[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

Exploiting Lexical Resources and Linguistic Tools in Cross-Language Information Retrieval: the EuroSearch approach

Eugenio Picchi[1], Carol Peters[2]

[1] Istituto di Linguistica Computazionale, CNR, Pisa, Italy

[2] Istituto di Elaborazione della Informazione, CNR, Pisa, Italy

Abstract

Cross-language Information Retrieval (CLIR) is a very new research area in which methodologies and tools developed for Natural Language Processing (NLP) are being integrated with techniques and results coming from the Information Retrieval (IR) field. The EuroSearch project provides an excellent testbed for the application and testing of different kinds of cross-language retrieval methodologies. In EuroSearch, a federation of industrial search engines (Arianna for Italy, Ole for Spain and EuroSpider for Switzerland), using different kinds of search mechanisms, will provide their users with functionalities for cross-language querying. The paper describes and motivates the technology that will be adopted in the implementation of the multilingual interface of the federation, focusing on an approach that enhances the potential of a lexicon-based search through the integration of a corpus-based methodology.

1. Introduction

The rapid growth of the Information Society has meant that vast amounts of information of all types - scientific, economic, literary, news, etc. - are now readily available over the networks and, in particular, through the Internet and the World Wide Web. However, both providers and seekers of information on the Web who are not, or who are non-native, English speakers are relatively disadvantaged compared to their English-speaking counterparts. The EuroSearch project aims to restore the linguistic and cultural equilibrium on the Web by building a pan-European federation of national search and categorization services. Initially comprising services from Italy, Spain and Switzerland, EuroSearch will provide a multilingual searching service, permitting users to enter queries in their own, or their preferred language, and to carry out search and information retrieval over some or all of the federation's national sites. Each national site will be responsible for maintaining and operating a search service dedicated to its own language(s), so that the needs of each language community will be catered for by the native speakers of that language. The project thus seeks to make a concrete contribution to the realisation of an open, multilingual Information Society by multiplying and extending access to Web information in and across a number of European languages[1].

Up to now, the study of efficient query mechanisms for Web search engines has been a priority of the Information Retrieval world. However, with the growth in awareness of the need for multilingual access functionalities on the global networks, the importance of methodologies and resources developed by the language engineering community is gaining recognition. In this paper, we present the strategy being adopted to develop the multilingual component of the EuroSearch project. This involves an integration of resources and tools developed for NLP tasks (e.g. lexical databases, morphologies, procedures for mono- and bilingual corpus management and analysis) with techniques typical of Information Retrieval.

The paper is organised as follows. Section 2 will explain why the particular configuration of the EuroSearch federation has determined the decision to adopt two complementary cross-language search strategies: lexicon-based and similarity thesaurus techniques, and Section 3 will describe how this decision is implemented. In Sections 4 and 5, we focus on the lexicon-based methodology and its enhancement through a corpus-based technique. This component will be used for global Web searches. It depends heavily on pre-constructed lexical data archives and already existing linguistic tools developed for NLP purposes and our aim, in this paper, is to show how such resources can find important applications in CLIR. The final section will briefly present the other cross-language search component implemented by the federation. This uses similarity thesaurus technology, which is based on ideas first studied in monolingual IR. It will be employed in the project for searching domain-specific collections.

2. The EuroSearch Federation

The EuroSearch federation will initially comprise services from Italy (with Arianna), Spain (through Ole) and Switzerland (using EuroSpider). The languages involved are thus Italian, Spanish, French and German plus English, as it is recognized that these three countries all produce a considerable amount of Web documentation in English and that the users of the federation may want to access English documents. It is hoped that new partners will join the federation in the future.

[1] Eurosearch (LE4-8303), an eighteen month project of the Language Engineering programme of the European Commission, began in January 1998. The industrial partners are Italia On-Line - Pisa (Coordinator), CINET - Barcelona, EuroSpider Information Technology - Zurich; the academic members are the Italian National Research Council (CNR) - Pisa, and the University of Dortmund.

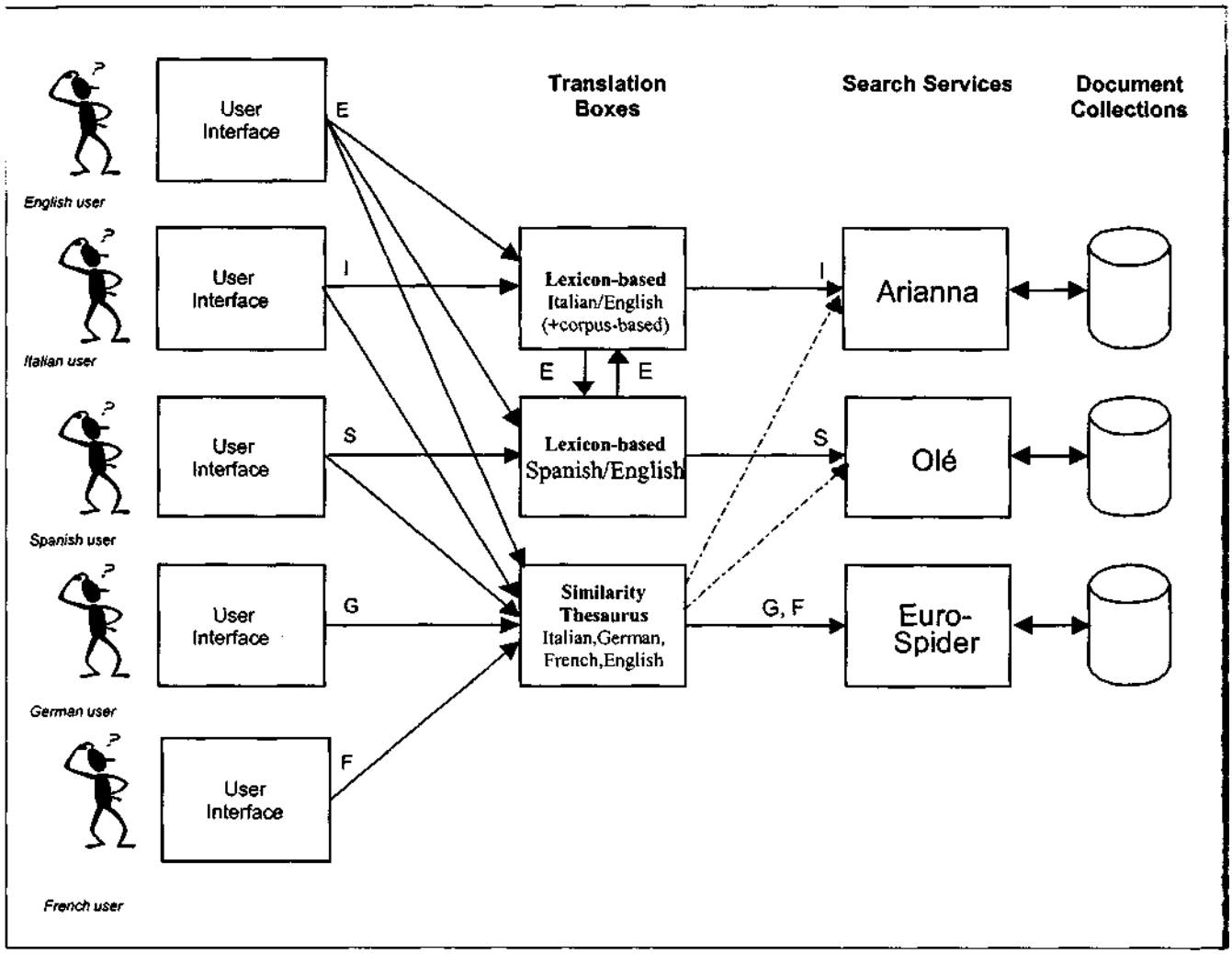

Figure 1: Logical Flow between Users, Cross-Language Components, Search Services and Document Collections.

The industrial project partners provide different kinds of retrieval services: Arianna and EuroSpider offer functions for full-text querying - Arianna through word-form indexes and EuroSpider with a stemmed index - whereas Ole provides a catalog-based service and querying is through indexed keywords. The information space covered by the partners also differs: while Arianna manages a search engine and catalog covering the Italian Web space and Ole runs a subject-oriented catalog of Spanish sites, EuroSpider provides multilingual access to vertical domain databases such as the Intranet of the Swiss Federal Institute of Technology, documents from the Swiss Federal Court, or collections of news documents.

The differences in the partners' document collections and indexing mechanisms means that it is necessary to implement different multilingual search strategies, depending on the collection to be queried. We have to be able to search efficiently both over collections covering the entire range of information available on the Web sites of a nation, and in highly domain-specific collections. Therefore, in order to optimise retrieval performance and in accordance with recent trends in the field of cross-language information retrieval, the multilingual search component of EuroSearch will consist of an integration of lexicon- and corpus-based search mechanisms (see (Oard & Dorr, 1996; Oard, 1997) for a survey of recent work in the field). Two distinct types of searching will be activated: a term-based search using a multilingual lexicon from which target language (L2) translation equivalences are identified for each term entered in a local language query (L1) and searched in the target language Web documents; more sophisticated search mechanisms which use information on cross-language equivalences extracted from multilingual document bases in order to retrieve L2 documents on the basis of an L1 query.

3. Cross-Language Querying in EuroSearcb

Figure 1 shows the logical flow between the users of the federation, the cross-language components and **the** document collections. Users can select their preferred interface and query language. Each site will have a local "translation box" installed. This "box" will be configured according to the requirements of the local search engine. In the case of lexicon-based searches, the translation boxes will translate queries between the local language and the pivot language (English)2. The translated query will then be sent to the site of the document collection to be queried, where (unless the document collection is in English) it will then be translated from the pivot language to the target language. The figure shows the current implementation which permits cross-language term-based querying in Italian or Spanish over document collections in Italian, Spanish and also

2 We have adopted the pivot language concept in order to facilitate the insertion of additional query languages.

English. In addition, for Italian/English document collections, the term-based query can be enhanced by the integration of a corpus-based query technique. The similarity thesaurus technique is being implemented to permit cross-language querying over domain-specific collections in Italian, German, French and English; it is hoped that functions that permit the cross-language querying of collections of news documents in different languages including Spanish will also be included before the end of the project lifetime (see the dashed line in the figure). Additional "translation boxes" will be installed when new partners join the federation, and configured accordingly.

The search strategy that is adopted for users at different sites will depend on the collection they wish to query. Users will thus enter their query in the local language and indicate which collection they wish to query. For example, the lexicon-based approach must be adopted when queries are addressed to the Ole database given that the application of a corpus-based methodology is not possible as no full text index is currently provided, whereas the similarity thesaurus technique will be adopted for querying over domain-specific collections.

In the next two sections we will describe the lexicon-based approach and its enhancement using a corpus-based technique.

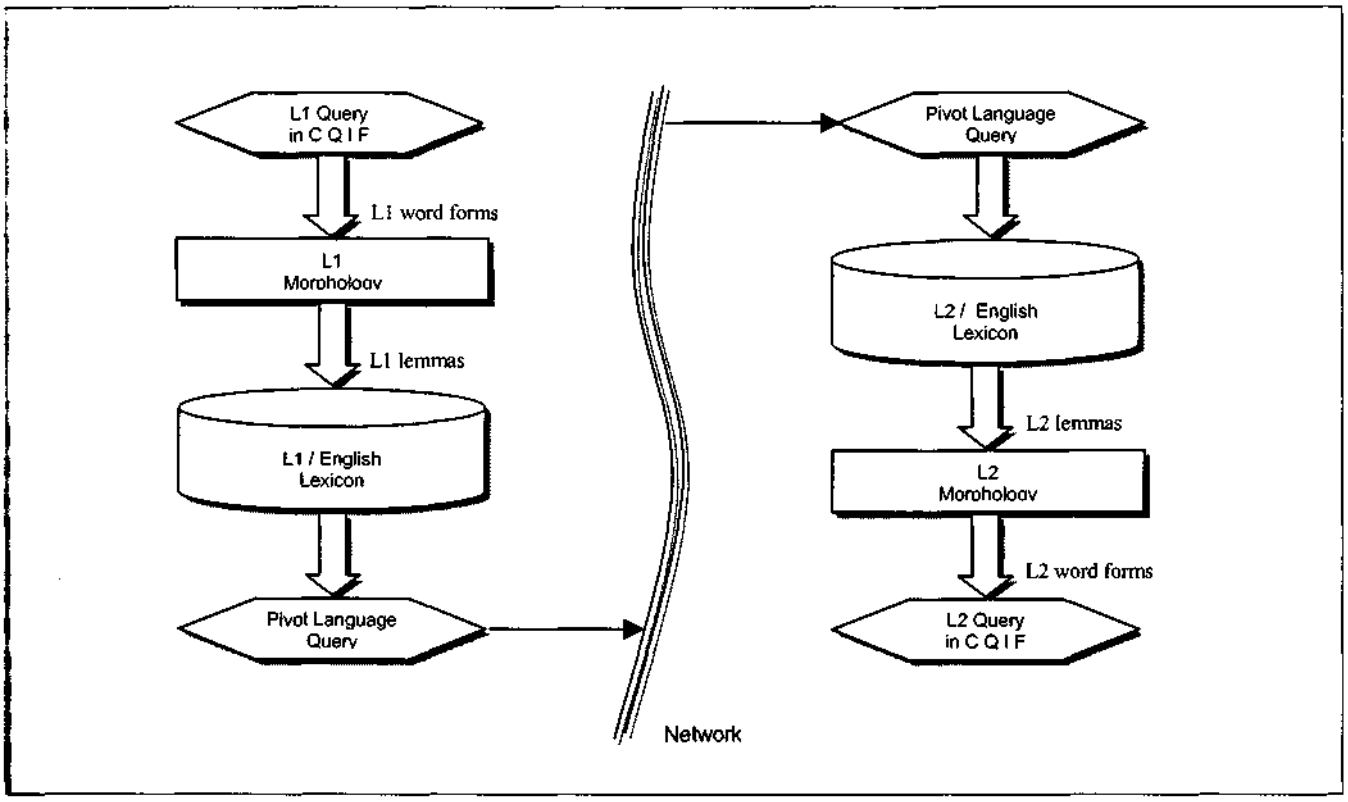

Figure 2: Lexicon-based Query Translation

4. Lexicon-based Technology

4.1 Approach

The EuroSearch Multilingual Lexicon will be constructed on the basis of a core general language vocabulary with the addition of the most frequent and significant terms used in Web queries and in Web documents. In its initial configuration the multilingual lexicon will only cover Italian, Spanish and English. However, the design of the lexicon will be open in order to permit the inclusion of additional languages in the future. For this reason, it has keen decided to adopt English as a pivot language.

Figure 2 shows how the user's query (in the Common Query Input Format (CQIF)) will be translated from his/her language (L1) into the English pivot and then into the language (L2) of the site specified in the query. This decision has been necessary in order to ensure that other languages can be included in the lexicon in the future with a minimum of effort. It has implied a trade-off between the higher level of precision provided by using separate bilingual dictionaries for each pair of languages included in the multilingual lexicon against the costs of constructing and maintaining such a (potentially) high number of dictionaries. The multilingual lexicon thus consists of a set of bilingual Local Language/English dictionaries, with procedures that map between the English datasets in the different dictionaries; the "pivot" consists of the linked set of English datasets. In order to guarantee cross-language transfer, each entry and word sense included in the pivot should have correspondences in all the other languages included in the lexicon. In each language, entries with more than one sense division have semantic indicators in the source language in order to facilitate the identification of cross-language equivalences.

As shown in the figure, to permit lookup in any source language in the multilingual lexicon and the generation of all possible forms for each suggested translation equivalent so that they can be searched in the relevant target language

documents, morphological procedures are also needed for each language.

4.2 Morphological Procedures

The morphological system that we use is PiMorfo, developed at ILC-CNR, Pisa (Picchi, 1996). The system consists of a set of language independent procedures which operate on a suitably encoded description of a language in order to recognise and produce word-forms in that language. The language description is formulated in two files: a lexicon file containing a list of base lemmas with associated morphological information and an inflection code - this file is directly linked to the data files of the multilingual lexicon; a rule file containing the rules which specify the correspondences between underlying lexical items and surface forms. The program is reversible; the same lexicon and set of rules is used for recognition and generation. Each new term added to the lexicon, is analysed by the system and an inflection code is then assigned. The system includes an on-line display and editor which can be used to view the generation of the word forms for any lemma in the lexicon and to add to or to correct either the inflectional or morphosyntactic codes if necessary. In the context of the project, the system will run for Italian, Spanish and English. It has also been developed for French and Latin.

4.3 Multilingual Lexicon

4.3.1 Vocabulary

The language used in Web searching and retrieval can be considered as a special language with particular characteristics, e.g. a restricted vocabulary employing only the major grammatical categories, a frequent use of domain-specific terms, a high occurrence of multiwords. Our multilingual lexicon will thus consist of an already existing core general language vocabulary - we have acquired two general purpose Italian/English, Spanish/English bilingual machine readable dictionaries. These dictionaries are now being implemented for the purposes of the project, e.g. unnecessary information is eliminated; archaic, infrequent and regional terms and word-senses are excluded. A special-purpose vocabulary is also being built up on the basis of a study of Web-specific terms and added to the dictionaries. Statistics on vocabulary usage in Web queries and documents from the Italian and Spanish Web services (including place names, proper nouns, acronyms, and the catalogue keywords used by the Spanish service) are now being acquired and studied with respect to the acquisition of appropriate terms for the lexicon.

4.3.2 Mapping through the Pivot Language

The problem with using a pivot language is that it introduces an extra level of possible ambiguity when passing from the source to the target language. Procedures will be written to facilitate the cross-language mapping through the pivot and reduce as far as possible the ambiguity. Thus in order to translate a term from Italian to Spanish, for example, the following path is traced:

Bilingual Dict: It/Eng **Bilingual Dict: Eng/Span**
It. entry → Eng. trans → Eng. entry → Span, trans

This is very straight forward with monosemous terms:
It **architettura** *sf* → Eng architecture →
Eng **architecture** *n* → Span arquitectura
Mapping is between equivalent entries with equivalent parts-of-speech.

More complex when we have polysemous entries, as below:

	Bilingual Dict: It/Eng	**Bilingual Dict: Eng/Span**
cancro *sm*		
1.	(Med) cancer →	**cancer** n (Med) cancer
2.	(Bot) canker →	**canker** *n* (Bot) cancro
3.	(Astron.) Cancer →	**cancer** *n* (Astron.) Cáncer

Even more complex when the polysemy is multiple, as in the following example:

	Bilingual Dict: It/Eng	**Bilingual Dict: Eng/Span**
calcio *sm*		
1.	(pedata) kick →	**kick** *n* **1.** (gen) patada, puntapié **2**. (of firearm) culatazo
2.	(Sport) football →	**football** *n* **1**. (game) fútbol, balonpié **2**. (ball) balon
3.	(di fucile) butt →	**butt** *n* **1**. (end) extremo **2.** (of gun) culata **3.** (of cigarette) colilla
4.	(Chim) calcium →	**calcium** *n* (Chim) calcio

In such cases, where possible, the mapping procedure will use the information provided by the Semantic Indicators to trace an L1 - pivot - L2 path. For each English translation of an L1 term, the equivalent entry on the English/L1 side of the dictionary will be read and the information provided by the Semantic Indicators will be used to identify the most appropriate L2 translations. The procedure will use a robust string matching technique and map only through entries with equivalent grammatical categories (identified over languages through a mapping table). It can be seen that for sense 2 of *calcio* above there is no direct string matching between the sense indicators for the Sport/game meaning of football sense. In such cases - when there is no clear indication of sense equivalence - the procedure accepts all possible target language senses. However, for common equivalent semantic indicators in the English data sets (such as the "Sport/game" case), we intend to implement mapping table, as for the grammatical categories.

4.4 Query Analysis and Translation

In Section 4.3.3 we have seen how our bilingual lexical data sets will be mapped through the pivot language. In this section we describe how the query terms are analysed by the lexicon-based translation server and translated from the query language to the target language(s).

As shown in Figure 2, the query is received in the Common Query Input Format. The first step is to eliminate stop words. These include words on the stop word list but also the minor grammatical categories, e.g. prepositions, functional words, etc.; this implies a morphological analysis. Only lexically significant words will be processed as query terms. For each term, the base lemma is identified. If the query term is not in the base form, the morphology will identify its source lemma (i.e. equivalent to a dictionary

headword). The output of this phase is a set of lemmas/dictionary entries. These will be read into the local "translation box" which contains the local language (L1)/English dictionary; English translations will be produced as intermediate output, accompanied by any Semantic Indicators, and passed to the English/target language (L2) dictionary. Target language translation candidates will be produced as output. All possible forms **will** be generated by the morphological procedures for each target language term so that they can be searched in the target language database.

The source language query term will also be passed, not translated, as target language output. This is in order to deal with untranslatable queries (see below). In the testing stage, the advantages/disadvantages of also searching for the original query terms in the target language documents will be evaluated. It is likely that this option will only be activated for terms not found in the lexicon, or queries identified by the user as proper nouns.

From the description of the L1 - pivot - L2 mapping procedure above, it can be seen that considerable "noise" i.e. spurious translations) can be produced in this phase. This must be reduced as far as possible. Here below we give examples of queries and describe how they are handled by the lexicon-based translation

4.4.1 Unambiguous Queries

The most simple case is when we are handling monosemous terms as in the "architecture" example above.

Spanish Query on Italian Document Collection
Query term: **arquitectura**
Morph analysis → arquitectura *sf* Pivot trans: architecture
Pivot entry: architecture *n* Target trans: architettura
Morph expansion: architettura, architetture
Italian documents will be searched for all occurrences of "architettura|architetture"

The query could be more specific:
Query terms: arquitectura + italiana
Morph. analysis → arquitectura *sf* + italiano *agg*|*sm*
Pivot translations: architecture + Italian
Pivot entries: architecture *n* + Italian *adj/n*
Target translation: architettura + italiano
Morph expansion: (architettura,architetture) + (italiano, italiana, italiani, italiane).
Italian documents will be searched for all occurrences of (architettura|architetture) in close co-occurrence with (italiano|italiana|italiani|italiane). It would have been possible to search only for co-occurrences of forms in inflectional agreement, i.e. in this case (architettura +italiana) | (architetture+italiane) but this can be reductive; a relevant sequence such as "architettura da maestri italiani" would be excluded from the search.

4.4.2 Ambiguous Queries

Sense disambiguation is one of the main problems in processing free-text query terms. This is already true in monolingual querying; clearly the problem is intensified when an extra passage - the cross-language step - is added. In multilingual querying, if we have an ambiguous query term, two or more distinct sets of candidate translations can be selected; the passage through the pivot language can also increase the potential "noise", as shown above. We have to study methods to constrain the risks of an explosion of multiple senses as far as possible. To a large extent, if several query terms are processed simultaneously (i.e. they form part of a single query) much ambiguity and noise will be eliminated automatically. For example, an Italian user could formulate a query using the terms: "calcio" and "rigore"; while both words are ambiguous in Italian, it is most likely that the cooccurrence in documents of the possible translation candidates will occur only for the "football" senses. However, unfortunately, experience shows that a large percentage of queries on the Web actually consist of single query terms.

4.4.3 Query Refinement

When the user enters an ambiguous query term or set of terms, the query will be processed and the first set of results will be presented; at the same time the user can be asked if he wants to perform a query refinement. For example, by adding another term, as shown here:

Italian Query on Spanish Document Collection
Query term: **cancro**
Morph analysis → cancro *sm*
Pivot trans.: cancer (Med), canker(Bot),
Cancer (Astron.,Astrologia)
Pivot entry: cancer *n* (Med) Target trans: cáncer
and Pivot entry: canker *n* (Bot) Target trans:cancro
and Pivot entry: Cancer *n* (Astron., Astrologia)
Target trans: Cáncer

The Spanish document collection will be searched for documents which can concern Medical, Botanical and Astrological topics. However, in this case, the ambiguity is precisely the same as that which would have been obtained in an Italian monolingual search (thus the pivot adds no "noise").

In a monolingual query, the user interested in the medical sense of "cancro" but retrieving many documents on fortune-telling, would probably better define his query by the addition of another relevant term, e.g.
Query terms: **cancro** + (**polmonare**| **polmone**)
Morph analysis → cancro *sm* polmonare *agg* polmone *sm*
Pivot translation: cancer (Med), canker (Bot), Cancer (Astron., Astrologia)
and pulmonary
or lung
Pivot entries: cancer, canker, Cancer *n* →
Target trans: cáncer, cancro,Cáncer (see above)
and Pivot entry: pulmonary *adj* →
Target translation: pulmonar
or Pivot entry: lung *n* →
Target translation: pulmón

and the Spanish document collection will be searched for documents containing the keywords (cáncer|cancro|Cáncer) in close occurrence with (pulmonar|pulmón). It is thus far more likely that documents referring to lung cancer rather than other topics will be retrieved.

Another way of eliminating ambiguity in a query term, is for the user interface to request the user to perform an interactive sense disambiguation. When a highly ambiguous query term, such as Italian *calcio* is entered, the query can be processed but at the same time the user can be sent a message that asks if he/she wants to perform a query refinement - in this case a disambiguation of the query term.

A menu can be presented to the user which displays the different senses of the query term with associated Semantic Indicators in the Query language and he/she can click the intended sense e.g. for an Italian user querying "calcio" the display would be:

calcio 1. pedata
2. sport
3. di fucile
4. chimica

In 99% of cases, sense no. 2 would be selected!

4.4.4 Untranslatable Queries

Not all query terms are translatable. The term may not be included in the lexicons and/or it may be a proper noun (person or place name). As stated above, all query terms - whether contained in the lexicons or not - are also output by the translation processor in their source format and can thus be searched through the target language database.

4.4.5 Treatment of Multiwords in the Queries

Multiwords will be recognized in queries if they are identified as such by the user, according to the indications provided by the user interface. In this case, they will be looked up in the lexicon as a single lexical item; however, if there is no entry for the multiword then the separate elements will be translated and searched in co-occurrence in the target language. One of the tasks of the project will be the implementation of procedures for the recognition of multiwords in the Web documents; the most frequent will be included in the multilingual lexicon.

5. Enhancement using Comparable Corpus Methodology

The simple lexicon-based query translation described above has clear limits. When a term in the query is not included in the lexicon, no translation can be found. In Section 4.4.4 above we have indicated how the lexicon-based translation server will handle these cases; however, we will also be implementing an experimental methodology which uses data extracted from document archives to expand the terms in the query by associating a vocabulary of related terms. In this way, we can provide a relevance ranking of our results (documents containing a higher proportion of the term correlated vocabulary are considered more relevant) and query terms which are not included in the multilingual lexicon can also be searched.

This strategy is based on the assumption that (i) words acquire sense from their context, (ii) words used in a similar way will be semantically similar, (iii) and that this is also true in a cross-language setting. It follows that, if it is possible to establish equivalences between several items contained in two different contexts (i.e. documents), even in different languages, there is a high probability that the contexts themselves are somewhat similar. Thus, given a particular term or set of terms in the documents in one language, the aim is to be able to identify contexts which contain equivalent or related expressions in the collections in other languages. To do this, we isolate the vocabulary related to that term in the documents in the first language (which we call the source language) - hypothesising that lexically equivalent terms will be associated with a similar vocabulary in the target languages. The application of this method to an Italian/English comparable corpus of parliamentary texts is documented in (Peters and Picchi, 1997).

In EuroSearch, the challenge is thus to extend and test a methodology, originally developed to run on sets of domain-specific texts in different languages, for cross-language searching over a much wider vocabulary, as represented by the set of documents accessible on national Web sites. We will experiment it for Italian/English cross-language Web querying. Unfortunately, we are unable to test it for Italian/Spanish searching as the Ole provides a catalog-based search through keyword, without a full-text indexing of documents.

5.1 Methodology

For any term of interest, T, searched in one language, our objective is to be able to retrieve a ranked list of documents containing equivalent terms in the other languages. Thus when T is entered, the system will automatically construct a context window containing T and up to "n" lexically significant words (nouns and verbs only) to the right and left of T for the set of documents in our collection. The value for "n" can be varied. For each of these co-occurrences of T morphological procedures identify the base lemma(s), i.e. each word-form is analysed in order to match it against equivalent forms and to identify the relevant entry that will be looked up in the multilingual lexicon. The significance of the correlation between its collocates (i.e. significantly co-occurring terms) and T is then calculate using a statistical procedure. We are currently using Church and Hanks' Mutual Information Index (Church & Hanks, 1990).

The set of most significant collocates derived makes up the vocabulary, V, that is considered to characterize our term, T, in the document collection. For example, the twelve most significant collocates obtained using this method for *cancro* in an Italian corpus of news documents were *tropico, polmone, retto, tumore, disturbo, seno, malato, aids, tipo, anima, cura, ricerca* (tropic, lung, rectum, disorder, breast, patient, aids, type, soul, treatment, research) and for *calcio* were *sale, campionato, giornaliero, tifo, videogioco, mondiale, tribu, partita, gioco, minuto, rigore, squadra* (salt, championship, daily, fanaticism, videogame, world, tribe, match, game, minute, penalty, team).

The next step is to establish an equivalent target-language vocabulary for T. This vocabulary represents the set of potential significant collocates for T in the target language. This is done by looking up each item of vocabulary V in the Italian/English database and extracting the entire set of possible translation equivalents. The target language vocabulary for T is thus significantly larger than the source language vocabulary. Words or expressions that can be considered as lexically equivalent to our selected term in the source language texts will then be searched in the document sets in other languages, i.e. we do this by searching for those contexts in the target language collections in which there is a significant presence of the target language vocabulary for T. The significance is determined on the basis of a statistical procedure that assesses the probability for different sets of target language cooccurrences to represent lexically equivalent contexts for T. The target language documents retrieved are listed in descending order of relevance to our original query term. In

the experiments we have performed so far, the creation of target language vocabularies for any term of interest, T, has been performed on-line. In a real-world retrieval context, such as EuroSearch, the creation of these vocabularies will be done off-line, periodically, in order to optimize the on-line search times.

5.2 Search Term Disambiguation

The construction of the source language vocabulary which characterizes our term T permits us to obtain a clustering of the most relevant terms connected to T. If the document collection contains a predominant sense for the term then the vocabulary should represent this sense - secondary senses that appear rarely will not interfere with this. This is the case of the example of *cancro* above; apart from the first term (Tropic), and possibly the tenth term *anima*, all the others, in some way, refer to the medical sense. If, in the collection, there is more than one relevant sense for T then we would expect two or more distinct clusterings of significant collocates. We are currently working on the definition of a technique that should make it possible to separate distinct senses of the same word in a document collection on the basis of their collocates; for each collocate we will build the set of most strongly related words and compare these to identify overlapping. In this way, we hope to distinguish between the sense of *calcio* characterized by words such as salt and daily (plus diet and milk which appeared further down in our list of significant collocates) and the very different sense identified by collocates such as championship, match, team, penalty, etc.

5.3 Target Term Disambiguation

When constructing the target language vocabularies of significant collocates for the source language term being searched, our procedure will take as input all the translation candidates provided by the multilingual lexicon, regardless of sense distinctions. We denote these as "translations blocks". It must be remembered that we are searching in the target language collection for documents with a significant co-occurrence of items from this vocabulary. Spurious or inappropriate translations are eliminated by the fact that we normally do not find them together with a significant number of other items from the target language terms proposed. For example, let us look again at the case of *cancro* above. Our target language vocabulary, for the twelve most significant collocates, will consist of the translations listed above, which include tropic, clearly irrelevant for the dominant medical sense, soul and also words such as bother, inconvenience, noise, interference (translations of non-medical senses of disturbo). However, only if such terms are found together with other terms from the target language vocabulary - particularly unlikely for tropic, soul, or noise, will they become significant for the search.

This makes it possible for us to perform a sense disambiguation on the target terms proposed. Thus, our approach helps us to identify the correct sense of the target language translation candidates and to provide a ranking of th**e** best target language matches for the query term searched (For more details, see Peters and Picchi, op cit.).

Using this methodology, we can enhance the results of cross-language lexicon-based querying as we can also search for terms for which cross-language equivalences are not included in our multilingual lexicon.

6. Similarity Thesaurus Technology

As has been stated, the objective of this paper has been to describe a method for cross-language information retrieval that uses lexical and linguistic resources and methodologies in a real-world application. Before concluding, however, we briefly present the other multilingual search component that will be employed in the EuroSearch project.

The similarity thesaurus technology is based on ideas originally developed for monolingual query expansion (see Qui, 1995). A multilingual similarity thesaurus contains entries that link terms in one language (L1) to a list of "similar" terms in another language (L2), each assigned with a value giving an estimate of similarity. This estimate is based on statistical occurrence, i.e. basically on how often the terms co-occur in similar texts taken from training data. The process for calculating the thesaurus is fully automatic. A similarity thesaurus that provides a mapping from terms in language L1 to similar terms in language L2 allows a query formulated by the user to be transferred into the target language by substituting the query terms with some of their most similar counterparts. A distinctive property of this approach is that the target language query produced is not really a translation of the user's search request, but a reformulation using terms that are likely to retrieve documents relevant to the user's information needs (for a full description of this technology, see Sheridan & Ballerini, 1996; Sheridan et al, 1997).

The similarity thesaurus technique does not use lexical resources nor does it require the morphological tools needed by the lexicon-based approach to match the word-forms found in queries and documents to the lemmas given as dictionary headwords. As long as suitable training data is available, the thesauri can be built automatically and a relatively simple stemming algorithm is sufficient to match terms in texts. As can be seen in Figure 1, the similarity thesuarus technology will be employed in EuroSearch for the querying of domain-specific collections, initially in Italian, French, German and English - possibly with the inclusion of Spanish before the end of the project. For details on how this technology will be implemented in the Eurosearch project, see Peters et al. (1998).

Of course, any method will have advantages and disadvantages. The main problem in the lexicon-based approach is probably represented by the adequacy of the lexical resources to provide sufficient coverage of the vocabulary used in both Web queries and documents, and the consequent need to update this vocabulary. The need to implement a pivot language in order to provide a truly multilingual base for our lexicon also introduces additional difficulties. To a large extent, we attempt to overcome these by activating a cross-language query expansion with the addition of corpus-extracted data on significant collocates and the translation of the data into the target language(s). The major drawback of the similarity thesaurus technique is that it is only applicable to particular domains and it is difficult to envisage its extension to general purpose querying.

7. Final remarks

We have described the strategies being adopted by a European federation of Web search services in the implementation of a multilingual interface. A particular feature of this federation is that each member is already providing a monolingual retrieval service and that the particular characteristics of the existing search engines and document collections had to be respected. For this reason, we could not implement a single cross-language strategy but had to adapt our proposal to meet the requirements of the various local situations. We had to be able to query collections covering the entire national Web space or specific domains; to query using full-text query mechanisms and keyword-type searches. At the same time, we had to design a system that could incorporate new partners and additional languages into the federation. For these reasons, it has been necessary to implement both lexicon and corpus-based techniques; the strategy activated at query time will depend on the collection that the user intends to search.

8. References

Church, K. & Hanks, P. (1990). Word association norms, mutual information and lexicography. Computational Linguistics, 16(1), *22-29*.

Oard, D.W. & Dorr, B.J. (1996). A Survey of Multilingual Text Retrieval, University of Maryland Technical Report. UMIACS-TR-96-19. {http://www.glue.umd.edu/oard/research.html}

Oard, D.W. (1997). Alternative Approaches for Cross-Language Text Retrieval. In Proceedings of AAAI Symposium on Cross-Language Text and Speech Retrieval. {http://www.glue.umd.edu/oard/research.html}

Peters, C. Picchi, E. (1997). Using Linguistic Tools and Resources in Cross-Language Retrieval. In Proceedings of AAAI Symposium on Cross-Language Text and Speech Retrieval. {http://www.clis.umd.edu/filter/sss/papers/}

Peters, C., Picchi, E., Braschler, M. & Schäuble, P. (1998). EuroSearch: Multilingual Components Specification. LE4-8303 EuroSearch Deliverable 3.2.

Picchi, E. (1996) PiMorfo: sistema di analisi morfologica. Technical Report, ILC-CNR, Pisa, {http://www.ilc.pi.cnr.it/dbt/pimorfo.htm}

Qiu, Yonggang (1995). Automatic Query Expansion based on a Similarity Thesaurus. PhD. Thesis, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.

Sheridan, P. & Ballerini, J-P (1996). Experiments in Multilingual Information Retrieval Using the SPIDER System. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 58-65.

Sheridan, P., Braschler, M. & Schäuble, P (1997). Cross-Language Information Retrieval in a Multilingual Legal Domain. In C.Peters & C.Thanos (Eds), Proceedings of First European Conference on Research and Advanced Technology for Digital Libraries, LNCS 1324, Springer, pp 253-268.