Terminology Retrieval: towards a synergy between thesaurus and free text searching

Anselmo Peñas, Felisa Verdejo and Julio Gonzalo

Dpto. Lenguajes y Sistemas Informáticos, UNED
{anselmo,felisa,julio}@lsi.uned.es

Abstract. Multilingual Information Retrieval usually forces a choice between free text indexing or indexing by means of multilingual thesaurus. However, since they share the same objectives, synergy between both approaches is possible. This paper shows a retrieval framework that make use of terminological information in free-text indexing. The Automatic Terminology Extraction task, which is used for thesauri construction, shifts to a searching of terminology and becomes an information retrieval task: *Terminology Retrieval.* Terminology Retrieval, then, allows cross-language information retrieval through the browsing of morpho-syntactic, semantic and translingual variations of the query. Although terminology retrieval doesn't make use of them, controlled vocabularies become an appropriate framework for terminology retrieval evaluation.

1 Introduction

The organization of information for later retrieval is a fundamental area of research in Library/Information Sciences. It concerns to understand the nature of information, how humans process it and how best to organize it to facilitate use. A number of tools to organize information have been developed, one of them is the information retrieval thesaurus. A thesaurus is a tool for vocabulary control. Usually it is designed for indexing and searching in a specific subject area. By guiding indexers and searchers about which terms to use, it can help to improve the quality of retrieval. Thus, the primary purposes of a thesaurus are identified as promotion of consistency in the indexing of documents and facilitating searching. Most thesauri have been designed to facilitate access to the information contained within one database or group of specific databases. An example is the ERIC[1] thesaurus, a gateway to the ERIC documents database containing more than 1.000.000 abstracts of documents and journal articles on education research and practice. Via the ERIC interface, one can navigate the thesaurus and use the controlled vocabulary for more accurate and fruitful search results. Thesaurus were a resource used primarily by trained librarians

[1] http://www.ericfacility.net/extra/pub/thessearch.cfm

This work has been partially supported by European Schools Treasury Browser project (ETB, http://www.eun.org/etb) and HERMES project (TIC2000-0335-C03-01).

obtaining good performance. However nowadays on-line database searching is carried out by a wider and less specialized audience of Internet users and recent studies [4] claim that most end-users obtained poor results, missing highly relevant documents. Nevertheless there is a strong feeling in the documentalist field that the use of a thesaurus is a central issue for raising the quality of end-users results [7] specially in a multilingual context where natural language ambiguity increases, producing additional problems for translingual retrieval. A multilingual thesaurus guarantees the control of the indexing vocabulary, covering each selected concept with a preferred term, a descriptor, in each language, and ensuring a very high degree of equivalence among those terms in different languages. However, multilingual thesauri construction and maintenance is a task with a very high cost, which motivates the exploration of alternative approaches based on free text indexing and retrieval.

In this paper, on one hand, we show how NLP techniques have a part to play both in thesaurus construction and in free text searching in specialized collections; on the other hand, we describe an evaluation framework for an NLP-based full-text multilingual search system where a thesaurus resource is used as a baseline. The structure of the paper is the following: Section 2 reports the developed methodology implying NLP techniques to support the construction of the European Schools Treasury Browser (ETB) multilingual thesaurus in the field of education. This methodology easily shifts to a new strategy with IR shared objectives: *terminology retrieval*. Section 3 introduces Website Term Browser (WTB) [5], a browsing system that implements this strategy for searching information in a multilingual collection of documents. The system helps users to cross language barriers including terminology variations in different languages. In order to assess the performance of WTB we have designed an evaluation framework for the *terminology retrieval* task, that takes profit of the ETB multilingual thesaurus. Section 4 presents this evaluation framework and shows the results obtained. The conclusion points out a direction in which NLP techniques can be a complement or an alternative to thesaurus based retrieval.

2 From Terminology Extraction to Terminology Retrieval

Thesaurus construction requires collecting a set of salient terms. For this purpose, relevant sources including texts or existing term lists have to be identified or extracted. This is a task combining deductive and inductive approaches. Deductive procedures are those analyzing already existing vocabularies, thesauri and indexes in order to design the new thesaurus according to the desired scope, structure and level of specificity; inductive approaches analyze the real-world vocabularies in the document repositories in order to identify terms and update the terminologies. Both approaches can be supported by automatic linguistic techniques. Our work followed the inductive approach to provide new terminology for the ETB thesaurus, starting with an automatic Terminology Extraction (TE) procedure [6]. Typically, TE (or ATR, Automatic Terminology Recognition) is divided in three steps [2], [3]:

1. Term extraction via morphological analysis, part of speech tagging and shallow parsing. We distinguish between one word terms (mono-lexical

terms) and multi-word terms (poly-lexical terms), extracted with different techniques detailed in [6].
2. Term weighting with statistical information, measuring the term relevance in the domain.
3. Term selection. Term ranking and truncation of lists by thresholds of weight.

These steps require a previous one in which relevant corpora is identified, automatically collected and prepared for the TE task. After collecting terms, documentalists need to decide which ones are equivalent, which are finally selected and which other terms should be introduced to represent broad concepts or to clarify the structure of semantic relations between terms. The main semantic relations are hierarchical (represented as BT and NT) and RT to express an associative relationship. To support documentalists decisions, a web-based interface making use of hyperlinks was provided. Through this interface, access to candidate terms contexts as well as their frequency statistics were provided.

This was the methodology employed to term extraction task and thesaurus construction. However, while the goal in the Terminology Extraction is to decide which terms are relevant in a particular domain, in a full text search users decide which are the relevant terms according to their information needs, i.e. the user query gives the relevant terms. In this case, the automatic terminology extraction task oriented to text indexing should favor recall rather than precision of the extracted terms. This implies:

1. Terminology list truncation is not convenient.
2. Relaxing of poly-lexical term patterns is possible.

And also suggests a change of strategy. From a thesaurus construction point of view, TE procedure shifts to *term searching* becoming a new task: *terminology retrieval*. From a text retrieval perspective, *retrieved terminology* becomes an intermediate information level which provides document access bridging the gap between query and collection vocabularies even in different languages. This framework, shared for both tasks, needs:

1. A previous indexing to permit phrase retrieval from query words.
2. Expansion and translation of query words in order to retrieve terminology variations (morpho-syntactic, semantic and translingual).

This strategy has been implemented in the WTB described in the next section.

3 Website Term Browser

The system, *Website Term Browser* (*WTB*) [5], applies NLP techniques to perform automatically the following tasks:

1. Terminology Extraction and indexing of a multilingual text collection.
2. Interactive NL-query processing and retrieval.
3. Browsing by phrases considering morpho-syntactic, semantic and translingual variations of the query.

Terminology Extraction and Indexing. The collection of documents is automatically processed to obtain a large list of terminological phrases. Detection of phrases in the collection is based on syntactic patterns. Selection of phrases is based

on document frequency and phrase subsumption. Such processing is performed separately for each language (currently Spanish, English, French, Italian and Catalan). We reused, in a relaxed way, the terminology extraction procedure originally meant to produce a terminological list to feed a thesaurus construction process[6].

Query processing and retrieval. Cross-language retrieval is performed by translating the query to the other languages in the collection. Word translation ambiguity can be drastically mitigated by restricting the translation of the components of a phrase into words that are highly associated as phrases in the target language [1]. This process is generalized in the Website Term Browser as follows:

1. Lemmatized query words are expanded with semantically related words in the query language and all target languages using the EuroWordNet lexical database [8] and some bilingual dictionaries.
2. Phrases containing some of the expanded words are retrieved. The number of expansion words is usually high, and the use of semantically related words (such as synonyms) produce a lot of noise. However, the retrieval and ranking of terms via phrasal information discards most inappropriate word combinations, both in the source and in the target languages.
3. Unlike batch cross-language retrieval, where phrasal information is used only to select the best translation for words according to their context, in this process all salient phrases are retained for the interactive selection process.
4. Documents are also ranked according to the frequency and salience of the relevant phrases they contain.

Browsing by phrases. *Figure 1* shows the WTB interface. Results of the querying and retrieval process are shown in two separate areas: a ranking of phrases that are salient in the collection and relevant to the user's query (on the left part) and a ranking of documents (on the right part). Both kinds of information are presented to the user, who may browse the ranking of phrases or directly click on a document.

Phrases in different languages are shown to users ranked and organized in a hierarchy according to:

1. Number of expanded terms contained in the phrase. The higher the number of terms within the phrase, the higher the ranking. In the monolingual case, original query terms are ranked higher than expanded terms.
2. Salience of the phrase according to their weight as terminological expressions. This weight is reduced to within-collection document frequency if there is no cross-domain corpus to compare with.
3. Subsumption of phrases. For presentation purposes, a group of phrases containing a sub-phrase are presented as subsumed by the most frequent sub-phrase in the collection. That helps browsing the space of phrases similarly to a topic hierarchy.

Figure 1 shows an example of searching. The user has written the English query *"adult education"* in the text box. Then, the system has retrieved and ranked related terminology in several languages (Spanish, English, French, Italian and Catalan). This terminology was extracted automatically during indexing, and now has been retrieved from the query words and their translations. In the example, the user has selected the *Spanish* tab as target language where there are three different top terms (folders): *"formación de adultos"*, *"adultos implicados en el proceso de enseñanza"* and *"educación de adultos"*. The second one (*"adultos implicados en el proceso de*

enseñanza") is not related to the concept in the query, but the *term browsing facility* permits to discard it without effort. Top term folders contain morpho-syntactic and semantic variations of terms. For example, the preferred Spanish term in the ETB thesaurus is *"educación de adultos"*. However, in this case, besides the preferred term, *WTB* has been able to offer some variations:

- *Morpho-syntactic variation*: *"educación permanente de adultos", "educación de personas adultas"*.
- *Semantic variation: "formación de adultos", "formación de personas adultas"*

In the example, the user has expanded the folder *"educación de adultos"* and has selected the term *"educación de las personas adultas"*, obtaining (on the right handside) the list of documents containing that term.

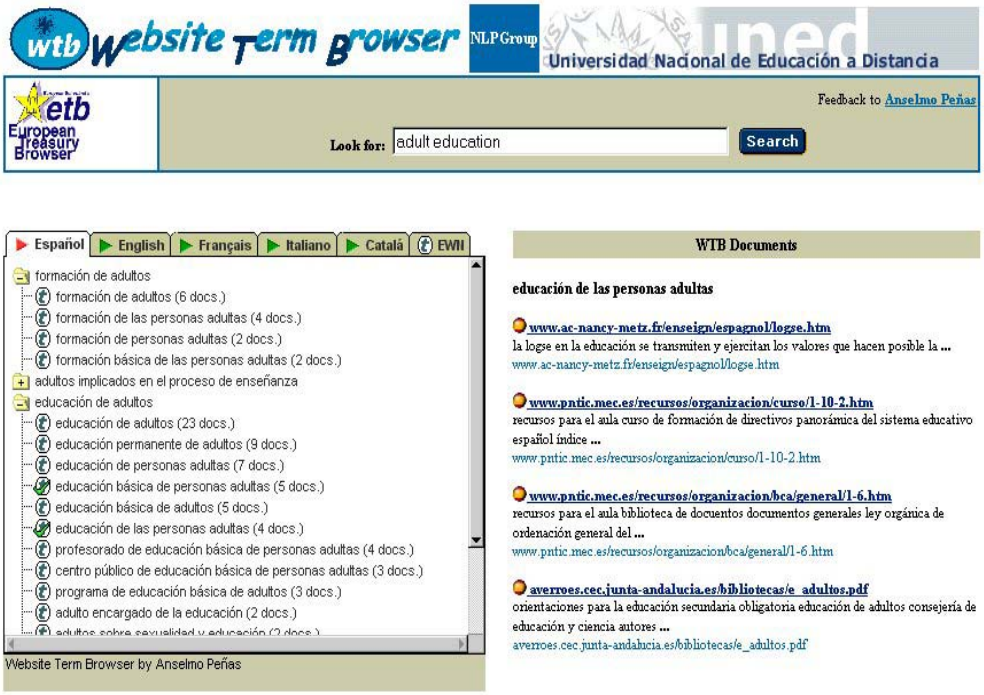

Fig. 1. Website Term Browser interface

4 Evaluation

The usefulness of *term browsing* versus document ranking was already evaluated in [5] from the users perspective. Now the evaluation is aimed to establish the system coverage for translingual terminology retrieval compared with the use of a multilingual handcrafted thesaurus for searching purposes. The second main point of this evaluation aims to study the dependence between the quality of our results, the quality of used linguistic resources and the quality of WTB processing. While NLP

techniques feed Terminology Extraction and thesaurus construction, now a thesaurus becomes a very useful resource to give feedback and evaluate the linguistic processes in a retrieval task.

The evaluation has been performed comparing the WTB terminology retrieval over a multilingual web pages collection, with the European Schools Treasury Browser (ETB) thesaurus. The multilingual collection comprises 42,406 pages of several European repositories in the educational domain (200 Mb) with the following distribution: Spanish 6,271 docs.; English 12,631 docs.; French 12,534 docs.; Italian 10,970 docs.

	ESP	ENG	FRA	ITA
	terapia	therapy	thérapie	terapia
therapy	-terapeutico -terapia -terapéutica	-therapy -treatment	-thérapie -traitement	-cura -curar -terapia -trattamento
terapia	-terapeutico -terapia -terapéutica	-therapeutics -therapy -treatment	-thérapie -traitement	-cura -curar -terapia -trattamento
thérapie	-terapeutico -terapia -terapéutica	-therapy -treatment	-thérapie -traitement	-cura -curar -terapia -trattamento
terapia	-terapeutico -terapia -terapéutica	-therapeutics -therapy -treatment	-thérapie -traitement	-cura -curar -terapia -trattamento

Fig. 2. Interface for qualitative evaluation of terminology retrieval (mono-lexical terms)

The ETB thesaurus alpha version used in the evaluation has 1051 descriptors with its translations to each of the five considered languages (English, Spanish, French, Italian and German). German hasn't been considered in the evaluation because no linguistic tools were available to us for that language. Each ETB thesaurus descriptor has been used as a WTB query. The thesaurus preferred translations have been compared with the WTB retrieved terms in each language. In such a way, precision and recall measures can be provided. Approximately half of the thesaurus descriptors are phrases (poly-lexical terms) which can be used to evaluate the WTB terminology retrieval. Thesaurus mono-lexical terms permit the coverage evaluation of linguistic resources used in the expansion and translation of query words.

Qualitative evaluation. *Figure 2* shows the interface for the qualitative evaluation. This interface is aimed to facilitate inspection on the system behavior, in order to detect errors and suggest improvements on WTB system. The first column contains the thesaurus terms in each language (in the example, *therapy, terapia, thérapie and terapia)*. Each of them are the preferred terms, or descriptors, in the thesaurus and have been used as WTB queries. The retrieved terms in each target language are shown in the same row. For example, when searching WTB with *therapy* (English term), in the first column, the system retrieves *terapeutico, terapia y terapéutica*, in Spanish (same row, second column); it also retrieves *therapy* and *treatment* in English (same row, third column).

Quantitative evaluation. If the preferred term in the thesaurus has been retrieved by WTB, then it is counted as a correctly retrieved term. Then, precision and recall measures can be defined in the following way:

- *Recall*: number of retrieved descriptors divided by the number of descriptors in the thesaurus.
- *Precision*: number of retrieved descriptors divided by the number of retrieved terms.

Figure 2 shows that there are correct terms retrieved by WTB different from the preferred terms (descriptors) in the thesaurus. Hence, the proposed recall and precision measures are lower bounds to the real performance. For example, among the retrieved terms by the English query *"adult education"*, only the Spanish term *"educación de adultos"* adjusts to the preferred term in the thesaurus. However, there are some morpho-syntactic variations *("educación de adultas", "educación de los adultos*"), semantic variations *("formación de adultos")*, and related terms *("formación básica de las personas adultas")* which are correctly retrieved terms but not counted as such.

WTB retrieved terms have been directly extracted from texts and, for that reason, recall will depend on the coverage of thesaurus descriptors in the test collection. Although the test collection is very close to the thesaurus domain, it's not possible to guarantee the presence of all thesaurus terms in all languages in the collection. Indeed, thesaurus descriptors are indexes to abstract concepts, which are not necessarily contained in the texts being indexed. *Table 1* shows the coverage of thesaurus descriptors in the test collection where exact matches have been considered (including accents).

Table 1. Thesaurus descriptors in the test collection

Coverage	Spanish	English	French	Italian
Mono-lexical descriptors found in the collection	84.3%	81.9%	82.3%	81.1%
Poly-lexical descriptors found in the collection	56.5%	57.5%	54.2%	42.6%

Mono-lexical term retrieval. Since mono-lexical term expansion and translation only depend of lexical resources, potential retrieval capabilities can be evaluated independently of the collection, just counting the mono-lexical thesaurus descriptors present in the lexical resources used (EuroWordNet lexical database and bilingual dictionaries). This comparison gives and idea of the domain coverage by the lexical resources. *Table 2* shows presence of thesaurus descriptors in the lexical resources (monolingual case, in diagonal) and their capability to go cross-language. The first column corresponds to the source languages and the first row corresponds to the target languages. The cell values correspond to the percentage of mono-lexical thesaurus descriptors recovered in the target language from the source language descriptor. *Table 2* shows that recall for the Spanish/ English pairs is significantly higher than the rest. The reason is that Spanish and English languages have been complemented with bilingual dictionaries while French and Italian only use EuroWordNet relations. Since monolingual cases show a good coverage, numbers point out that there is a lack of connections between different language hierarchies in

EuroWordNet. In conclusion, with the currently used resources, we can expect a poorer behavior of WTB translingual retrieval implying French and Italian.

Table 2. Potential recall of mono-lexical descriptors with WTB lexical resources

Recall	Spanish	English	French	Italian
Spanish	91.6%	83.7%	60.9%	64.3%
English	80.4%	97.2%	63.9%	63.9%
French	66.3%	61.8%	85.5%	55.9%
Italian	67.9%	62.2%	53.9%	96.7%

Poly-lexical term retrieval. WTB poly-lexical term retrieval depends of the previously extracted phrases from the document collection and therefore, depends on the coverage of thesaurus descriptors in the test collection. Coverage of thesaurus descriptors in the test collection in the monolingual case (*Table 1, last row*), gives an upper bound for recall in the translingual cases. *Table 3* show WTB recall for each pair of languages in percentage over this upper bound for the target language.

Table 3. WTB recall in % respect collection coverage (poly-lexical terms)

Recall	Spanish	English	French	Italian
Spanish	63.1%	45.8%	19.9%	16.3%
English	40.2%	66.5%	14.7%	7.4%
French	12.5%	15.6%	40.3%	7.8%
Italian	17.1%	17.2%	8.9%	39.3%

As shown in *Table 3* English/ Spanish pairs show better behavior than other pairs of languages. The reason for this relies in that poly-lexical term retrieval is based in the combination of mono-lexical terms, and this depends on the lexical resources used. Again, just in the case of English/ Spanish pairs, EuroWordNet has been complemented with bilingual dictionaries and, for that reason, these pairs of languages present the best behavior in both mono and poly-lexical term retrieval. However, differences between mono and poly-lexical terms recall need further consideration. While mono-lexical terms correspond to nouns, which are well covered by EuroWordNet hierarchies, most poly-lexical terms include adjective components which aren't covered so well by EuroWordNet. This lack has been also corrected only for English/ Spanish pairs using bilingual dictionaries and this is an additional factor for a better recall.

The best recall is obtained for Spanish as source language. The reason relies in that, for this language, WTB uses a morphological analyzer which gives all possible lemmas for the query words. All these lemmas are considered in expansion, translation and retrieval. In this way, possible lemmatization errors are avoided both in query and texts, and increases the number of possible combinations for poly-lexical term retrieval. However, the recall values are quite low even in monolingual cases and thus, a broader study explaining loss of recall is needed. As said, WTB poly-lexical term retrieval depends on the previous extracted phrases and thus, not only depends on the test collection, but also on phrase extraction, indexing and retrieval procedures. *Table 4* shows the loss of recall due to phrase extraction and indexing procedures. There are several factors which lead to a loss of recall:

1. *Phrase extraction procedure.* Loss of recall due to not exhaustive syntactic patterns and wrong part-of-speech tagging. The loss of recall due to a wrong phrase extraction procedure is represented by the differences between first and second rows and oscillates between 2.8% for Spanish and 17.3% for French.
2. *Phrase indexing*. Loss of recall due to wrong phrase components lemmatization. The loss of recall due to wrong indexing (mainly wrong lemmatization of phrases components in texts) oscillates between 2% for English and 34% for French.
3. *Phrase retrieval*. Loss of recall due to wrong lemmatization, expansion and translation of query words, and wrong discarding in phrase selection and ranking of terms. WTB discards retrieved terms with document frequency equal to 1 in order to improve precision in the terms shown to users. This fact produces a loss of recall between 12.9% for Spanish and 36.7% for Italian.
4. *Mismatching caused by accents and case folding*. WTB doesn't need to separate documents in different languages. For this reason the loss of recall due to accents mismatching is difficult to quantify here because it produces a big confusion between languages. For example, there are lots of terms in English equal to the French ones without accents. Similar occurs between Italian and Spanish.

All this factors show that not only lexical resources must be improved, but also linguistic processing tools as lemmatizers and part-of-speech taggers.

Table 4. Loss of recall in WTB poly-lexical term retrieval by steps in the processing

Poly-lexical descriptors	Spanish	English	French	Italian
found in the collection	56.5%	57.5%	54.2%	42.6%
found among extracted phrases	54.9%	50.1%	44.8%	40.0%
(loss of recall due to phrase extraction)	(-2.8%)	(-12.9%)	(-17.3%)	(-6.1%)
retrieved with WTB	40.9%	49.1%	29.2%	26.4%
(loss of recall)	(-27.6%)	(-14.6%)	(-46.1%)	(-38%)
(loss of recall due to phrase indexing)	(-25.5%)	(-2%)	(-34.8%)	(-34%)
retrieved with WTB discarding df=1	35.6%	38.2%	21.8%	16.7%
(loss of recall)	(-36.9%)	(-33.5%)	(-59.7%)	(-60.7%)
(loss of recall due to phrase selection)	(-12.9%)	(-22.1%)	(-25.3%)	(-36.7%)

Regarding precision, in the worst case, there is one preferred descriptor in average among ten retrieved terms, and three in the best case. Term discrimination is an easy and very fast task which is helped in the WTB interface through the term organization into hierarchies. In fact, about 70% of the retrieved relevant descriptors are retrieved in the top level of the hierarchies. This is a good percentage to ensure fast discrimination of retrieved terms.

5 Conclusions

Terminology Retrieval gives a shared perspective between terminology extraction and cross-language information retrieval. From thesaurus construction point of view, the Automatic Terminology Extraction procedures shift to term searching. From text retrieval perspective, *retrieved terminology* becomes an intermediate information level which provides document access crossing the gap between query and collection

vocabularies even in different languages. This strategy has been implemented in the Website Term Browser. The evaluation framework for *terminology retrieval* has been established in this paper, being able to detect where processing and resources can be improved. While NLP techniques feed Automatic Terminology Extraction for thesaurus construction, now, in a retrieval framework, a thesaurus provides a baseline for *terminology retrieval* evaluation and gives feedback on the quality, coverage and use of the linguistic tools and resources.

The qualitative evaluation shows that WTB is able to retrieve a considerable amount of appropriate term variations not considered in the thesaurus. Thus, terminology retrieval becomes a very good complement to thesauri in the multilingual retrieval task. The quantitative evaluation results are a lower bound of the real recall and precision values because correct term variations, different from the preferred thesaurus descriptors, are not taken into account. Results show a high dependence of WTB terminology retrieval with respect to the used linguistic resources showing that EuroWordNet relations between different languages must be improved. Results also show the loss of recall due to phrase extraction, indexing and retrieval processes. Future work must study the loss of recall due to accent mismatching. We conclude that, when appropriate resources and linguistic tools are available, WTB show a reasonable good behavior, although there is place for improvement.

Future work will refine the evaluation framework and include the study of infrequent thesaurus descriptors (especially those not found in the collection). For these purposes, the construction of a new test collection is planned querying Internet search engines with the thesaurus descriptors. The crawling of the listed documents will ensure a collection with a thesaurus coverage of 100% and will permit separate processing and results for each language including accent mismatching evaluation.

References

1. Ballesteros, L. and Croft W. B. Resolving Ambiguity for Cross-Language Information Retrieval. Proceedings of the 21st ACM SIGIR Conference. 1998
2. Bourigault, D. Surface grammatical analysis for the extraction of terminological noun phrases. Proceedings of 14th International Conference on Computational Linguistics, COLING'92. 1992; 977-981.
3. Frantzi, K. T. and S. Ananiadou. The C-value/NC-value domain independent method for multiword term extraction. Natural Language Processing. 1999; 6(3)
4. Hertzberg, S. and Rudner L. The quality of researchers' searches of the ERIC Database. Education Policy Analysis Archives. 1999.
5. Peñas, A. Gonzalo J. and Verdejo F. Cross-Language Information Access through Phrase Browsing. Proceedings of NLDB 2001, Madrid, Lecture Notes in Informatics (LNI), (GI-Edition). 2001; P-3:121-130.
6. Peñas, A. Verdejo F. and Gonzalo J. Corpus-based Terminology Extraction applied to Information Access. Corpus Linguistics 2001; Lancaster, UK. 2001.
7. Trigari, M. Multilingual Thesaurus, why? E. Schools Treasury Browser. 2001.
8. Vossen, P. Introduction to EuroWordNet. Computers and the Humanities, Special Issue on EuroWordNet. 1998.