Website Term Browser:
Overcoming language barriers in text retrieval

Anselmo Peñas, Felisa Verdejo and Julio Gonzalo

Dpto. Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia (UNED)
http://nlp.uned.es
Ciudad Universitaria, s/n
28040 Madrid, Spain
{anselmo,felisa,julio}@lsi.uned.es

Abstract: Current search systems fail to satisfy users when the relevant information is written in a foreign language; when the user is not aware of the relevant -perhaps specialized - terminology for a given topic; or when the user need is fuzzy and requires assisted search once inside an appropriate web portal. This paper describes an interactive multilingual search system that alleviates such limitations, through the browsing of phrases in different languages after being automatically extracted from the text collection. The evaluation of WTB has been focussed in two aspects: the capability to offer translingual terminology to users, and the usefulness of phrase browsing. In this sense, the evaluation shows that users consider the new level of terminological information useful, as it complements the traditional document ranking outcome.

1 Introduction

The organization of information for later retrieval is a fundamental area of research in Library/Information Sciences. It is related to understand the nature of information, the way humans process it, and to find optimal ways of organizing and storing it to facilitate its usage. A number of conceptual tools to organize information have been developed, one of them being the information retrieval thesaurus. A thesaurus is a tool for vocabulary control, and it is usually 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 enhancement of the search process. 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.

Thesaurus were a resource used primarily by trained librarians 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 [6] 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 [9] specially in a multilingual context where natural language ambiguity increases, producing additional problems for translingual retrieval. 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.

Current search engines based on free text indexing are quite efficient at finding precise information, but there are still a number of common searching scenarios which are not properly supported. We will discuss the following three scenarios:

1. The requested information is available only in a foreign language.
2. The user is not aware of the appropriate wording for the search..
3. The user need is vague or not completely defined.

The requested information is available only in a foreign language

Even if the user is able to read documents in some foreign language(s) (*passive vocabulary*) he might not be able to formulate adequate queries in such language(s) (*active vocabulary*), or he might just ignore in which language he will find the information he is seeking for.

A Spanish speaker, for instance, may understand the specifications of a handheld computer written in English; but he may not be able in advance to use the term “handheld” to retrieve updated info about leading-edge handheld computers. Using Spanish terms (“ordenador de bolsillo”) he will retrieve only documents in Spanish, probably missing a substantial amount of the most recent information.

This search scenario is very common between languages that belong to one family. A French speaker, for instance, can typically grasp the contents of Spanish, Portuguese or Italian documents, but is unable to formulate queries in any of these languages. But even for strictly monolingual users there might be a need for *cross-language* searches. Imagine a Dutch lawyer making a comparative study of local regulations in European countries regarding some specific issue. He needs, first, a cross-language search facility that retrieves regulations written (at least) in nine different languages; once he finds the documents, he can use human or machine translation services into Dutch to make the information usable for him.

The user is not aware of the appropriate wording for the search

Imagine that someone is interested in educational resources for disabled people. Such description, "educational resources for disabled people" might be perfectly understandable for a non-initiated person… but will lead to topics, such as "computer accessibility", which are only marginally related to what the user looks for. The missing piece here is a better knowledge of the specialized terminology in the field of education. The right wording to access the relevant pages is *special needs education*, so that only an expert in the field will be able to conduct precise searches. This terminology gap is not a crucial problem in (specialized) textual databases that are accessed only by trained experts, such as MEDLINE (the largest source of information in medicine). But it is a crucial problem in the web, where the diversity and depth of the information is maximal and the training of the users is minimal in average.

The user need is vague or not completely defined

Search engines are good at solving precise information needs, such as "Where can I buy soja milk online in the New York area?" or "I need a map of Rome with the highlights for tourists". But for typical, more vague requests, search engines might guide the user to an appropriate web portal, but then navigation becomes the only way of refining and changing the information need [4].

Imagine, for instance, someone that has become interested in digital cameras, and wants to get a grasp of this (for him) new technology. For the query "digital camera", a search engine retrieves a portal entitled "Digital cameras resource" as the first hit. That's a perfect match for such a general query. Once inside the portal, the user may follow a range of hyperlinks including "Forum", "News", "Buyer's guide", "FAQ", "Links", "Other sites", etc. Now the user decides to start by reading about the basics of a digital camera, narrowing his information need. Given the choices available, he starts by trying "buyer's guide". Unfortunately, this page offers detailed explanations for every relevant feature of a digital camera, but not a general introduction for newcomers. The user comes back to the portal entry and tries again, using the "FAQ" option. Again, what he gets is an unstructured set of answers to particular questions, which does not fulfil his needs. Finally, the user decides to come back to the search engine and pose a new, more defined query: "How a digital camera works". The search engine leads him to a page entitled "How digital cameras work? ", satisfying his request. What we want to highlight with this example is that, at any time, a web surfer only has hyperlink vs. search facilities, but not an adequate combination of both that permits a progressive refinement/shift of the information needs along the web browsing process. More specifically, the question is how to combine search and navigation once inside a web portal.

1.1 Interaction and language barriers

In scenarios 1 and 2, language barriers (foreign languages and specialized terminology) prevents search systems from getting at the right information. In scenario 3, search systems are lacking interactive mechanisms to help the user explore and refine his/her query according to the more specific or related topics implicitly available in the search space/portal contents.

In this article, on one hand, we show how NLP techniques have a part to play both in thesaurus-based searching and in free text searching. 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 the *Website Term Browser (WTB)*, a search/browsing system that implements this strategy for searching information in a multilingual collection of documents. *WTB* is an interactive multilingual searching facility that provides, besides documents, a set of terminological phrases related to the query as an alternative way of accessing information. Such expressions match and refine the user needs according to the contents, language and terminology in the collection.

In order to assess the performance of WTB we have designed an evaluation framework shown in Section 4. In the first part, a multilingual thesaurus has been used as a baseline to evaluate translingual terminology retrieval. In the second, usefulness of phrase browsing has been evaluated through the recording of users interaction.

2 From Automatic 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 analysing 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 analyse 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 Spanish terminology for the ETB thesaurus, starting with an automatic Terminology Extraction (TE) procedure. 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.
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 in the thesaurus. The main semantic relations in the thesaurus 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 for the term extraction task and the thesaurus construction. However, while the goal in automatic Terminology Extraction (TE) is to decide which terms are relevant in a particular domain, in a full text search, users are the ones who can decide which are the relevant terms according to their information needs: the user query gives the relevant terms. In this case, the automatic Terminology Extraction task is oriented to determine all possible candidates in texts that could match the user needs even with different wording or language. This perception changes the automatic terminology extraction methodology: the process should favour recall rather than precision of term extraction. 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, the *retrieved terminology* becomes an intermediate information level which provides document access and bridges the gap between query and collection vocabularies even in different languages.

The framework, shared for both tasks, needs:

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

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

3 The Website Term Browser

The WTB system is related to a growing research area known as *phrase browsing*, which refers to interactive systems that help the user refine his query with adequate phrases, rather than providing a static set of potentially relevant documents [11]. Comparing to other phrase browsing systems, *WTB* has the following, distinct features:

- WTB is fully multilingual, currently handling searches in (and across) Spanish, English, French, Italian and Catalan. There are only a few interactive search systems that support multilinguality [7], and none of them, to our knowledge, is based on phrase searching. Phrases are, however, ideal for cross-language searching [1], because word co-occurrence in a phrase significantly reduces the problem of translation ambiguity (words are ambiguous and usually have many possible translations, and only a few are appropriate given the context).

- Phrases handled by the system are not mere collocations (i.e., groups of words that tend to co-occur together), but nominal expressions detected with natural language processing techniques that include *morphological analysis* (to collapse inflectional variants of words), *shallow part of speech tagging* (assignment of grammatical role to ambiguous terms) and *shallow parsing* (recognition of grammatical expressions). A strength of WTB is that the use of language technologies does not compromise the scalability of the system, which has already been tested for sites up to the gigabyte scale.

- The system is able to relate phrases that do not have words in common, looking for syntactic/semantic/translingual variants of the query terms using a multilingual semantic network (EuroWordNet), co-occurrence restrictions and statistical information.

- WTB has been evaluated in a real working environment (the UNED university portal), comparing the usefulness of the phrases suggested by WTB to the document ranking returned by the Google search engine. The analysis of thousands of interactive searching sessions strongly suggests that WTB phrases are very useful for searching/browsing the UNED portal. This is a strong point for our system, as interactive retrieval systems rarely show significant differences in performance when evaluated comparatively [5].

Website Term Browser (*WTB*), applies NLP techniques to perform automatically the following tasks:

1. Phrase extraction and indexing of a multilingual text collection.
2. Query processing and retrieval.
3. Interaction through the browsing of terminological phrases considering morpho-syntactic, semantic and translingual variations of the query.

In the remainder of this section each part of the system is explained in more detail.

3.1 Terminology Extraction and Indexing

The collection of documents is automatically processed to obtain a large list of potential terminological phrases. Phrase extraction is based on matching syntactic patterns over the texts tagged on their part of speech (*Table 1*). For example, a word tagged as an adjective (e.g. *special*) followed by a word tagged as a noun *(e.g. needs)* fits with the English pattern (Adjective Noun [Noun]) (e.g. *special needs [education]*) and it is retained for further consideration. Such processing is performed separately for each language. *WTB* currently handles Spanish, English, French, Italian and Catalan.

Lemmatising (base form of words) is preferred to stemming (suffix stripping) in order to keep accurate morphology links for languages with rich inflectional morphology, and to facilitate access to the lexical resources and dictionaries. The *part of speech tagging* needed for the pattern recognition, is performed in a shallow task-oriented way in compromise with the computational cost of such processing. For example, the English and Spanish collections of the Cross-Language Evaluation Forum (1 gigabyte size*)* have been processed successfully within *Website Term Browser*. Selection of phrases is based on document frequency and phrase subsumption. The complete indexing processing follows these steps:

1. Text pre-processing and listing of words.
2. Word tagging (oriented to phrase detection).
3. Phrase detection and lemmatization of components.
4. Document indexing and frequency statistics (term and document frequencies).
5. Phrase selection with criteria based on subsumption and lexicalization degree.
6. Phrase indexing to allow phrase retrieval from lemmas and document retrieval from phrases.

3.2 Query processing and retrieval

The inference of semantic and translingual variations of the query, requires the consideration of synonyms and candidate translations of the query words. However, word ambiguity complicates this task. *Figure 1* shows an example of the query expansion and translation problem, where the ambiguity of the query terms introduces a high level of noise in the translation. In such cases, determining the appropriate translation expressions needs further processing.

One way to drastically mitigate the expansion/translation ambiguity is to consider the query as a phrase and restrict the single term translations to those candidates which co-occur in some salient phrases in the target language. The referenced phrases are those extracted in the previous step. Then, the phrase retrieval process consists in matching appropriate combinations of candidate lemmas, synonyms and translations of the original query words. The process is shown in *Figure* 2 where the steps are the following:

1. *Pre-processing*: the query is tokenised and lemmatised. All possible lemmas are retained.
2. *Query expansion and translation*: lemmas are expanded and translated with semantically related terms using the EuroWordNet lexical database [10] and some recent extensions.
3. *Phrase retrieval*: phrases containing some of the expansion terms (mainly synonyms) are retrieved. The number of expansion terms is usually high, and the use of semantically related terms (such as synonyms or meronyms) produces a lot of noise. However, the ranking via phrasal information discards most inappropriate combinations, both in the source and in the target languages.
4. *Term ranking:* unlike non-interactive cross-language retrieval, where phrasal information is used only to select the best translations for individual terms according to their context, in WTB all salient phrases are retained for the interactive selection process. The phrases are ranked according to:

- Number of expanded query words they contain,
- Weight as lexicalised expressions in terms of document frequency.
- 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.

5. *Document ranking*: documents are ranked according to the frequency and salience of the relevant phrases they contain. For the comparative evaluation of the system, this document ranking is replaced by the results of Google, offering the user both WTB phrases and Google's document ranking.

3.3 Interaction through phrase browsing

The relevant phrases in every language (which are morpho-syntactic, semantic and translingual variations of the query contained in the web pages) are selected, organised and presented to the user hierarchically, together with the documents but in a different area. The user then selects the phrase that better addresses or refines his query and obtains the pages containing such phrase (or any morpho-syntactic variant).

Figure 3 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. In the example, 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.

Figure 4 is a snapshot of the search interface over a collection of international news in English, Spanish and Catalan. The user has written a Spanish query ("tratados de prohibición de pruebas nucleares"). After the query expansion and translation, relevant phrases have been retrieved in the three languages. For instance, "test ban treaty" is an English phrase, "prohibición total de ensayos nucleares" is a Spanish phrase, and "prohibició total de proves nuclears" is a Catalan phrase. The user has selected the English phrase "nuclear non-proliferation treaty" and WTB has unfolded the corresponding sub-hierarchy of terms, altering the document ranking in order to give the user the list of documents containing that phrase. This example shows also that documents can be described with all the terminology inside the document which is close to the query.

4 Evaluation

The evaluation of WTB has been focussed in two aspects: the capability to offer translingual terminology to users, and the usefulness of phrase browsing. The first evaluation takes a multilingual thesaurus as a baseline to evaluate translingual terminology retrieval. The usefulness of phrase browsing has been evaluated recording more than 2,000 sessions in a real work environment. Both evaluations are described in the following sections.

4.1 Translingual terminology retrieval

This evaluation is first aimed to establish the system coverage for translingual terminology retrieval compared with the use of a multilingual handcrafted thesaurus for searching purposes. The evaluation also aims to study the dependence between the quality of 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.

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 (polylexical 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.

4.1.1 Qualitative evaluation

Figure 5 shows the interface for the qualitative evaluation. This interface is aimed to facilitate inspection on the system behaviour, 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).

4.1.2 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 3 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 2* shows the coverage of thesaurus descriptors in the test collection where exact matches have been considered (including accents).

4.1.3 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 3* 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 3* 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 behaviour of WTB translingual retrieval implying French and Italian.

4.1.4 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 2, last row*), gives an upper bound for recall in the translingual cases. *Table 4* show WTB recall for each pair of languages in percentage over this upper bound for the target language.

As shown in *Table 4* English/ Spanish pairs show better behaviour 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 behaviour 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 analyser which gives all possible lemmas for the query words. All these lemmas are considered in expansion, translation and retrieval. In this way, possible lemmatisation 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 required.

4.1.5 Loss of recall

As said before, 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* ___ 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 selection.* 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.
3. *Phrase indexing and retrieval*. Loss of recall due to:
    a. Wrong phrase components lemmatisation.
    b. Wrong lemmatisation, expansion and translation of query words.
    c. *Mismatching caused by accents and case folding*.

   The loss of recall due to phrase indexing and retrieval oscillates between 2% and 34.8% depending on the languages.

Regarding the lower bound of precision (correct term variation are not counted), there is one to three preferred descriptors in average among the first ten. 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 rather good percentage to ensure fast discrimination of retrieved terms.

4.2 Usefulness of phrase browsing

The evaluation of interactive retrieval systems can be an elusive task, as proved by previous TREC experiences. Previous designs to evaluate interactive cross-language systems do not suit the *Website Term Browser*, as they are devoted to evaluate a) how the systems helps choosing adequate target language terms when the user has no familiarity with that language [8], and b) how the system shows documents written in a foreign language so that the user can judge about their relevance without knowing the language [7]. The *Website Term Browser*, however, is intended for users that a) have a reasonable passive vocabulary in the target languages, but b) are not necessarily familiarised with the domain-specific terminology used in the collection of documents.

The evaluation has been conducted indexing the public documents in UNED[1] domain (*uned.es)*. This domain contains about 40,000 web pages written in Spanish (predominantly) and English (mostly containing research information). We have prepared a web interface for searching pages in this domain, that has been made available to all teachers and students in the university. Again, three areas are distinguished:

[1] Distance Learning University of Spain, http://www.uned.es

1. The query area, where the user poses a new query.
2. The document area, where the traditional document ranking is shown. In this area the users click a title to explore its page.
3. The term area, where the salient phrases related to the query are shown. In this area two additional actions can be performed:
    a. click a phrase to list the documents associated with, and
    b. click the re-consult link to use the phrase as a new query in an external search engine (currently Google).

Given a query, the search interface presents two kinds of information to users: the relevant phrases (in both languages) in the left side and, in the right hand side, the ranked documents found by the *Google*[2] search engine in the *uned.es* domain. At any time after the first query, the user can:

a) EXPLORE DOCUMENT, select a document to view its contents,
b) EXPLORE PHRASE, select a phrase to view the related documents,
c) RECONSULT WITH PHRASE, query *Google* again with any of the phrases displayed.

The hypothesis is that users will only use phrasal information when Google does not fulfil the information need directly, and one or more phrases seem useful suggestions to the user. To verify this hypothesis, all interactions with the system are logged. The interactions are grouped in sessions, where a session begins with a query to the system, continues with any combination of actions in a), b) or c), and ends when the user leaves the system or poses a new query.

Tables 6, 7 and *8* show some statistical information about the first 4731 search sessions made in the system. The results indicate that phrasal information is helpful in the searching process. After posing a query of more than one word, EXPLORE PHRASE is the first action in 54.9% of the sessions, whereas EXPLORE DOCUMENT (thus preferring Google outcome) is the first action in 38.5% of the sessions. This suggests that phrases give better expectations of relevance than *Google's* ranking.

The last row in the table provides evidence about how useful phrases are compared to *Google's* ranking. It considers sessions where the last action is a document exploration, which include successful sessions ending in an appropriate document. In 47% of that sessions, with more than 1 word queries, the previous action was a phrase selection, while in a 45.59% of them, the previous action was a *Google's* ranking. This is a strong indication that WTB phrasal information can substantially complement the document rankings provided by the standard search engines.

5 Conclusions

Terminology Retrieval gives a shared perspective between terminology extraction and multilingual information retrieval. From a thesaurus construction point of view, the Automatic Terminology Extraction procedures shift to term searching. From a text retrieval perspective, the *retrieved terminology* becomes an intermediate information level which provides document access bridging the gap between query and collection vocabularies, even across different languages. This strategy has been implemented in the Website Term Browser.

The Website Term Browser makes use of phrasal information to process queries and suggest relevant, complex terms in a fully multilingual setting. It is conceived as a tool to explore the contents of web portals, and we have shown that it may enhance search engines in three scenarios: when the relevant

[2] Google, http://www.google.com

information is written in a foreign language; when the user is not aware of the adequate terminology in the search domain; and when the user need is fuzzy and the system may help refining his needs according to the portal contents. Altogether, the WTB system is a step towards overcoming language barriers in web searches.

The system integrates simple Natural Language Processing techniques with a low computational cost: morpho-syntactic information (including lexical databases, lemmatisation, part of speech tagging and shallow parsing), multilingual semantic knowledge (via the EuroWordNet database) and implicit disambiguation of translation candidates.

The evaluation framework for the translingual *terminology retrieval* has been established being able to detect where linguistic 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 behaviour, although there is place for improvement.

The usefulness of the system has been evaluated over more than 2000 interactive retrieval sessions with real users in the UNED.es university domain. The results show that phrasal information, as suggested by the WTB system, is preferred by users as the best indicator for relevant contents in a significant percentage of searching sessions.

6 Acknowledgments

This material is based on work supported by European Schools Treasury Browser project (IST-1999-11781) and Hermes Project (TIC2000-0335-C03-01).

7 References

[1] Ballesteros, L. and Croft W. B. Resolving Ambiguity for Cross-Language Information Retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1998; 64-71.

[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. Journal of Natural Language Processing. 1999; 6(3):145-180.

[4] Hearst, M. Next generation web search: setting our sites. IEEE Data Engineering Bulleting,

Special Issue on Next Generation Web Search, Luis Gravano (Ed.). 2000.

[5] Hersh, W. Turpin A. Price S. Kraemer D. Chan B. Sacherek L. and Olson D. Do batch and user evaluations give the same results? *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. 2000; 17-24.

[6] Hertzberg, S. and Rudner L. The quality of researchers' searches of the ERIC Database. Education Policy Analysis Archives. 1999.

[7] Oard, D. Evaluating Cross-Language Information Retrieval: Document selection. Cross-Language Information Retrieval and Evaluation: Proceedings of CLEF'2000: Springer-Verlag; 2001.

[8] Ogden, W. Cowie J. Davis M. Ludovik E. Molina-Salgado H. and Shin H. Getting information from documents you cannot read: an interactive cross-language text retrieval and summarisation system. Joint ACM DL/SIGIR Workshop on Multilingual Information Discovery and Access: 1999.

[9] Trigari, M. Multilingual Thesaurus, why? European Schools Treasury Browser. 2001.

[10] Vossen, P. Introduction to EuroWordNet. Computers and the Humanities, Special Issue on EuroWordNet. 1998.

[11] Wacholder, N. and Manning N. The technology of Phrase Browsing apllications. SIGIR FORUM. 2001; 35(1):18.

8 Tables

1. N N	1. A N [N]	N: noun
2. N A	2. N N [N]	A: adjective
3. N [A] Prep N [A]	3. A A N	Prep: preposition
4. N [A] Prep Art N [A]	4. N A N	Art: article
5. N [A] Prep V [N [A]]	5. N Prep N	V: infinitive

Table 1. Syntactic patterns for Spanish and English

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%

Table 2. Thesaurus descriptors present in the text collection

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%

Table 3. Potential recall of mono-lexical descriptors according to the WTB lexical resources

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%

Table 4. Recall of poly-lexical descriptors in translingual retrieval

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 and retrieval)	(-25.5%)	(-2%)	(-34.8%)	(-34%)
retrieved with WTB discarding doc.freq.=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%)

Table 5. Loss of recall in WTB poly-lexical term retrieval

4731 Sessions			
	Sessions without query (empty):	4.71%	
	Sessions without interaction:	46.29%	
	Sessionswith interaction:	48.99%	
		EXPLORE DOCUMENT is used in	74.63%
		RECONSULT WITH PHRASE is used in	16.00%
		EXPLORE PHRASE is used in	64.71%

Table 6. Recorded sessions over WTB and distribution of actions

	all queries (2318)	1 word queries (886)	**> 1 word queries** **(1432)**
EXPLORE DOCUMENT	42%	47%	**39%**
EXPLORE PHRASE	51%	45%	**55%**
RECONSULT WITH PHRASE	7%	8%	6%

Table 7. First action after query

	all queries (1429)	1 word queries (567)	**> 1 word queries** **(862)**
Google ranking	50%	57%	**46%**
EXPLORE PHRASE	44%	38%	**47%**
RECONSULT WITH PHRASE	6%	5%	7%

Table 8. Source of last document explored

9 Figures

Query	Tratados de	Prohibición de	Pruebas	Nucleares
Expansion Terms	acuerdo capitulación concertación convenio cuidar, pacto manejar procesar	embargo entredicho interdicción interdicto proscripción	cata, catadura degustación ensayo escandallo experimento gustación muestreo, tanteo	Nuclear
Translation Terms	accord discourse handle manage pact process treat treatise treaty	ban interdiction prohibition proscription	demonstrate establish, exhibit experiment experimentation fall, fitting indicate, point present, proof prove, run sample, sampling shew, show, taste test, trial, try	Nuclear
Translated Query	Nuclear fitting interdiction manage? Nuclear taste proscription process? Nuclear test ban treaty?			

Figure 1.Ambiguity in query expansion and translation

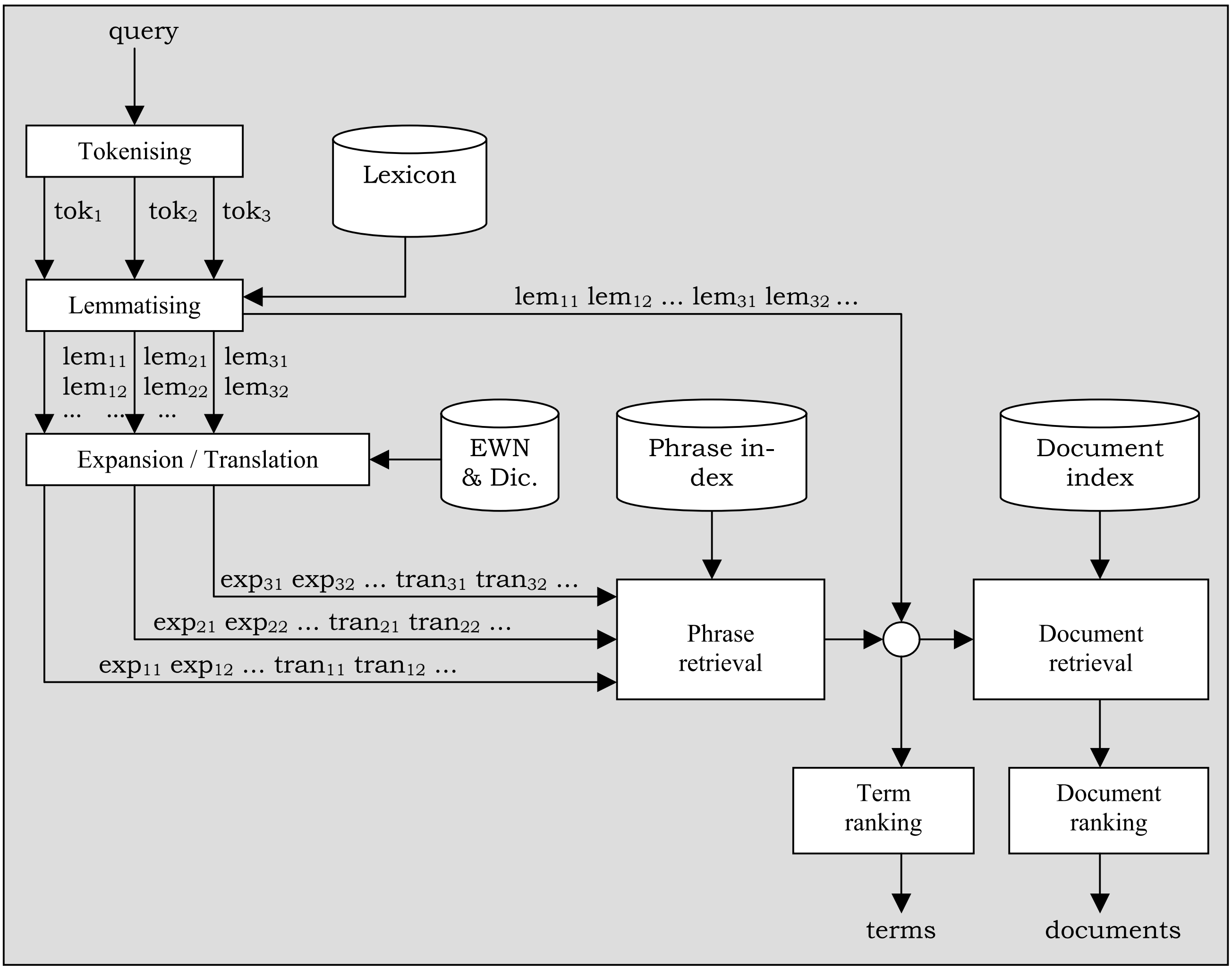
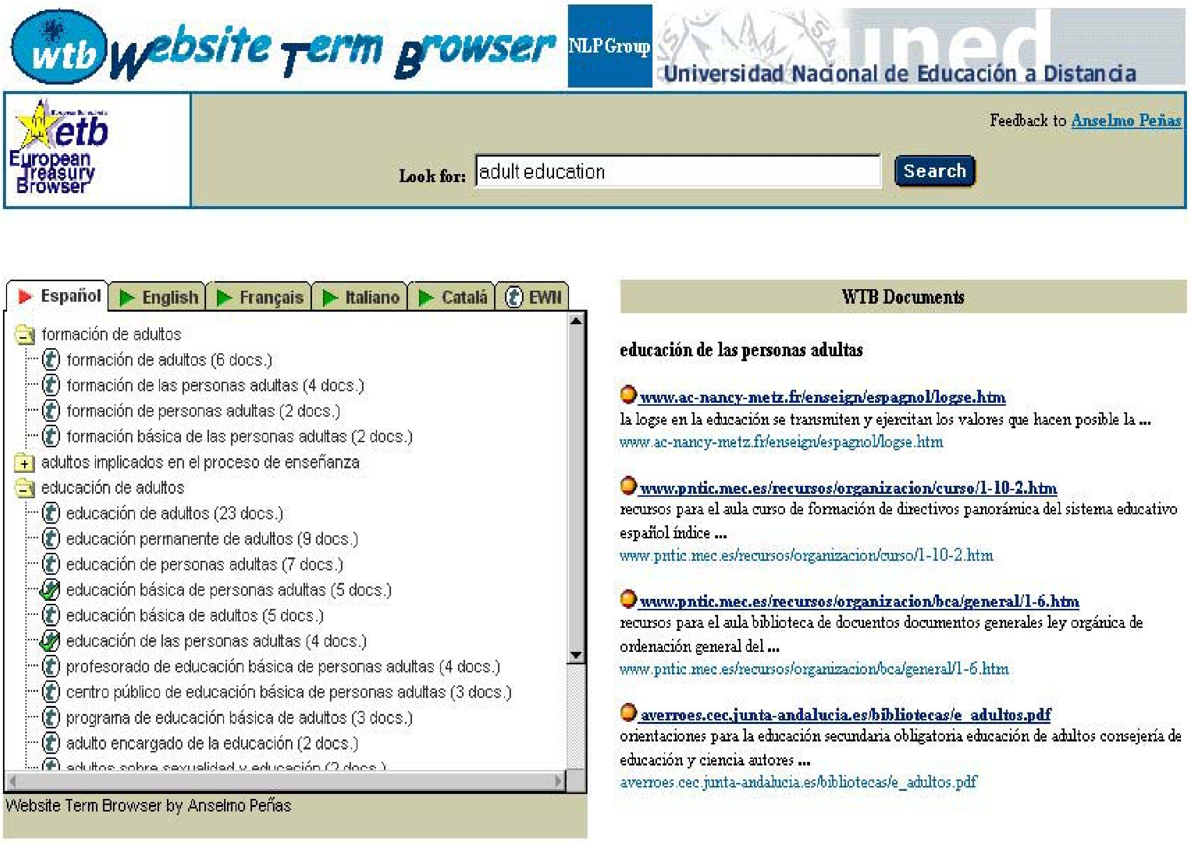


Figure 2. Retrieval process in Website Term Browser


wtb Website Term Browser
NLP Group
uned Universidad Nacional de Educación a Distancia
etb European Treasury Browser
Feedback to Anselmo Peñas
Look for:
Search
Español
English
Français
Italiano
Catalá
EWN
WTB Documents
Website Term Browser by Anselmo Peñas

Look for: adult education

- formación de adultos
  - formación de adultos (6 docs.)
  - formación de las personas adultas (4 docs.)
  - formación de personas adultas (2 docs.)
  - formación básica de las personas adultas (2 docs.)
- adultos implicados en el proceso de enseñanza
- educación de adultos
  - educación de adultos (23 docs.)
  - educación permanente de adultos (9 docs.)
  - educación de personas adultas (7 docs.)
  - ✓ educación básica de personas adultas (5 docs.)
  - educación básica de adultos (5 docs.)
  - ✓ educación de las personas adultas (4 docs.)
  - profesorado de educación básica de personas adultas (4 docs.)
  - centro público de educación básica de personas adultas (3 docs.)
  - programa de educación básica de adultos (3 docs.)
  - adulto encargado de la educación (2 docs.)
  - adultos sobre sexualidad y educación (2 docs.)

educación de las personas adultas

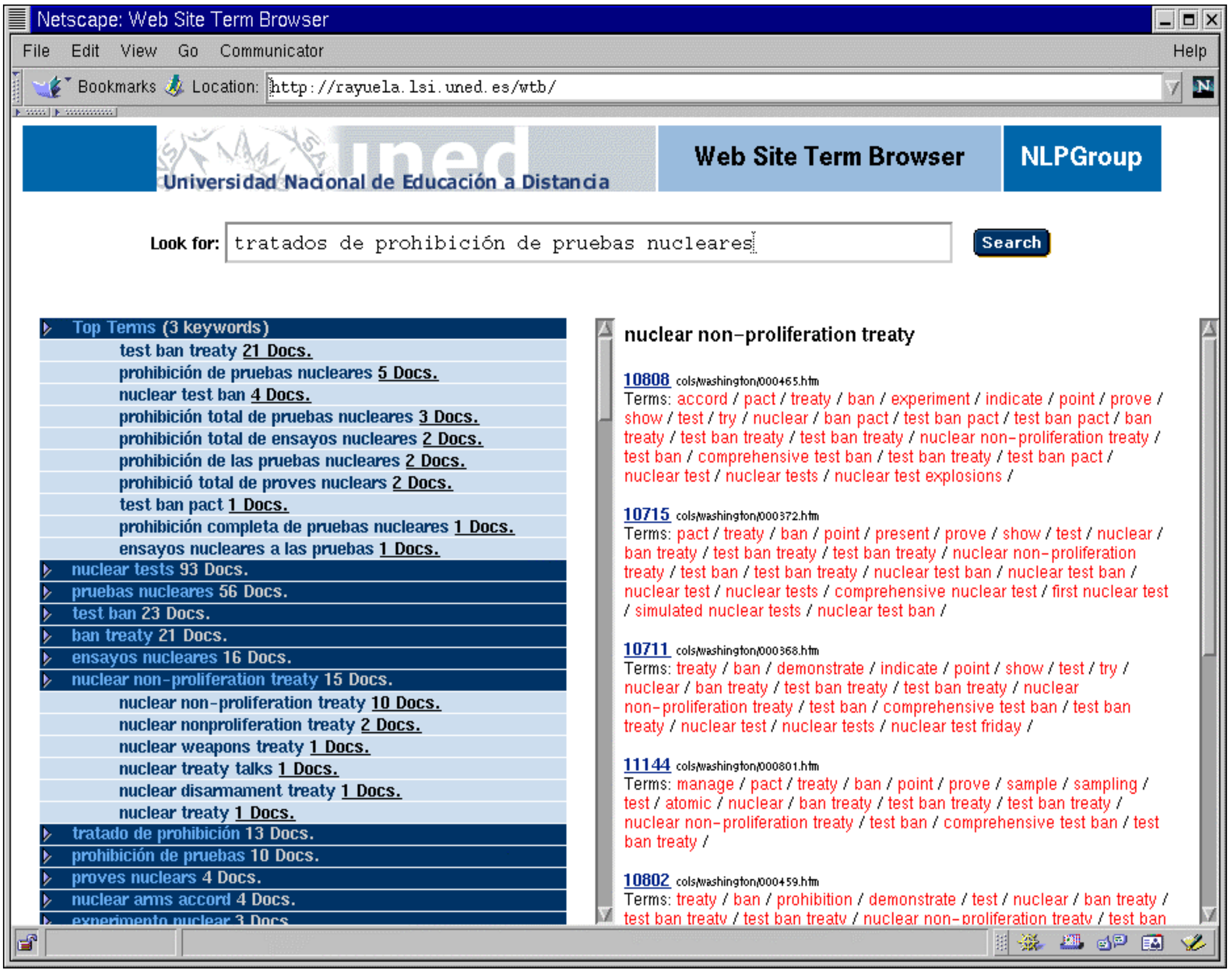
● **www.ac-nancy-metz.fr/enseign/espagnol/logse.htm**
la logse en la educación se transmiten y ejercitan los valores que hacen posible la ...
www.ac-nancy-metz.fr/enseign/espagnol/logse.htm

● **www.pntic.mec.es/recursos/organizacion/curso/1-10-2.htm**
recursos para el aula curso de formación de directivos panorámica del sistema educativo español índice ...
www.pntic.mec.es/recursos/organizacion/curso/1-10-2.htm

● **www.pntic.mec.es/recursos/organizacion/bca/general/1-6.htm**
recursos para el aula biblioteca de docuentos documentos generales ley orgánica de ordenación general del ...
www.pntic.mec.es/recursos/organizacion/bca/general/1-6.htm

● **averroes.cec.junta-andalucia.es/bibliotecas/e_adultos.pdf**
orientaciones para la educación secundaria obligatoria educación de adultos consejería de educación y ciencia autores ...
averroes.cec.junta-andalucia.es/bibliotecas/e_adultos.pdf

Figure 3. Website Term Browser interface for web pages in education



Netscape: Web Site Term Browser
File Edit View Go Communicator Help
Bookmarks
Location: http://rayuela.lsi.uned.es/wtb/
Universidad Nacional de Educación a Distancia
Web Site Term Browser
NLPGroup
Search

Look for: tratados de prohibición de pruebas nucleares

Top Terms (3 keywords)

- test ban treaty 21 Docs.
- prohibición de pruebas nucleares 5 Docs.
- nuclear test ban 4 Docs.
- prohibición total de pruebas nucleares 3 Docs.
- prohibición total de ensayos nucleares 2 Docs.
- prohibición de las pruebas nucleares 2 Docs.
- prohibició total de proves nuclears 2 Docs.
- test ban pact 1 Docs.
- prohibición completa de pruebas nucleares 1 Docs.
- ensayos nucleares a las pruebas 1 Docs.

nuclear tests 93 Docs.

pruebas nucleares 56 Docs.

test ban 23 Docs.

ban treaty 21 Docs.

ensayos nucleares 16 Docs.

nuclear non-proliferation treaty 15 Docs.

- nuclear non-proliferation treaty 10 Docs.
- nuclear nonproliferation treaty 2 Docs.
- nuclear weapons treaty 1 Docs.
- nuclear treaty talks 1 Docs.
- nuclear disarmament treaty 1 Docs.
- nuclear treaty 1 Docs.

tratado de prohibición 13 Docs.

prohibición de pruebas 10 Docs.

proves nuclears 4 Docs.

nuclear arms accord 4 Docs.

experimento nuclear 3 Docs.

nuclear non-proliferation treaty

10808 cols/washington/000465.htm
Terms: accord / pact / treaty / ban / experiment / indicate / point / prove / show / test / try / nuclear / ban pact / test ban pact / test ban pact / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban / comprehensive test ban / test ban treaty / test ban pact / nuclear test / nuclear tests / nuclear test explosions /

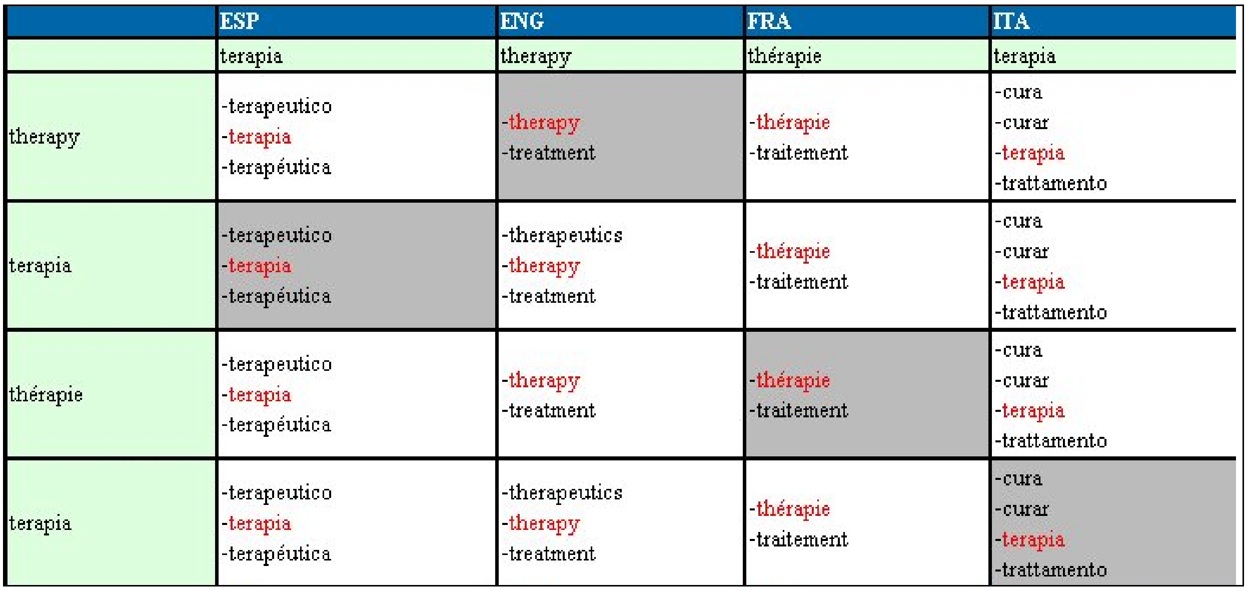
10715 cols/washington/000372.htm
Terms: pact / treaty / ban / point / present / prove / show / test / nuclear / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban / test ban treaty / nuclear test ban / nuclear test ban / nuclear test / nuclear tests / comprehensive nuclear test / first nuclear test / simulated nuclear tests / nuclear test ban /

10711 cols/washington/000368.htm
Terms: treaty / ban / demonstrate / indicate / point / show / test / try / nuclear / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban / comprehensive test ban / test ban treaty / nuclear test / nuclear tests / nuclear test friday /

11144 cols/washington/000801.htm
Terms: manage / pact / treaty / ban / point / prove / sample / sampling / test / atomic / nuclear / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban / comprehensive test ban / test ban treaty /

10802 cols/washington/000459.htm
Terms: treaty / ban / prohibition / demonstrate / test / nuclear / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban

Figure 4. Website Term Browser interface for a Multilingual News Repository



	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

Figure 5. Interface for qualitative evaluation of translingual terminology retrieval