification of Infrequent Word Senses*

   In this paper we show that an unsupervised method for ranking word senses automatically can be used to identify infrequently occurring senses. We demonstrate this using a ranking of noun senses derived from the BNC and evaluating on the sense-tagged text available in both SemCor and the SENSEVAL-2 English all-words task. We show that the method does well at identifying senses that do not occur in a corpus, and that those that are erroneously filtered but do occur typically have a lower frequency than the other senses. This method should be useful for word sense disambiguation systems, allowing effort to be concentrated on more frequent senses; it may also be useful for other tasks such as lexical acquisition. Whilst the results on balanced corpora are promising, our chief motivation for the method is for application to domain specific text. For text within a particular domain many senses from a generic inventory will be rare, and possibly redundant. Since a large domain specific corpus of sense annotated data is not available, we evaluate our method on domain-specific corpora and demonstrate that sense types identified for removal are predominantly senses from outside the domain.

3. [Wang, 2004] *Automatic Acquisition of English Topic Signatures Based on a Second Language*

   This paper presents a novel approach for automatically acquiring English topic signatures. Given a particular concept, or word sense, a topic signature is a set of words that tend to co-occur with it. Topic signatures can be useful in a number of Natural Language Processing applications, such as Word Sense Disambiguation and Text Summarisation. Our method takes advantage of the different way in which word senses are lexicalised in English and Chinese, and also exploits the large amount of Chinese text available in corpora and on the Web. We evaluated the topic signatures on a WSD task, where we trained a second-order vector co-occurrence algorithm on standard WSD datasets, with promising results.

4. [Agirre and Martinez, 2004d] *Unsupervised WSD based on automatically retrieved examples: The importance of bias*

   This paper explores the large-scale acquisition of sense-tagged examples for Word Sense Disambiguation (WSD). We have applied the "WordNet monosemous relatives"

method to construct automatically a web corpus that we have used to train disambiguation systems. The corpus-building process has highlighted important factors, such as the distribution of senses (bias). The corpus has been used to train WSD algorithms that include supervised methods (combining automatic and manually-tagged examples), minimally supervised (requiring sense bias information from hand-tagged corpora), and fully unsupervised. These methods were tested on the Senseval-2 lexical sample test set, and compared successfully to other systems with minimum or no supervision.

5. [Agirre and Martinez, 2004b] *The effect of bias on an automatically-built word sense corpus*

   The goal of this paper is to explore the large-scale automatic acquisition of sense-tagged examples to be used for Word Sense Disambiguation (WSD). We have applied the "monosemous relatives" method on the Web in order to build such a resource for all nouns in WordNet. The analysis of some parameters revealed that the distribution of the word senses (bias) in the training and test corpus is a determinant factor. Provided there is a method to approximate the bias for each word sense, the results we obtained for English are comparable to the use of hand-tagged data (Semcor), which is a very interesting perspective for lesser studied languages.

6. [Agirre *et al.*, 2004b] *Exploring portability of syntactic information from English to Basque*

   This paper explores a crosslingual approach to the PP attachment problem. We built a large dependency database for English based on an automatic parse of the BNC, and Reuters (sports and finances sections). The Basque attachment decisions are taken based on the occurrence frequency of the translations of the Basque (verb-noun) pairs in the English syntactic database. The results show that with this simple technique it is possible to transfer syntactic information from a language like English in order to make PP attachment decisions in another language, in this case Basque.

7. [Agirre and de Lacalle, 2004] *Publicly available topic signatures for all WordNet nominal senses*

   Topic signatures are context vectors built for word senses and concepts. They can be automatically acquired from the web for any concept hierarchy using the "monosemous relative" method. Topic signatures have been shown to be useful in Word Sense Disambiguation, for modeling similarity between word senses, classifying new terms in hierarchies and also building hierarchical clusters of word senses for a given word. In this work we present a publicly available resource which comprises both automatically extracted examples for all WordNet 1.6 noun senses and topic signatures built based on those examples. We gathered around 700 sentences per each noun in WordNet. When the monosemous relatives are used to build a sense corpus for polysemous words, they comprise an average of around 3,500 sentences per word

sense. The size of the topic signatures thus constructed is of around 4,500 words per word sense.

8. [Atserias *et al.*, 2004b] *Cross-Language Acquisition of Semantic Models for Verbal Predicates*

   This paper presents a semantic-driven methodology for the automatic acquisition of verbal models. Our approach relies strongly on the semantic generalizations allowed by already existing resources (e.g. Domain labels, Named Entity categories, concepts in the SUMO ontology, etc). Several experiments have been carried out using comparable corpora in four languages (Italian, Spanish, Basque and English) and two domains (FINANCE and SPORT) showing that the semantic patterns acquired can be general enough to be ported from one language to the other language.

9. [Fernández *et al.*, 2004] *Automatic Acquisition of Sense Examples using ExRetriever*

   A current research line for word sense disambiguation (WSD) focuses on the use of supervised machine learning techniques. One of the drawbacks of using such techniques is that previously sense annotated data is required. This paper presents ExRetriever, a new software tool for automatically acquiring large sets of sense tagged examples from large collections of text and the Web. ExRetriever exploits the knowledge contained in large-scale knowledge bases (e.g., WordNet) to build complex queries, each of them characterising particular senses of a word. These examples can be used as training instances for supervised WSD algorithms.

10. [Cuadros *et al.*, 2004] *Automatic Acquisition of Sense Examples using ExRetriever*

    A promising research line for word sense disambiguation (WSD) focuses on the use of supervised machine learning techniques. One of the drawbacks of using such techniques is that they requires previously sense annotated data. This paper presents ExRetriever, a new software tool for automatically acquiring large sets of sense tagged examples from large collections of text (e.g. the Web). ExRetriever exploits large-scale knowledge bases (e.g., WordNet) to build complex queries, each of them characterising particular senses of a word. These examples can be used as training instances for supervised WSD algorithms.

11. [D'Avanzo *et al.*, 2004] *Keyphrase Extraction for Summarization Purposes: The LAKE System at DUC-2004*

    We report on ITC-irst participation at Task 1 (very short document summaries) at DUC-2004. We propose to exploit a keyphrase extraction methodology in order to identify relevant terms in the document. The LAKE algorithm first considers a number of linguistic featues to extract a list of well motivated candidate keyphrases, then uses a machine learning framework to select significat keyphrases for a document. With respect to other approaches to keyphrase extraction, LAKE makes use of linguistic processors such as multiword and named entities recognition, which are not usually exploited.

12. [D'Avanzo *et al.*, 2005] *Automatic acquisition of domain information for lexical concepts*

    In this paper we adopted Latent Semantic Kernels to perform a Term Categorization task, and we applied this technique to assign domain labels to monosemous words. Results show that the proposed technique is e ective, achieving an accuracy of about 43% for all the monosemus terms in a corpus. We also reported an error analysis showing that most of the misclassi cation errors are related the the fuzzy nature of domain distinctions. In particular we identi ed a set of "families" in the WordNet Domains categories that makes di cult the classi cation task.

5 Papers related to WP6: Word Sense Disambiguation

1. [Arranz *et al.*, 2005] *Multiword Expressions and Word Sense Disambiguation*

   See the abstract in section 2.

2. [McCarthy *et al.*, 2004b] *Finding Predominant Senses in Untagged Text*

   See the abstract in section 4.

3. [McCarthy *et al.*, 2004c] *Using automatically acquired predominant senses for word sense disambiguation*

   In word sense disambiguation (WSD), the heuristic of choosing the most common sense is extremely powerful because the distribution of the senses of a word is often skewed. The first (or predominant) sense heuristic assumes the availability of hand-tagged data. Whilst there are hand-tagged corpora available for some languages, these are relatively small in size and many word forms either do not occur, or occur infrequently. In this paper we investigate the performance of an unsupervised first sense heuristic where predominant senses are acquired automatically from raw text. We evaluate on both the SENSEVAL-2 and SENSEVAL-3 English all-words data. For accurate WSD the first sense heuristic should be used only as a back-off, where the evidence from the context is not strong enough. In this paper however, we examine the performance of the automatically acquired first sense in isolation since it turned out that the first sense taken from SemCor outperformed many systems in SENSEVAL-2.

4. [Villarejo *et al.*, 2004] *The MEANING system on the English Allwords task*

   This paper reports the work done in building a WSD system for the Senseval-3 English all-words task, which integrates several supervised machine learning word sense disambiguation modules, and several knowledge–based (unsupervised) modules. The supervised modules have been trained exclusively on the SemCor corpus, while the unsupervised modules use WordNet-based lexico–semantic resources integrated in

the Multilingual Central Repository (MCR). The architecture of the system is quite simple. Raw text is passed through a pipeline of linguistic processors (tokenizers, POS tagging, named entity extraction, and parsing) and then a Feature Extraction module codifies examples with features extracted from the linguistic annotation and MCR. The supervised modules have priority over the unsupervised and they are combined using a weighted voting scheme. For the words lacking training examples, the unsupervised modules are applied in a cascade sorted by decreasing precision. The tuning of the combination setting has been performed on the Senseval-2 allwords corpus. Two systems were presented: 1) Meaning-allwords, which is an ensemble of all supervised and unsupervised classifiers, and 2) Meaning-simple, which is a combination of the better supervised and unsupervised systems.

5. [Castillo *et al.*, 2004] *The TALP Systems for Disambiguating WordNet Glosses*

   This paper describes the TALP systems presented at Senseval-3 task 12 "Word-Sense Disambiguation of WordNet Glosses". Our method combines a set of knowledge-based heuristics integrating several information sources and techniques. From the ten systems presented at the task, our system has obtained the first position of recall.

6. [Wang, 2004] *Automatic Acquisition of English Topic Signatures Based on a Second Language*

   See the abstract in section 4.

7. [Màrquez *et al.*, 2004c] *On the Quality of Lexical Resources for Word Sense Disambiguation*

   Word Sense Disambiguation (WSD) systems are usually evaluated by comparing their absolute performance, in a fixed experimental setting, to other alternative algorithms and methods. However, little attention has been paid to analyze the lexical resources and the corpora defining the experimental settings and their possible interactions with the overall results obtained. In this paper we present some experiments supporting the hypothesis that the quality of lexical resources used for tagging the training corpora of WSD systems determines in some way the quality of the results. In order to verify this initial hypothesis we have developed two kinds of experiments. At the linguistic level, we have tested the quality of lexical resources in terms of the annotators' agreement degree. From the computational point of view, we have evaluated how those different lexical resources determine the quality of the WSD methods. We have carried out these experiments using three different lexical resources as sense inventories and a fixed WSD system based on Support Vector Machines.

8. [Escudero *et al.*, 2004] *The TALP System for the English Lexical Sample Task*

   This paper describes the TALP system on the English Lexical Sample task of the Senseval-3 event. The system is fully supervised and relies on a particular Machine

Learning algorithm, namely Support Vector Machines. It does not use extra examples than those provided by Senseval-3 organisers, though it uses external tools and ontologies to extract part of the representation features.

9. [Agirre and Martinez, 2004d] *Unsupervised WSD based on automatically retrieved examples: The importance of bias*

   See the abstract in section 4.

10. [Agirre and Martinez, 2004b] *The effect of bias on an automatically-built word sense corpus*

    See the abstract in section 4.

11. [Agirre and Martinez, 2004a] *The effect of bias on an automatically-built word sense corpus*

    Our group participated in the Basque and English lexical sample tasks in Senseval-3. A language-specific feature set was defined for Basque. Four different learning algorithms were applied, and also a method that combined their outputs. Before submission, the performance of the methods was tested for each task on the Senseval-3 training data using cross validation. Finally, two systems were submitted for each language: the best single algorithm and the best ensemble.

12. [Agirre and Martinez, 2004c] *Smoothing and Word Sense Disambiguation*

    This paper presents an algorithm to apply the smoothing techniques described in (Yarowsky, 1995) to three different Machine Learning (ML) methods for Word Sense Disambiguation (WSD). The method to obtain better estimations for the features is explained step by step, and applied to n-way ambiguities. The results obtained in the Senseval-2 framework show that the method can help improve the precision of some weak learners, and in combination attain the best results so far in this setting.

13. [Montoyo *et al.*, 2005] *Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods*

    In this paper we concentrate on the resolution of the lexical ambiguity that arises when a given word has several different meanings. This specific task is commonly referred to as word sense disambiguation (WSD). The task of WSD consists of assigning the correct sense to words using an electronic dictionary as the source of word definitions. We present two WSD methods based on two main methodological approaches in this research area: a knowledge-based method and a corpus-based method. Our hypothesis is that word-sense disambiguation requires several knowledge sources in order to solve the semantic ambiguity of the words. These sources can be of different kinds— for example, syntagmatic, paradigmatic or statistical information. Our approach combines various sources of knowledge, through combinations of the two WSD methods mentioned above. Mainly, the paper concentrates on how to combine

these methods and sources of information in order to achieve good results in the disambiguation. Finally, this paper presents a comprehensive study and experimental work on evaluation of the methods and their combinations.

14. [Gliozzo *et al.*, 2004b] *Unsupervised and supervised exploitation of semantic domains in lexical disambiguation*

    Domains are common areas of human discussion, such as economics, politics, law, science etc., which are at the basis of lexical coherence. This paper explores the dual role of domains in word sense disambiguation (WSD). On one hand, domain information provides generalized features at the paradigmatic level that are useful to discriminate among word senses. On the other hand, domain distinctions constitute a useful level of coarse grained sense distinctions, which lends itself to more accurate disambiguation with lower amounts of knowledge. In this paper we extend and ground the modeling of domains and the exploitation of WordNet Domains, an extension of WordNet in which each synset is labeled with domain information.We propose a novel unsupervised probabilistic method for the critical step of estimating domain relevance for contexts, and suggest utilizing it within unsupervised Domain Driven Disambiguation (DDD) for word senses, as well as within a traditional supervised approach. The paper presents empirical assessments of the potential utilization of domains in WSD at a wide range of comparative settings, supervised and unsupervised. Following the dual role of domains we report experiments that evaluate both the extent to which domain information provides e ective features for WSD, as well as the accuracy obtained by WSD at domain-level sense granularity. Furthermore, we demonstrate the potential for either avoiding or minimizing manual annotation thanks to the generalized level of information provided by domains.

15. [Gliozzo *et al.*, 2004a] *Unsupervised domain relevance estimation for word sense disambiguation*

    This paper presents Domain Relevance Estimation (DRE), a fully unsupervised text categorization technique based on the statistical estimation of the relevance of a text with respect to a certain category. We use a pre-de ned set of categories (we call them domains) which have been previously associated to WORDNET word senses. Given a certain domain, DRE distinguishes between relevant and non-relevant texts by means of a Gaussian Mixture model that describes the frequency distribution of domain words inside a large-scale corpus. Then, an Expectation Maximization algorithm computes the parameters that maximize the likelihood of the model on the empirical data. The correct identi cation of the domain of the text is a crucial point for Domain Driven Disambiguation, an unsupervised Word Sense Disambiguation (WSD) methodology that makes use of only domain information. Therefore, DRE has been exploited and evaluated in the context of a WSD task. Results are comparable to those of state-ofthe- art unsupervised WSD systems and show that DRE provides an important contribution.

16. [Gliozzo *et al.*, 2005] *Crossing parallel corpora and multilingual lexical databases for WSD*

    In this paper we propose the use of aligned corpora and multilingual lexical databases to automatically acquire sense tagged data, exploiting the polisemic di erential between two (or more) languages.

17. [Magnini *et al.*, 2004b] *A Semantic-Based Approach to Interoperability of Classification Hierarchies: Evaluation of Linguistic Techniques*

    Classification Hierarchies (CHs) are widely used to organize documents in a way that makes their retrieval easier. Common examples of CHs are Web directories, marketplace catalogs, and file systems. In this paper we discuss and evaluate CTXMATCH, an approach to interopelability that discovers mappings among CHs considering the semantic interpretation of their nodes. CTXMATCH performs a linguistic processing of the labels attached to the nodes, including tokenzation, Part of Speech tagging, multiword recognition and word sense disambiguation. We present an evaluation of the overall performance of the apporach over Web directories as well as systematic analysis of the linguistic modules invilved.

18. [Strapparava *et al.*, 2004] *Pattern abstraction and term similarity for Word Sense Disambiguation: IRST at Senseval-3*

    This paper summarizes IRST's participation in Senseval-3. We participated both in the English allwords task and in some lexical sample tasks (English, Basque, Catalan, Italian, Spanish). We followed two perspectives. On one hand, for the allwords task, we tried to re ne the Domain Driven Disambiguation that we presented at Senseval-2. The re nements consist of both exploiting a new technique (Domain Relevance Estimation) for domain detection in texts, and experimenting with the use of Latent Semantic Analysis to avoid reliance on manually annotated domain resources (e.g. WORDNET DOMAINS). On the other hand, for the lexical sample tasks, we explored the direction of pattern abstraction and we demonstrated the feasibility of leveraging external knowledge using kernel methods.

6 Papers related to WP7: Evaluation and Assessment

1. [Màrquez *et al.*, 2004c] *On the Quality of Lexical Resources for Word Sense Disambiguation*

   See the abstract in section 5.

2. [Màrquez *et al.*, 2004a] *Senseval-3: The Catalan Lexical Sample Task*

   In this paper we describe the Catalan Lexical Sample task. This task was initially devised for evaluating the role of unlabeled examples in supervised and semi-supervised learning systems for WSD and it is the counterpart of the Spanish Lexical Sample

task. It was coordinated also with other lexical sample tasks (Basque, English, Italian, Rumanian, and Spanish) in order to share part of the target words. Firstly, we describe the methodology followed for developing the specific linguistic resources necessary for the task: the MiniDir-Cat lexicon and the MiniCors-Cat corpus. Secondly, we briefly describe the seven participant systems, the results obtained, and a comparative evaluation between them.

3. [Màrquez *et al.*, 2004b] *Senseval-3: The Spanish Lexical Sample Task*

   In this paper we describe the Spanish Lexical Sample task. This task was initially devised for evaluating the role of unlabeled examples in supervised and semi-supervised learning systems for WSD and it was coordinated with five other lexical sample tasks (Basque, Catalan, English, Italian, and Rumanian) in order to share part of the target words. Firstly, we describe the methodology followed to develop the linguistic resources necessary for the task: the MiniDir-2.1 lexicon and the MiniCors corpus. Secondly, we summarize the participant systems, the results obtained, and a comparative analysis.

4. [Taulé *et al.*, 2004] *MiniCors and Cast3LB: Two Semantically Tagged Spanish Corpora*

   In this paper we present two Spanish corpora, MiniCors and Cast3LB, semantically tagged according to different annotation criteria and objectives. In order to guarantee the quality of the results, we have established a methodology for the development of these corpora. The resulting resources consist of a semantically tagged corpus according to the lexical sample task, and a semantically tagged corpus according to the all words task, both of them defined within the Senseval framework.

5. [Agirre *et al.*, 2004a] *The Basque lexical-sample task*

   In this paper we describe the Senseval 3 Basque lexical sample task. The task comprised 40 words (15 nouns, 15 verbs and 10 adjectives) selected from the Basque WordNet. 10 of the words were chosen in coordination with other lexical-sample tasks. The examples were taken from newspapers, an in-house balanced corpus and Internet texts. We additionally included a large set of untagged examples, and a lemmatised version of the data including lemma, PoS and case information. The method used to hand-tag the examples produced an inter-tagger agreement of 78.2% before arbitration. The eight competing systems attained results well above the most frequent baseline and the best system from Swarthmore College scored 70.4% recall.

6. [Magnini *et al.*, 2004a] *The Italian Lexical Sample Task at Senseval-3*

   The Italian lexical sample task at SENSEVAL-3 provided a framework to evaluate supervised and semi-supervised WSD systems. This paper reports on the task preparation which offered the opportunity to review and refine the Italian MultiWordNet and on the results of the six participants, focussing on both the manual and automatic tagging procedures.

7 Papers related to WP8: User Validation

1. [Artola *et al.*, 2004] *Laying Lexical Foundations for NLP: the Case of Basque at the Ixa Research Group*

   See the abstract in section 2.

References

[Aduriz *et al.*, 2004] I. Aduriz, M. Aranzabe, J. Arriola, A. Atutxa, A. Díaz de Ilarraza, K. Gojenola, M. Oronoz, and L. Uria. A cascaded syntactic analyser for basque. In *Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'04)*, Seoul, Korea, 2004.

[Agirre and de Lacalle, 2004] Eneko Agirre and Oier Lopez de Lacalle. Publicly available topic signatures for all wordnet nominal senses. In *Proceedings of the 4rd International Conference on Language Resources and Evaluations (LREC). Lisbon, Portugal.*, 2004.

[Agirre and Martinez, 2004a] Eneko Agirre and David Martinez. The basque country university system: English and basque tasks. In *Proceedings of the 3rd ACL workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL). Barcelona, Spain.*, 2004.

[Agirre and Martinez, 2004b] Eneko Agirre and David Martinez. The effect of bias on an automatically-built word sense corpus. In *Proceedings of the 4rd International Conference on Language Resources and Evaluations (LREC). Lisbon, Portugal.*, 2004.

[Agirre and Martinez, 2004c] Eneko Agirre and David Martinez. Smoothing and word sense disambiguation. In *EsTAL - España for Natural Language Processing. Alicante, Spain. Published in the Springer Verlag Lecture Notes in Computer Science. Editors: José Luis Vicedo, Patricio Martínez-Barco, Rafael Muñoz, et al. Copyright Springer-Verlag.*, 2004.

[Agirre and Martinez, 2004d] Eneko Agirre and David Martinez. Unsupervised wsd based on automatically retrieved examples: The importance of bias. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Barcelona, Spain.*, 2004.

[Agirre *et al.*, 2004a] Eneko Agirre, Itziar Aldabe, Mikel Lersundi, David Martinez, Eli Pociello, and Larraitz Uria. The basque lexical-sample task. In *Proceedings of the 3rd ACL workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL). Barcelona, Spain.*, 2004.

[Agirre *et al.*, 2004b] Eneko Agirre, Aitziber Atutxa, Koldo Gojenola, and Kepa Sarasola. Exploring portability of syntactic information from english to basque. In *Proceedings of the 4rd International Conference on Language Resources and Evaluations (LREC). Lisbon, Portugal.*, 2004.

[Alegria *et al.*, 2004a] I. Alegria, O. Ansa, X. Artola, N. Ezeiza, K. Gojenola, and R. Urizar. Representation and treatment of multiword expressions in basque. In *Proceedings of the ACL workshop on Multiword Expressions. Barcelona.*, 2004.

[Alegria *et al.*, 2004b] I. Alegria, O. Arregi, I. Balza, N. Ezeiza, I. Fernandez, and R. Urizar. Design and development of a named entity recognizer for an agglutinative language. In *Proceedigns of the IJCNLP-04 workshop on Named Entity Recognition*, Hong Kong, China, 2004.

[Ansa *et al.*, 2004] O. Ansa, X. Arregi, B. Arrieta, N. Ezeiza, I. Fernandez, A. Garmendia, K. Gojenola, B. Laskurain, E. Martínez, M. Oronoz, A. Otegi, K. Sarasola, and L. Uria. Integrating nlp tools for basque in text editors. In *Workshop on International Proofing Tools and Language Technologies. University of Patras. Greece.*, 2004.

[Aranzabe *et al.*, 2004] M. Aranzabe, J.M. Arriola, and A. Díaz de Ilarraza. Towards a dependency parser of basque. In *Proceedings of the Coling 2004 Workshop on Recent Advances in Dependency Grammar*, 2004. Geneva, Switzerland.

[Arranz *et al.*, 2005] Victoria Arranz, Jordi Atserias, and Mauro Castillo. Multiword expressions and word sense disambiguation. In *CICLING'2005*, 2005.

[Artola *et al.*, 2004] X. Artola, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, A. Sologaistoa, and A. Soroa. Eulia: a graphical web interface for creatin, browsing and editing linguistically annotated corpora. In *LREC 2004. Workshop on "XML-based Richly Annotated Corpora". Lisbon, Portugal.*, 2004.

[Artola, 2004] Xavier Artola. Laying lexical foundations for nlp: the case of basque at the ixa research group. In *4th International SALTMIL (ISCA SIG) LREC workshop on "First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation" 24 May 2004, Lisbon, Portugal*, 2004.

[Atserias *et al.*, 2004a] J. Atserias, S. Climent, and G. Rigau. Towards the meaning top ontology: Sources of ontological meaning. In *4rd International Conference on Language Resources and Evaluations (LREC)*, 2004.

[Atserias *et al.*, 2004b] J. Atserias, B. Magnini, O. Popescu, E. Agirre, E. Pociello, G. Rigau, J. Carroll, and R. Koeling. Cross-language acquisition of semantic models for verbal predicates. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC-2004*, 2004. Lisbon, Portugal.

[Atserias *et al.*, 2004c] J. Atserias, G. Rigau, and L. Villarejo. Spanish wordnet 1.6: Porting the spanish wordnet across princeton versions. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, 2004.

[Bentivogli *et al.*, 2004a] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In *Proceedings of COLING 2004 Workshop on "Multilingual Linguistic Resources"*, 2004. Geneva, Switzerland.

[Bentivogli *et al.*, 2004b] L. Bentivogli, P. Forner, and E. Pianta. Evaluating cross-language annotation transfer in the multisemcor corpus. In *Proceedings of COLING 2004*, 2004. Geneva, Switzerland.

[Carreras *et al.*, 2004] X. Carreras, L. Màrquez, and G. Chrupala. Hierarchical recognition of propositional arguments with perceptrons. In *Proceedings of the 8th Conference on Computational Natural Language Learning, CoNLL-2004*, 2004. Boston, MA, USA.

[Carreras *et al.*, 2005] X. Carreras, L. Màrquez, and J. Castro. Filtering-ranking perceptron learning for partial parsing. *Machine Learning. Special issue on Speech and Natural Language Processing*, ??(??):??–??, 2005.

[Castillo *et al.*, 2004] M. Castillo, F. Real, J. Atserias, and G. Rigau. The talp systems for disambiguating wordnet glosses. In *Proceedings of the Senseval-3 ACL-SIGLEX Workshop*, 2004. Barcelona, Spain.

[Cuadros *et al.*, 2004] M. Cuadros, J. Atserias, M. Castillo, and G. Rigau. Automatic acquisition of sense examples using exretriever. In *Proceedings of the IBERAMIA Workshop on Lexical Resources and The Web for Word Sense Disambiguation*, Puebla, Mexico, 2004.

[D'Avanzo *et al.*, 2004] E. D'Avanzo, B. Magnini, and A. Vallin. Keyphrase extraction for summarization purposes: The lake system at duc-2004. In *Proceedings of the Document Understanding Conference (DUC-2004)*, 2004. Boston, USA.

[D'Avanzo *et al.*, 2005] E. D'Avanzo, A. Gliozzo, and C. Strapparava. Automatic acquisition of domain information for lexical concepts. In *Proceedings of the MEANING workshop*, 2005. Trento, Italy.

[de Ilarraza *et al.*, 2004] A. Díaz de Ilarraza, A. Garmendia, and M. Oronoz. Abar-hitz: An annotation tool for the basque dependency treebank. In *Proceedings of the 4rd International Conference on Language Resources and Evaluations (LREC)*, 2004. Lisbon, Portugal.

[de Ilarraza *et al.*, 2005] A. Díaz de Ilarraza, K. Gojenola, and M. Oronoz. Design and development of a system for the detection of agreement errors in basque. In *CICLing-2005, Sixth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico*, 2005.

[Escudero *et al.*, 2004] G. Escudero, L. Màrquez, and G. Rigau. The talp system for the english lexical sample task. In *Proceedings of the Senseval-3 ACL-SIGLEX Workshop*, 2004. Barcelona, Spain.

[Fernández *et al.*, 2004] J. Fernández, M. Castillo, G. Rigau, J. Atserias, and J. Turmo. Automatic acquisition of sense examples using exretriever. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, 2004.

[Giménez and Màrquez, 2004] Jesús Giménez and Lluís Màrquez. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of the 4th LREC Conference*, 2004. Lisbon, Portugal.

[Gliozzo *et al.*, 2004a] A. Gliozzo, B. Magnini, and C. Strapparava. Unsupervised domain relevance estimation for word sense disambiguation. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, 2004. Barcelona, Spain.

[Gliozzo *et al.*, 2004b] A. Gliozzo, C. Strapparava, , and I. Dagan. Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. *Computer Speech and Language*, 18(3):275–299, 2004.

[Gliozzo *et al.*, 2005] A. Gliozzo, C. Strapparava, and M. Ranieri. Crossing parallel corpora and multilingual lexical databases for wsd. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CiCling-05)*, 2005. Mexico City, Mexico.

[Magnini *et al.*, 2004a] B. Magnini, D. Giampiccolo, and A. Vallin. The italian lexical sample task at senseval-3. In *Proceedings of the Senseval-3 ACL-SIGLEX Workshop*, 2004. Barcelona, Spain.

[Magnini *et al.*, 2004b] B. Magnini, M. Speranza, and C. Girardi. A semantic-based approach to interoperability of classification hierarchies: Evaluation of linguistic techniques. In *Proceedings 20th International Conference on Computational Linguistics (Coling'04)*, 2004. Geneva, Switzerland.

[McCarthy *et al.*, 2004a] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Automatic identification of infrequent word senses. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004*, pages 1220–1226, Geneva, Switzerland, 2004.

[McCarthy *et al.*, 2004b] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain, 2004.

[McCarthy *et al.*, 2004c] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Using automatically acquired predominant senses for word sense disambiguation. In *Proceedings of the Senseval-3 ACL-SIGLEX Workshop*, pages 151–154, 2004. Barcelona, Spain.

[Montoyo *et al.*, 2005] A. Montoyo, A. Suarez, G. Rigau, and Palomar. Combining knowledge- and corpus-based word-sense-disambiguation methods. *Journal of Artificial Intelligence Research*, 23:288–330, 2005.

[Màrquez *et al.*, 2004a] L. Màrquez, M. Taulé, M.A. Martí, M. García, F. Real, and D. Ferrés. Senseval-3: The catalan lexical sample task. In *Proceedings of the Senseval-3 ACL-SIGLEX Workshop*, 2004. Barcelona, Spain.

[Màrquez *et al.*, 2004b] L. Màrquez, M. Taulé, M.A. Martí, M. García, F. Real, and D. Ferrés. Senseval-3: The spanish lexical sample task. In *Proceedings of the Senseval-3 ACL-SIGLEX Workshop*, 2004. Barcelona, Spain.

[Màrquez *et al.*, 2004c] L. Màrquez, M. Taulé, L. Padró, L. Villarejo, and M.A. Martí. On the quality of lexical resources for word sense disambiguation. In *Proceedings of the EsTAL Conference*, 2004. Alicante, Spain.

[Palomar *et al.*, 2004] M. Palomar, M. Civit, A. Díaz de Ilarraza, L. Moreno, E. Bisbal, M. Aranzabe, A. Ageno, M.A. Martí, and B. Navarro. 3lb: Construcción de una base de árboles sintáctico-semánticos para el catalán, euskera y castellano. In *XX. Congreso de la SEPLN*, 2004.

[Pianta and Bentivogli, 2004a] E. Pianta and L. Bentivogli. Annotating discontinuous structures in xml: the multiword case. In *Proceedings of the LREC 2004 Satellite Workshop on "XML-based richly annotated corpora"*, 2004. Lisbon, Portugal.

[Pianta and Bentivogli, 2004b] E. Pianta and L. Bentivogli. Knowledge intensive word alignment with knowa. In *Proceedings of COLING 2004*, 2004. Geneva, Switzerland.

[Ranieri *et al.*, 2004] M. Ranieri, E. Pianta, and L. Bentivogli. Browsing multilingual information with the multisemcor web interface. In *Proceedings of the LREC 2004 Satellite Workshop on "The Amazing Utility of Parallel and Comparable Corpora"*, 2004. Lisbon, Portugal.

[Strapparava *et al.*, 2004] C. Strapparava, A. Gliozzo, and C. Giuliano. Pattern abstraction and term similarity for word sense disambiguation: Irst at senseval-3. In *Proceedings of the Senseval-3 ACL-SIGLEX Workshop*, 2004. Barcelona, Spain.

[Taulé *et al.*, 2004] M. Taulé, M. Civit, N. Artigas, M. García, L. Màrquez, M.A. Martí, and B. Navarro. Minicors and cast3lb: Two semantically tagged spanish corpora. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC-2004*, 2004. Lisbon, Portugal.

[Villarejo *et al.*, 2004] L. Villarejo, L. Marquez, E. Agirre, D. Martinez, B. Magnini, C. Strapparava, D. McCarthy, A. Montoyo, and A. Suarez. The meaning system on the english allwords task. In *Proceedings of the Senseval-3 ACL-SIGLEX Workshop*, 2004. Barcelona, Spain.

[Wang, 2004] Xinglong Wang. Automatic acquisition of English topic signatures based on a second language. In *Proceedings of the Student Research Workshop at ACL 2004*, 2004. Barcelona, Spain.