Lexical enrichment of WordNet with Classification Systems using Specification Marks Method*

Andrés Montoyo[1], Manuel Palomar[1] and German Rigau[2]

[1]Department of Software and Computing Systems, University of Alicante, Alicante, Spain
{montoyo, mpalomar}@dlsi.ua.es
[2]TALP Research Center, Universitat Politécnica de Catalunya, 08028 Barcelona, Spain
g.rigau@lsi.upc.es

Abstract: This paper presents an automatic method and interface to enrich semantically WordNet with categories from general domain classification systems. The method is performed in two consecutive steps. First, a lexical knowledge word sense disambiguation process. Second, a set of rules to select the main concepts as representatives for each category. The method has been applied to label automatically WordNet synsets with Subject Codes from a standard news agencies classification system. Experimental results show than the proposed method achieves more than 95% accuracy selecting the main concepts for each category. The interface has been implemented using programming language C++ and providing a visual framework.

1 Introduction and Motivation

Lexical resources are an essential component of language enabled systems. They are one of the main ways of representing the knowledge which applications use in Natural Language Processing (NLP) system, such as Information Retrieval (IR), Information Extraction (IE), Machine Translation (MT), Natural Language Interface or Text Summarization.

Many researchers have proposed several techniques for taking advantage of more than one lexical resource, that is, integrating several structured lexical resources from pre-existing sources.

Byrd in [Br89], proposes the integration of several structured lexical knowledge resources derived from monolingual and bilingual Machine Read Dictionaries (MRD) and Thesaurus. The work reported in [Ro89] used a mapping process between two thesaurus and two sides of a bilingual dictionary. Knight in [Kk93], provides a definition match and hierarchical match algorithms for linking WordNet [Mg90] synsets and LDOCE [Pp87] definitions. Knight and Luk in [KL94], describe the algorithms for merging complementary structured lexical resources from WordNet, LDOCE and a Spanish/English bilingual dictionary. A semiautomatic environment for linking DGILE [Am87] and LDOCE taxonomies using a bilingual dictionary is described in [Aa94]. A semi-automatic method for associating Japanese entries to an English ontology using a

* This research has been partially funded by the UE Commission (NAMIC IST-1999-12302) and the Spanish Research Department (TIC2000-0335-C03-02 and TIC2000-0664-C02-02).

Japanese/English bilingual dictionary is described in [OH94]. An automatic method to enrich semantically the monolingual Spanish dictionary DGILE, using a Spanish/English bilingual dictionary and WordNet is described in [Rg94]. Several methods for linking Spanish and French words from bilingual dictionaries to WordNet synsets are described in [RA95]. A mechanism for linking LDOCE and DGILE taxonomies using a Spanish/English bilingual dictionary and the notion of Conceptual Distance between concepts are described in [RRT95]. The work reported in [CC98] used LDOCE and Roget's Thesaurus to label LDOCE. A robust approach for linking already existing lexical/semantic hierarchies, in particular WordNet 1.5 onto WordNet 1.6, is described in [DPR00]. Magnini and Cavaglia in [MC00] presented an augmented version of nominal part of WordNet, whose synsets have been semi-automatically annotated with one or more subject field codes.

This paper is motivated by two reasons: i) to enrich WordNet with domain labels of classification systems (like IPTC[1] Subject Codes (Version IPTC/1)) to establish semantic relations among word senses and words grouped by their category, ii) the observation that empirical studies about the application of the Specification Marks Method [MP00] proves that it works well for words associated with a semantic category. At the same time the classification systems are now widely used by the NLP tasks. These systems are used by libraries to organise their books, thesauri appear on-line, and on the World Wide Web to organise information by subject. As an alternative approach to improve the lexical knowledge base of WordNet 1.6, this paper presents an automatic method to enrich semantically WordNet 1.6. with categories or classes from that classification systems. The method has been applied to automatically label the WordNet´s synsets with IPTC Subject Codes (Version IPTC/1). Although, this method can also be applied to other classification systems such as Library of Congress Classification(LC)[2], Dewey Decimal Classification (DDC)[3] or Roget's Thesaurus.

The organisation of this paper is as follows: After this introduction, in Section 2 we describe the technique used (Word Sense Disambiguation (WSD) using Specification Marks Method) and its application. In Section 3 we describe the rules used in the method for labelling the noun taxonomy of the WordNet. In section 4, we describe the user interface which allows the enrichment of WordNet. In Section 5, some experiments related to the proposal method are presented, and finally, conclusions and an outline of further lines of research are shown.

2 Specification Marks Method

WSD with Specification Marks is a method for the automatic resolution of lexical ambiguity of groups of words, whose different possible senses are related. The disambiguation is resolved with the use of the WordNet lexical knowledge base (1.6). The method requires the knowledge of how many of the words are grouped around a specification mark, which is similar to a semantic class in the WordNet taxonomy. The word-sense in the sub-hierarchy that contains the greatest number of words for the

---

[1] The IPTC Subject Reference System has been developed to allow Information Providers access to a universal language independant coding system for indicating the subject content of news items. (http://www.iptc.org)

[2] http://lcweb.loc.gov/catalog

[3] http://www.noblenet.org/wakefield/rdewey.htm

corresponding specification mark will be chosen for the sense-disambiguating of a noun in a given group of words.
The algorithm with Specification Marks consists basically of the automatic sense disambiguating of nouns that appear within the context of a sentence and whose different possible senses are related. Its context is the group of words that co-occur with it in the sentence and their relationship to the noun to be disambiguated. The input for the WSD algorithm will be the group of words $w=\{w_1, w_2, ..., w_n\}$. Each word w_i is sought in WordNet, each one has an associated set $s_i=\{s_{i1}, s_{i2}, ..., s_{in}\}$ of possible senses. Furthermore, each sense has a set of concepts in the IS-A taxonomy (hypernym/hyponym relations). First, the concept that is common to all the senses of all the words that form the context is sought. We call this concept the Initial Specification Mark (ISM), and if it does not immediately resolve the ambiguity of the word, we descend from one level to another through WordNet´s hierarchy, assigning new Specification Marks. The number of concepts that contain the sub-hierarchy will then be counted for each Specification Mark. The sense that corresponds to the Specification Mark with highest number of words will then be chosen as the sense disambiguation of the noun in question, within its given context.
We should like to point out that after having evaluated the method, we subsequently discovered that it could be improved with a set of heuristics, providing even better results in disambiguation. The set of heuristics are Heuristic of Hypernym, Heuristic of Definition, Heuristic of Common Specification Mark, Heuristic of Gloss Hypernym, Heuristic of Hyponym and Heuristic of Gloss Hyponym. Detailed explanation of the method can be found in [MP01], while its application to NLP tasks are addressed in [Pm01].

3 Proposal for WordNet Enrichment

The classification systems provide a means of arranging information so that it can be easily located within a library, World Wide Web, newspapers, etc. Materials are usually classified by their category or class. Therefore, the field of human knowledge is divided into major categories, these are divided into subsections, and so on. The classification scheme is structured according to the state of current human knowledge.
On the other hand, WordNet presents word senses that are too fine-grained for NLP tasks. We define a way to deal with this problem, describing an automatic method to enrich semantically WordNet 1.6. with categories or classes from the classification systems using the Specification Marks Method. Categories, such as Agriculture, Health, etc, provide a natural way to establish semantic relations among word senses.

3.1 Method

In this section we describe, in detail, the method employed for enriching WordNet 1.6. The group of words pertaining to a category, that is, to be disambiguated come from different files of the classification systems. These groups of nouns are the input for the WSD module. This module will consult the WordNet knowledge base for all words that appear in the semantic category, returning all of their possible senses. The disambiguation algorithm will then be applied and a new file will be returned, in which the words have the correct sense as assigned by WordNet. After a new file has been obtained, it will be the input for the rules module. This module will apply a set of rules for finding out the main-concept in WordNet. This main-concept in WordNet is labelled with its corresponding category of the classification system. This process is illustrated in Figure 1.

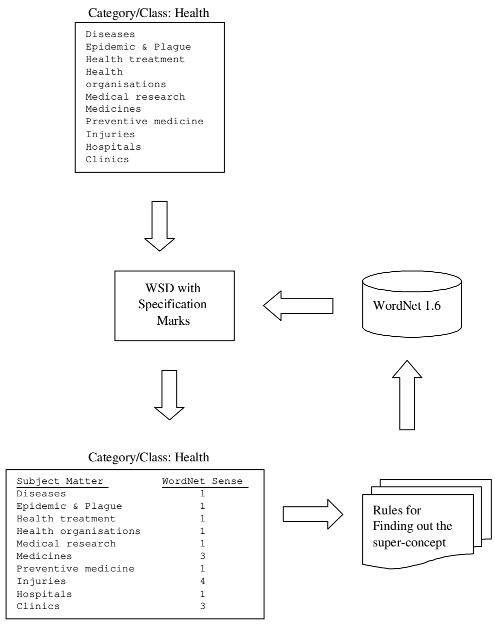
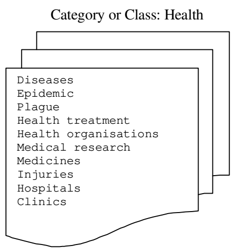


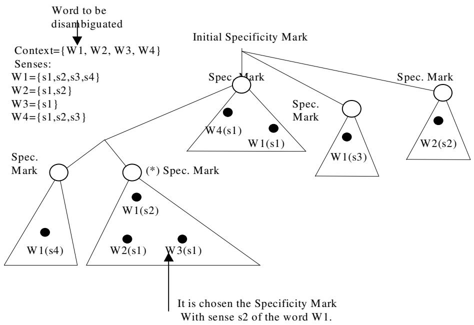
Figure 1 : Process of WordNet enrichment

The method performs the following steps to enrich and label WordNet.

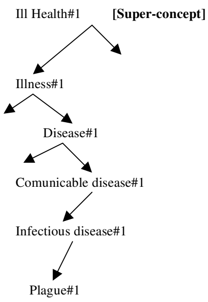
Step 1. Starting with the categories of the classification systems. We would like to clear up any ambiguities at this stage. There are words in the categories that form two words or more. These word combinations of two or more words are not in WordNet, therefore it would be impossible to disambiguate. To resolve this problem we use the utility of WordNet "Find Keywords by Substring" (grep). This substring is a synset in WordNet and relates to the words of the category. (i.e., the substring "Health organization" isn´t in WordNet but finding it with this utility we obtain the substring "Health maintenance organization").



Step 2. To locate the synset or number sense associated with each one of the words of the category, using the Specification Marks Method. For example:



Step 3. To obtain the main-concept from each category, using the hyper/hyponym relationships in the taxonomy of WordNet. For example, the main-concept for disease is *ill_health*:

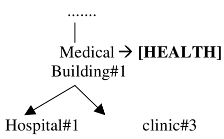


Step 4. To label the main-concept, obtained in WordNet, with the category belonging to the group of words in the classification systems. For example, the main-concept obtained in the step 3 is labeled with *Health*.

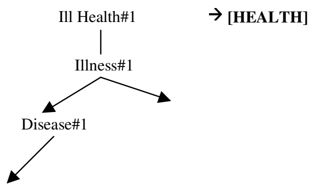
3.2 Main-Concept Rules

The way to combine the semantic categories of classification systems and WordNet would be to obtain the main-concept of WordNet for each group of words that belong to a semantic category. For obtaining these main-concepts we apply the following set of rules.

Rule 1. If a synset contains only hyponym words belonging to the category for disambiguating, it is chosen as the main-concept. The category is assigned to that main-concept as to full hyponyms and meronyms. For example, the category *Health* is made up of a group of words including *clinic* and *hospital* as it is shown in:

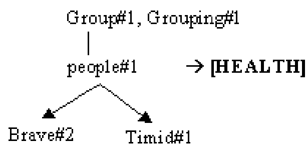


Rule 2. If the synset selected has a hypernym that is made up of the same word as the chosen entry, it is selected as the main-concept. The category is assigned to that main-concept as to full hyponyms and meronyms. For example, the synset *ill_health* is made up of *ill* and *health* and therefore it is a hypernym of *disease#1*, as it is shown in:



Rule 3. This rule resolves the problem of those words that are neither directly related in WordNet nor are in some composed synset of a hyper/hyponym relationships. We search for categories with the hyponym gloss. The hypernym of the word disambiguated is obtained in the taxonomy of WordNet. Then, all of the other words included in the category in some gloss of an immediate hyponym synset of WordNet are checked, and the label of the category is assigned to it. Also, this category label is assigned to all the hyponym and meronym relationships.

Rule 4. When the word to be disambiguated is next to the root level (one or two levels), that is, near the top of the taxonomy, this rule assigns the category only to this synset and at all its hyponyms and meronyms. For example, the category *Human Interest* is assigned to *people#1*, as it is shown in:



4 Interface

In order to build the enriched WordNet it is necessary the creation of a interface to label WordNet with categories from different available classification systems. This interface is made up of a set of computer programs that do all the work leading ultimately to a labelled lexical knowledge base of WordNet.

This section describes features of the design and implementation of the interface to obtain extensions and enhancements on the WordNet lexical database, with the goal of providing the NLP community with additional knowledge.

The design of the interface is composed of four processes: (i) selecting the classifycation systems and their categories, (ii) resolving the lexical ambiguity of each word, (iii) finding out the main-concept and (iiii) organization and format of the WordNet database. These processes are illustrate in the figure 2.

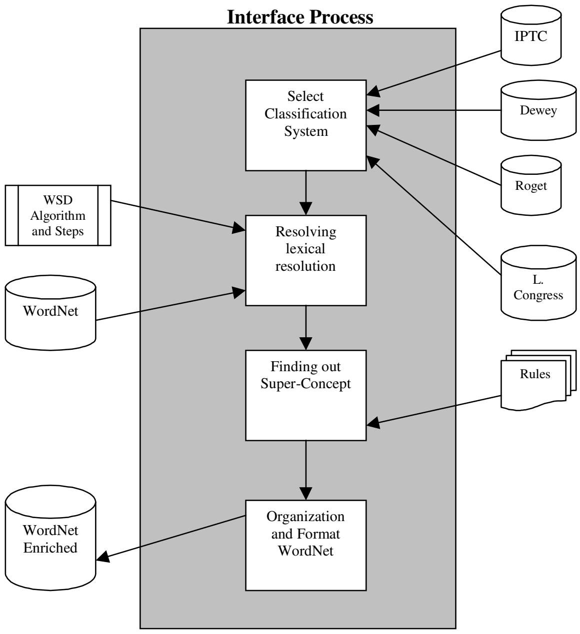


Figure 2 : Interface Process

In order to validate our study, we implemented the interface using programming language C++. It is shown in figure 3, with necessary given explanations below. And due to the physical distance between the different members of the group of investigation who use the interface, this has been developed to work through the local area network (LAN). The user interface offers the operations followed:

Select the classification system. A classification systems selection window contains option buttons. The user clicks on the appropriate button to select the desired classification system. We have considered the classification systems such as IPTC, Dewey classification, Library of Congress Classification and Roget´s.

Open category. The user clicks on this command button to select a category of the selected classification system in the previous step. The group of words that belong to the selected category appear in the left text window of the interface, named Input Category.

Run Interface. The processes, resolving the lexical ambiguity and finding out the main-concept, were implemented in a unique function. The command button Run Interface allows one to run this function, and the output information that belongs to the group of words of the selected category appear in the right text window of the interface, named Output Labelled Synsets. This output information is made up of WordNet Sense Word and Main-concept obtained for each word belonging to the category. For example:

WordNet Sense Word	Main-concept
{10129713} disease#1	{10120678} <IPTC.Health> ill Health

Save Category. If this command button is clicked, the information above is organizated, formatted and storaged in the WordNet lexical database for each main-concept, their full hyponyms and meronyms.

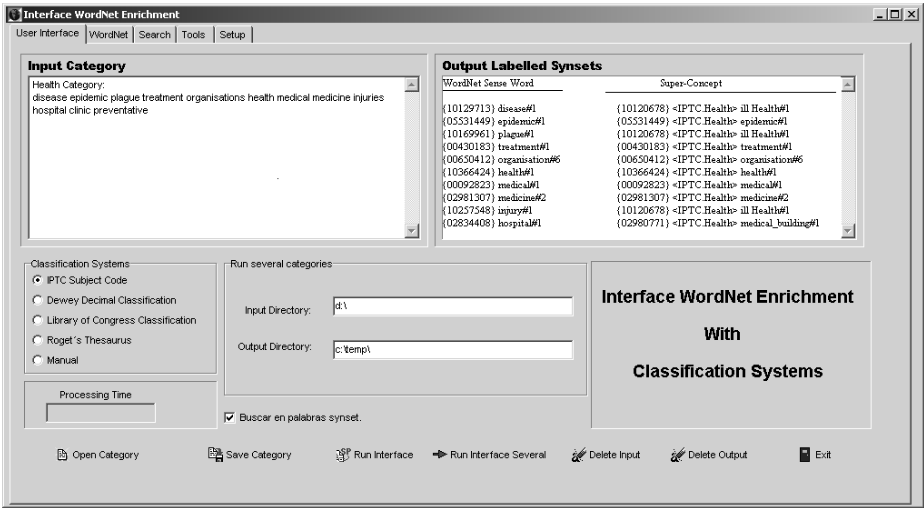


Figure 3: User Interface

5 Experiments And Results

In this section we will describe a set of experiments and the results obtained. The goal of the experiments is to assess the effectiveness of the proposed method to enrich semantically WordNet 1.6. with categories from the classification systems. A brief description of the resources used is included in this section to introduce the reader in the test environment.

5.1 Experiment 1

The first goal was to assess the effectiveness of the disambiguation of the Specification Marks method. It was carried out on random sentences taken from the Semantic Concordance Corpus (Semcor [Mg93]) and Microsoft Encarta Encyclopaedia Deluxe (Encarta), and the IPTC Subject Reference System (classification system). The method tested on IPTC Subject Reference System but this method can also be applied to other classification systems such as Library of Congress Classification(LC), Roget's Thesaurus or Dewey Decimal Classification (DDC).

These classification systems are divided in categories or classes, which are in turn subdivided into groups of words that are strongly related. In this work we intend to enrich WordNet 1.6. with synsets that have been annotated with one or more categories of the previous classification systems.

In the first approach we wanted to verify that, the Specification Mark Method can obtain successful results, and therefore this method can be applied successfully on any corpus.

The percentages of correct resolutions achieved with these two corpora were Semcor 67,4% and Encarta 65,9% respectively. We should like to make a clear distinction, however, it does not require any sort of training, no hand-coding of lexical entries, or the hand-tagging of texts. In the second approach we tested the Specification Mark Method on word clusters related by categories over IPTC Subject Reference System. The percentage of correct resolution was 96.1%. This successful percentage was because the method uses the knowledge of how many of the words in the context are grouped around a semantic class in the WordNet taxonomy. The experimental results[4] are those shown in the following table.

Texts	Corpus Names	Ratio	Values
Unrestricted Text	SEMCOR	Precision	67.4 %
		Recall	66.5 %
		Coverage	98.5 %
	ENCARTA	Precision	65.9 %
		Recall	65.1 %
		Coverage	98.8 %
Classification System	IPTC	Precision	96.1 %
		Recall	92.5 %
		Coverage	96.8 %

5.2 Experiment 2

Once it has been shown that the WSD Specification Marks Method works well with classification systems, we tested the method of combining the semantic categories of IPTC and WordNet.

Table 1 presents the totals results of the IPTC categories, computed as the amount of synsets of WordNet correctly labelled, synsets incorrectly labelled and words unlabelled (synsets are not in WordNet).

To evaluate the precision, coverage and recall of the method, we applied the rules of the section 2.2. and we hand checked the results for each word belonging to an IPTC category.

Categories IPTC	Total Number Words IPTC	Correctly Labelled Synsets	Incorrectly Labelled Synsets	Words Unlabelled
TOTAL	399	358	16	25

Table 1: Totals results of the IPTC categories

[4] Coverage is given by the ratio between total number of answered senses and total number of senses. Precision is defined as the ratio between correctly disambiguated senses and total number of answered senses. Recall is defined as the ratio between correctly disambiguated senses and total number of senses.

Precision is given by the ratio between correctly synsets labelled and total number of answered (correct and incorrect) synsets labelled. Coverage is given by the ratio between total number of answered synsets labelled and total number of words. Recall is given by the ratio between correctly labelled synsets and total number of words. The experimental results are those shown in the following table.

%	Coverage	Precision	Recall
WordNet Enrichment	93.7 %	95.7 %	89.8 %

We saw that if the Specification Mark Method disambiguates correctly and the rules of the section 2.2. are applied, the method works successfully. However, if the Specification Mark Method disambiguates incorrectly, the labelling of WordNet with categories of IPTC is also done incorrectly.

6 Conclusion and Further Work

This paper has shown the WSD Specification Marks Method to assign a category of a classification system to a WordNet synset and their descendants and meronyms. We enrich the WordNet taxonomy with categories of the classification system.
The experimental results, when the method is applied to IPTC Subject Reference System, indicate that this may be an accurate and effective method to enrich the WordNet taxonomy.
We have seen in these experiments a number of suggestive indicators. The WSD Specification Marks Method works successfully with classification systems, that is, categories subdivided into groups of words that are semantically related. Although, this method has been tested on IPTC Subject Reference Systems, but can also be applied to other systems that group words about a single category. These systems are Library of Congress Classification (LC), Roget's Thesaurus or Dewey Decimal Classification (DDC).
A relevant consequence of the application of the Method to enrich WordNet is the reduction of the word polysemy (i.e., the number of categories for a word is generally lower than the number of senses for the word). That is, category labels (i.e., HEALTH, SPORTS, etc), provide a way to establish semantic relations among word senses, grouping then into clusters. Therefore, this method intends to resolve the problem of the fined-grainedness [IV98] of WordNet´s sense distinctions.
Furthermore, now we able to perform variants of WSD systems using domain labels rather than synset labels [MS00].

References

[Aa94] Ageno A., Castellón I., Ribas F., Rigau G., Rodríguez H., and Samiotou A. 1994. TGE: Tlink Generation Environment. In proceedings of the 15th International Conference On Computational Linguistic (COLING´94). Kyoto, (Japan).

[Am87] Alvar M. 1987. Diccionario General Ilustrado de la Lengua Española VOX. Bibliograf S.A.. Barcelona, (Spain).

[Br89] Byrd R. 1989. Discovering Relationship among Word Senses. In proceedings of the 5th Annual Conference of the UW Centre for the New OED, pages 67-79. Oxford, (England).

[CC98] Chen J. and Chang J. 1998. Topical Clustering of MRD Senses Based on Information Retrieval Techniques. *Computational Linguistic* **24**(1): 61-95.

[DPR00] Daudé J., Padró L. And Rigau G. 2000. Mapping WordNets Using Structural Information. In Proceedings 38th Annual Meeting of the Association for Computational Linguistics(ACL00). Hong Kong. (Japan).

[Kk93] Knight K. 1993. Building a Large Ontology for Machine Translation. In proceedings of the ARPA Workshop on Human Language Technology, pages 185-190. Princenton.

[KL94] Knight K. and Luk S. 1994. Building a Large-Scale Knowledge Base for Machine Translation. In proceedings of the American Association for Artificial Inteligence.

[IV98] Ide N. and Véronis J. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. Computational Linguistics **24** (1): 1-40.

[MC00] Magnini B. and Cavaglia G. (2000) Integrating subject field codes into WordNet. In Proceedings of the LREC-2000, Athens, Greece.

[Mg90] Miller G. A., Beckwith R., Fellbaum C., Gross D., and Miller K. J. 1990. WordNet: An on-line lexical database. International Journal of Lexicography **3**(4): 235-244.

[Mg93] Miller G., Leacock C., Randee T. and Bunker R. 1993. A Semantic Concordance. Proc. 3rd DARPA Workshop on Human Language Tecnology, pages 303-308, Plainsboro, (New Jersey).

[MP00] Montoyo, A. and Palomar M. 2000. WSD Algorithm applied to a NLP System. In Proceedings 5th International Conference on Application of Natural Language to Information Systems (NLDB´2000). Versailles, (France).

[MP01] Montoyo, A. and Palomar, M. 2001. Specification Marks for Word Sense Disambiguation: New Development. 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001). México D.F. (México).

[MS00] Magnini B. and Strapparava C. 2000. Experiments in Word Domain Disambiguation for Parallel Texts. In Proceedings of the ACL Workshop on Word Senses and Multilinguality, Hong Kong, China.

[OH94] Okumura A. and Hovy E. 1994. Building japanese-english dictionary based on ontology for machine translation. In proceedings of ARPA Workshop on Human Language Technology, pages 236-241.

[Pm01] Palomar M., Saiz-Noeda M., Muñoz, R., Suárez, A., Martínez-Barco, P., and Montoyo, A. 2000. PHORA: NLP System for Spanish. In Proceedings 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001). México D.F. (México).

[Pp87] Procter P. 1987. Longman Dictionary of common English. Longm*an* Group. England.

[Rg94] Rigau G. 1994. An Experiment on Automatic Semantic Tagging of Dictionary Senses. In International Workshop the Future of the Dictionary. Grenoble, (France).

[RA95] Rigau G. and Agirre E.1995. Disambiguating bilingual nominal entries against WordNet. Seventh European Summer School in Logic, Language and Information (ESSLLI´95). Barcelona, (Spain).

[RRT95] Rigau G., Rodriguez H., and Turmo J. 1995. Automatically extracting Translation Links using a wide coverage semantic taxonomy. In proceedings fifteenth International Conference AI´95, Language Engineering´95. Montpellier, (France).

[Ro89] Risk O. 1989. Sense Disambiguation of Word Translations in Bilingual Dictionaries: Trying to Solve The Mapping Problem Automatically. RC 14666, IBM T.J. Watson Research Center. Yorktown Heights, (United State of America)