Opening the legal literature Portal to multilingual access

E. Francesconi, G. Peruginelli
ITTIG – Institute of Legal Information Theory and Technologies
Italian National Research Council,
Florence, Italy
Tel: +39 055 43999
Fax: +39 055 42272
e-mail: {francesconi, peruginelli}@ittig.cnr.it

Abstract:
The issues of multilingual access to legal information are examined, and strategies of cross-language retrieval to legal information resources are illustrated as additional services of the Portal to legal literature created by ITTIG. Consideration is given to the peculiarities of legal language as a technical language closely related to the diverse legal systems.
The paper describes a methodological approach planned for the Portal to provide a single point of access into disparate repositories where categories of law (i.e. trade law, constitutional law, criminal law) are the essential metadata to point to relevant material irrespective of the language used in a query. Categories of law of a specific legal system represent the way how retrieval can be satisfactorily achieved.
Strategies and techniques for translating legal queries to different target languages, eventually disambiguating ambiguous words by a machine learning approach are illustrated.
Keywords:
legal literature, cross-language retrieval, multilingual metadata, word sense disambiguation.

1 Introduction

Legal information has peculiarities due to its multifaceted nature, methods of use and the integration requirements of different legal information sources such as statutes, case-law and legal doctrine. In particular, the information retrieval of legal literature requires high quality indexing as well as appropriate searching methods and tools in order to be effectively accessed by diverse legal user communities.
The Portal of Italian legal literature developed by the Institute of Legal Information Theory and Technologies in 2002 and presented at the DC2003 Conference in Seattle has taken into account such requirements (1). The provision of a unified point of access to multiple legal doctrine resources through the exploitation of rich metadata and the development of tools for the discovery, selection and use of relevant legal material are the core functions of the Portal. At the same time, methodologies and tools have been developed to allow the communication of legal knowledge among general users, in an attempt to solve some of the difficulties caused by the complexity of legal language.
Efforts so far have been made to design the Portal's system, conduct a survey of Italian legal users and develop integrated tools for generating and capturing metadata for structured and semi-structured web documents (1).
A second phase has recently been launched, where emphasis is put on two distinct requirements, both addressing the need for international access through the legal Portal. These consist in: a) opening up the system to a wider user community, including foreign patrons, who must be given the possibility to access Italian legal material in their native language; b) providing multilingual access to foreign legal resources.
These objectives are based on the belief that the development of strategies and tools that enable access to information regardless of geographic or language barriers is a key factor for the truly global sharing of legal knowledge, thus making it possible for legal research and the legal profession to progress according to the requirements of modern society. With the rapid increase in globalization, transnational issues can be expected to arise today in virtually any legal context, therefore the retrieval of foreign law material with adequate multilingual tools becomes an essential requirement for success in modern law practice and research. Furthermore, the principle of multilingualism in the domain of law not only ensures democratic transparency and the equality of citizens' rights, but also guarantees legal certainty.

A two-phase approach has been developed for the implementation of the Portal's multilingual access function. Firstly, analysis has been carried out of cross-language retrieval approaches, their application to law and the multilinguality of metadata. Secondly, a practical approach to the retrieval of multilingual legal resources has been defined. Based on the features of the Portal's federation system, where documents coming from structured repositories and from the web are qualified using Dublin Core metadata set in its XML version, query translation and word disambiguation techniques have been implemented.

2 Multilingual information access approaches

In order that multilingual access be guaranteed, that is so that information can be searched, retrieved and presented effectively, without constraints due to the different languages and scripts used in documents and in metadata, both the users' native language and the multiplicity and wealth of world-wide languages have to be accommodated. What is needed is functionality like the thorough and proper handling of characters (their presentation, arrangement, transfer), putting queries into a preferred language and script, retrieving resources irrespective of the language used in searching and indexing, enabling world-wide communication no matter what the language. As a consequence multilingual facilities must include both multiple-language recognition and cross-language information retrieval (CLIR) (2)

Among the basic approaches to cross-language retrieval, based on controlled vocabulary and on free (or full-text) retrieval, no optimal solution exists: this is widely demonstrated by the wealth of applications, research and studies under way in the field of CLIR (3)

The use of controlled vocabulary and a multilingual thesaurus in a cross-language retrieval environment, where selected terms from each language are related to a common set of concept identifiers (4) involves labour-intensive work developing and managing such tools, especially in applications where diverse domains and material are to be managed.

Free-text searching and, connected with it, methods based on dictionaries or corpora (that is analysing existing collections of texts from which to extract information for the application specific-translation techniques) are an alternative approach (5). Here either the query is translated or the document when it is indexed, but limitations exist and vary according to the techniques used. Difficulties are the lack of equivalence in translation, the ambiguity which can arise from translating terms between languages when no context is provided and the availability of parallel and comparable corpora. This is particularly true in the domain of law.

3 Legal language peculiarities

In implementing the legal Portal's services the approaches to multilingual access are considered with reference to the peculiarities of legal language, which is a strictly technical language. Like other disciplines, law has its own lexicon, it uses ad-hoc terms and attributes specific meanings to terms taken from ordinary language. It is a sort of internal code allowing communication between legal experts, where terms and technical expressions respond to economic criteria, making concepts understandable by using a restricted vocabulary (6).

All this applies at a national level. At an international level the complexity and richness of diverse legal languages make understanding and exchange of concepts expressed in such languages a very difficult task. However, the main problem of legal terminology at an international level mainly regards the difference of legal concepts inherent to the diverse national legal systems. In fact merely translating words is likely to confuse users when a concept does not exist in the target legal system (7).

A basic difference between legal systems lies in the fact that, speaking in very general terms and without taking into account mixed legal systems, the Western world has long had two dominant legal traditions: Common law, with its beginnings in England, and Civil law, rooted in continental Europe[1]. At this stage of the project the Portal mainly deals with material drawn from these two basic systems.

Both systems have taken on a variety of cultural forms. Civil law systems have drawn their inspiration largely from their Roman law heritage and, by giving precedence to written law, have resolutely opted for a systematic codification of their general law.

Common law systems are instead based on English common law concepts and legal

[1] However no modern legal system is a "pure" representation of any type of system; all jurisdictions today represent mixed systems, to some extent.

organizational methods which assign a pre-eminent position to case-law, as opposed to legislation, as the ordinary means of expression of general law.
Quite a number of concepts belonging to different legal systems can hardly be transposed or even compared with one another.
Unlike technical and scientific disciplines, serious difficulties arise in translating law material due to the system-bound nature of legal terminology.
Consequent to this, categories of law (i.e. trade law, constitutional law, criminal law) are specific within a particular legal system. Therefore, categories of law of a specific legal system represent the way how retrieval can be satisfactorily achieved. Moreover, a legal system identifies one or more languages of a country where it is operative. At this stage of the study we consider a one to one mapping between a specific legal system and a given language.
As often there is no conceptual nor content similarity between the categories of law of the different legal systems, mapping between such law categories is necessary to reach proper contextualisation of the query in the diverse legal systems. An example illustrates the need for such mapping. The concepts related to property rights, such as the development of property law, land law, property questions on insolvency, intellectual property, etc. according to UK law belong to the field of property law, whereas in the Italian legal system these legal facts are regulated by private law, agricultural law and industrial law.

3.1 Equivalence between legal languages

In order that cross-language retrieval of legal material be effective, priority must be given to translating one legal language to another legal language and not to the ordinary words of the target language, as this can cause ambiguity and misunderstanding.
Therefore the whole process of interaction between legal languages can be identified as finding equivalents across legal systems (8).
If no acceptable equivalents can be found in the target-language, subsidiary solutions must be sought, such as no translation and use of source terms, paraphrasing, creating a neologism with explanatory notes.
The research for equivalence implies both a comparative study of the different legal systems and adequate knowledge of technical legal terminology. Some examples are provided, highlighting the complexity of comparing different legal languages.
Pure linguistic problems are likely to be encountered due to legal false friends. Some examples are given below. The terms «administrative tribunals» cannot be translated in French as «tribunaux administratifs». The English word for the French tribunal is Court and the administrative tribunals are administrative commissions which are comparable, *mutatis mutandis,* to the French «autorités administratives indépendantes».
The so-called «procacciatore d'affari», typical of the Italian legal system, has no equivalence in other systems. Translating it as «broker» (in English) or «pourvoyeur d'affaires» (in French) is incorrect. Unlike the Italian, these latter terms refer to an employee of a company for which he works and from which he gets a salary and not a commission.
Despite the difficulties in establishing the equivalence of legal concepts belonging to different legal systems, a compromise has been adopted in trying to favour the integration of diverse legal cultures, while respecting each national legal system.
What is needed is the identification of a common ground, namely common legal concepts and facts which, although not perfectly coinciding with those belonging to other systems, are conceptually close. It is up to legal users, once the material has been examined, to perceive the differences and peculiarities which make these resources unique. It is to be underlined that this does not necessarily lead to noise or unsuccessful searches, but allows for a first-phase search in context, useful to give evidence of the existence or non-existence of a specific concept in other legal systems.

4 Multi-language metadata approach

As discussed in (1) our Portal is aimed at integrating data coming from structured repositories and from Web documents in a unique point of access. In this paper, the architecture has been extended to deal with cross-language facilities.

4.1 Structured data and Web documents in a multi-language environment

Structured data coming from different repositories are usually provided with a specific metadata scheme, as well as a

specific language-dependent classification scheme. While metadata sets can be harmonized in DC, contents of metadata, especially the content of *dc:subject* can not (Section 3), except within a legal system, that in this study we consider correspondent to a language (Section 3).

Data providers are required to expose metadata, making repositories compliant to the DC metadata and ready to be harvested using the OAI-PMH protocol (1). At the end of this process the service providers are able to set up many XML metadata repositories as original languages considered (Fig. 1).

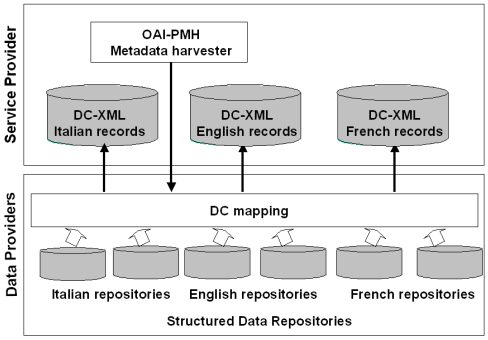

Fig. 1 The OAI harvesting of structured metadata in different languages.

On the other hand, Web documents do not usually contain any particular metadata scheme, nor any reliable or uniform HTML meta-tags, which can help the qualification of material of interest. As discussed in (1) in our architecture Web documents are selected and harvested with no prior agreement between service providers and data providers, using a focused crawler that selects documents within each legal system, corresponding to a language (Section 3). For this kind of documents, an automatic metadata generator capable of applying DC metadata has been developed and tested (1). Particularly a machine learning approach based on a naïve Bayes classifiers, has been used for *dc:subject.*

In our architecture the problem of extending the automatic metadata generator to deal with multi-language documents is the problem of extending the automatic classifier in a multi-language environment.

Multi-language automatic document classification has not widely studied on its own. In literature two main kinds of classifiers for multi-language document categorization can be distinguished (9):

1) A poly-language trained classifier: one classifier trained on documents written in different languages;
2) A single-language trained classifier: one classifier trained on documents written in language A and a translation of the most important terms of language B to A, in order to classify documents of language B.

Both these approaches assume that documents of different languages share the same classification scheme. However, as discussed in Section 3 this is not the case with us: we deal with documents of different legal systems, corresponding to different languages, each having a particular classification scheme, that usually cannot be harmonized one another.

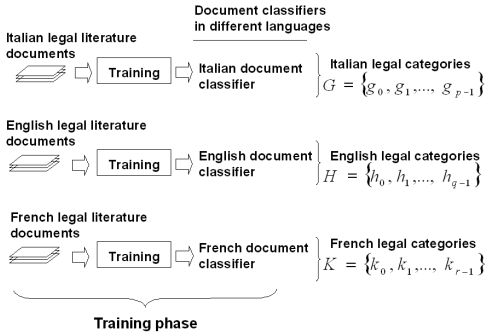

Fig. 2 Classifiers trained to apply *dc:subject* to Web documents within different language domains.

Therefore, in our architecture a naïve Bayes classifier is trained (Fig. 2) and used to apply *dc:subject* to Web documents within each language and classification scheme.

At the end of this process the service provider collects as many DC-qualified HTML metadata repositories as original languages (Fig. 3).

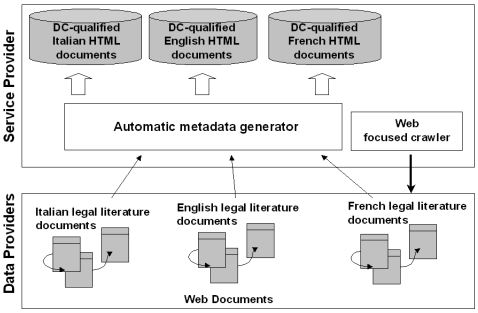

Fig. 3 The harvesting and DC qualification of Web documents in different languages.

4.2 Indexing legal documents

After having collected data from structured repositories and from Web documents the service provider obtains two types of archives containing data in different formats (XML records and HTML documents), sharing the

same DC metadata description scheme. Each type of archive is composed by sub-archives containing data in one language.
At this stage an indexing procedure, able to handle XML records and HTML documents (10), can be implemented with the aim of providing a uniform view and integrated access to the data of our Portal. The number of indexes obtained with the same DC metadata scheme is the same as the number of languages considered (Fig. 4).

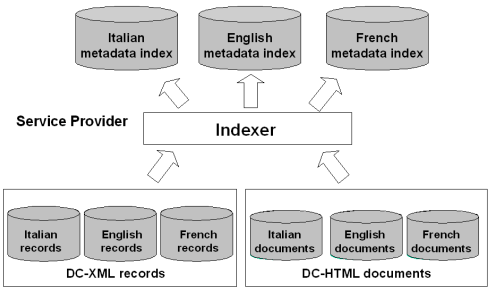

Fig. 4 Indexing of DC metadata from repositories using different formats and languages.

5 Access modalities

Having built the portal indexes in a multi-language environment, users may access data by these two modalities of query:

1) *metadata-based document querying* (MBDQ);
2) *keyword-based document querying* (KBDQ), combined with category (*category-based document querying* (CBDQ)).

Case 1) Advanced search: the user submits a query filling in the fields related to DC metadata.
Case 2) Simple search: the user submits a query, filling an unqualified text box with keywords. Moreover, in order to make the query more focused, the user may choose a legal category of the legal system associated with a language domain.
Dealing with querying and retrieval of multi-language documents, essentially involves the problem of query translation.
As discussed in Section 3.1, especially in legal domain, a word in query language can be ambiguous, having therefore different translations in a target language, each corresponding to a legal category in the target legal system (i.e. the Italian word "dolo" has two different translation into English: "fraud" and "malice", respectively belonging to private law and criminal law). The right sense of an ambiguous word in query language can be obtained only by word contextualization, giving the right sense to the context in terms of legal category Then such a legal category in the query legal system, can be mapped to the correspondent legal category in the target legal system, therefore the right translation of the ambiguous word can be obtained. If more than one category in the target legal system corresponds to the legal category of the query legal system, more than one translation of the ambiguous word are selected.
Therefore, the knowledge of a legal category in both the modalities of querying (MBDQ and KBDQ+CBDQ) is essential in order to identify the right translation of an ambiguous word.
Once query is translated in target languages and contextualized, documents of different languages are given back. The procedures used to obtain these results in MBDQ and in KBDQ+CBDQ modes are described respectively in Section 5.1 and 5.2 (Fig. 5 can be used as reference).

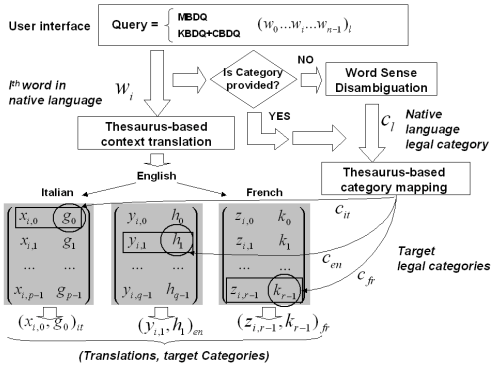

Fig. 5 Query translations of in MBDQ and KBDQ+CBDQ modalities

5.1 Query based on metadata (MBDQ)

MBDQ represents an "advanced search" standard modality of querying a qualified index. The user first of all is required to choose a legal system, thus implicitly identifying a language for queries, and a legal category, in terms of *dc:subject*, thus implicitly identifying the right translations of possible ambiguous words. Then each metadata field is filled with a set of words $(w_0, w_1, ..., w_{n-1})_l$, representing a context expressed in the query native language l, that has to be translated by a *thesaurus-based context translation* procedure.
Not every field has to be translated. In fact, DC metadata can be divided into query

language-dependent and query language-independent metadata. For example *dc:title* is query language-independent since, for example, the title of a document has to be queried in its native language, independently from the query-language. Therefore only the contents of query language-dependent metadata have to be translated. While in a multi-language environment *dc:subject* is usually query language-independent (or neutral (11)), within a multi-language legal domain this is not true (Section 3). For this reason *dc:subject* has to be translated, by a mapping of its values from a legal system to different target ones.

Also the content of *dc:description* (with its qualifiers, such as Abstract) is query language-dependent and it is a widely used access point: the information contained is often expressed using a semi-technical language; therefore *dc:description* element has been held as being as important to translate as *dc:subject* in the Portal functions.

The contents of *dc:subject* and *dc:description*, submitted in a native language are translated in a "pivot" language (English) (12). Then, from the "pivot" language, the query is translated again to the other languages used by the Portal.

The use of a "pivot" language in a N-language environment allows the reduction of the number of bilingual thesauri from a factor N^2 to a factor N, and also allows the solution of the problem of the non-availability of some bilingual thesauri.

As discussed in Section 5 the main problem with translation is that a single word (w_i) or expression in the native language can have different translations in a target language, depending on the context. For example, let us assume, without loosing generality, that w_i be an ambiguous single word of the context $(w_0, w_1, ..., w_{n-1})_l$ in *dc:description* in query native language l (Fig. 5). According to Fig. 5, different English translations $\{y_{i,0}, y_{i,1}, ..., y_{i,q-1}\}$ can be associated to w_i, each one corresponding to as many legal categories $\{h_0, h_1, ..., h_{q-1}\}$. For example, being the language l=Italian and w_i="dolo", possible translations in English are $y_{i,0}$="fraud" related to law category h_0="private law" and $y_{i,1}$="malice" related to law category h_1="criminal law". The right translation can be obtained only by knowing the sense, namely the category h, of the context in the query native language, where w_i is contained.

Such a context, or legal category, is required and is provided by the user using *dc:subject* element.

When a category c_l (Fig. 5) is selected in the *dc:subject* within a legal system, the problem arises of different classification schemes in different languages, corresponding to different legal systems (Section 3.1). The problem can be solved by using a *thesaurus-based category mapping*. In fact, when the category c_l is submitted as a query parameter, the category c_l is mapped in the corresponding, or the closest, categories in the "pivot" language, and from it to the other languages considered by the Portal ($c_l \Rightarrow c_{en} \Rightarrow \{c_{it}, c_{fr}\}$), using category thesauri. In accord with Fig. 5 and without loosing generality, let us assume that only one legal category $c_{en} = h_1$ in the English legal system corresponds to the legal category c_l ($c_l \Rightarrow c_{en} = h_1$). Consequently, the English translation $y_{i,1}$ (Fig. 5) related to the sense h_1, can be selected (in our example, the English word $y_{i,1}$="malice", related to law category h_1="criminal law" is selected as the right translation of the Italian word w_i="dolo"). If more than one category of the target legal system can be associated to c_l, all the corresponding translations of the current w_i are selected. When all the words of the current context are translated in *dc:description*, we obtain the translation of the submitted context $(w_0, w_1, ..., w_{n-1})_l$ from language l to Portal target languages. The value c_l in *dc:subject* is also mapped to the corresponding categories in the target languages. Now queries in different languages are ready to be dispatched to the related domain language indexes (Fig. 6).

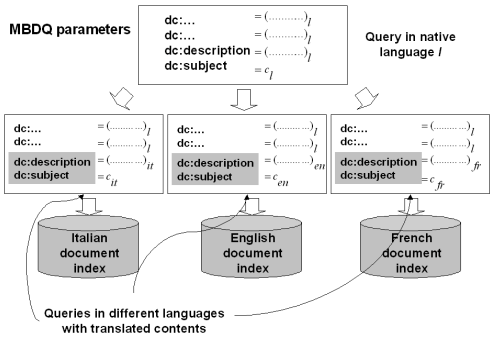

Fig. 6 MBDQ: results of query translation in different languages (in grey metadata whose content is translated).

5.2 Query based on keywords and legal categories (KBDQ+CBDQ)

A query based on keywords and legal categories represents the "simple search" modality of querying our Portal. In this mode the user is provided only with an unqualified text box to be filled with a context $(w_0, w_1, ..., w_{n-1})_l$ of words in a native language l. Words identifying the context will be translated into the target languages of the Portal (*thesaurus-based context translation*). Moreover, the user may or may not provide a legal category of the query legal system. Since category is essential for translation of ambiguous words, if a legal category is not provided, the system attempts to infer the correspondent legal category from the query context.

If the user selects a legal category c_l, among the values of *dc:subject* in the query legal system, a procedure of *thesaurus-based category mapping* is executed, as described in Section 5.1, obtaining the correspondences of c_l in Portal target legal systems (Fig. 5).

If the user fills only the unqualified text box, without choosing any value in *dc:subject*, the right sense to the query context is provided by a procedure of automatic word sense disambiguation, which assigns a legal category to a context as described in Section 5.3. The legal category thus identified in native query language, is then mapped to the related legal categories in target legal systems (*thesaurus-based category mapping*).

At the end of the process, the right translations of ambiguous words can be obtained, as discussed in Section 5.1 (Fig. 5), and as many different queries as target languages used by our Portal can be dispatched to the different language indexes (Fig. 7).

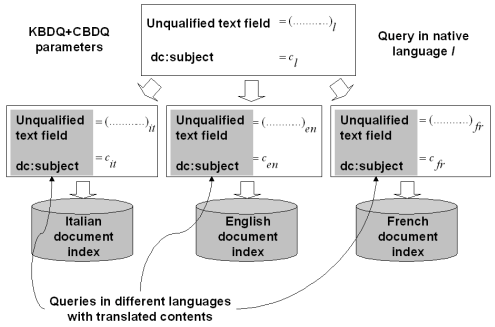

Fig. 7 KBDQ+CBDQ: results of query translation in different languages (in grey metadata whose content is translated).

5.3 Automatic word sense disambiguation

The problem of assigning the right meaning to a word in context is a problem of assigning the right sense to the context itself out of the various meanings that can be assigned to the ambiguous word.

According to literature lots of methods have been used to solve the problem of automatic disambiguation:

- Thesaurus-based disambiguation (13);
- Disambiguation based on sense definitions (14);
- Disambiguation based on translation in a second-language dictionary (15);
- Bayesian disambiguation (16).

In our Portal word disambiguation is a problem of context categorization with respect to the legal categories considered within a legal system. Moreover (16) context categorization is the same problem of document categorization, once we view contexts as documents and word sense a s categories. For these reasons in our system we use the same naïve Bayes classifiers described in Section 4.1, trained with labelled documents of different legal categories of a particular legal system and language. At the end of the training phase each category profile (in our case a vector of weighted terms relevant to it (1)) can also be considered as a context profile to be used for disambiguation function.

Given the set of categories, considered by our Portal, in a legal system, the naïve Bayes classifier used for word sense disambiguation is a ranking classifier which, for a given query context, returns the scores for the different categories. Each score represents the evidence

for placing a given context to a certain category.

It is important to execute automatic word disambiguation prior to translation, because, as discussed, correct word translation depends on contextualization activity of words in their native language.

6 Conclusions

Multilingual access to legal documents is problematic due to the legal system-bound nature of this type of information. In the legal literature Portal created by ITTIG an approach has been developed allowing cross-language retrieval of both structured and unstructured documents by exploiting content of *dc:subject* in diverse legal systems.

The designed functionality aims to provide a single point of access into disparate repositories where categories of law are the essential metadata to point to relevant material irrespective of the language used in a query. This is done through techniques able to translate legal queries to different target languages, eventually disambiguating ambiguous words by a machine learning approach. Basically, the approach gives the advantage of accessing multi-language legal documents respecting the identity and the peculiarities of different legal systems.

References

1. E. Francesconi, G. Peruginelli, "Integration between structured repositories and web documents", in *Proc. of the DC Conference 2003*, pp.99-107
2. C. Peters, E. Picchi, "Across languages, across cultures: issues in multilinguality and digital libraries", in *D-Lib Magazine,* May 1997. Retrieved July 7, 2004, from http://www.dlib.org/dlib/may97/peters/05peters.html
3. J. Mayfield, P. McNamee, "Three principles to guide CLIR research, 2002", Retrieved July 7, 2004, from http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR-18-mayfield.pdf
4. C. Fluhr, "Multilingual Information Retrieval", in *Survey of the State of the Art in Human Language Technology,* 1996. Retrieved July 7, 2004, from http://cslu.cse.ogi.edu/HLTsurvey/ch8node7.html
5. D. W. Oard, "Alternative approaches for Cross-Language Text Retrieval", 1997. Retrieved July 7, 2004, from http://www.ee.umd.edu/medlab/filter/sss/papers/oard/paper.html
6. C.J.P. Van Laer, "The applicability of comparative concepts", in *Electronic Journal of Comparative Law*, vol. 2.2, August 1998.
7. R. Sacco, " Droit et langue", in *Rapports italiens au XV Congrès international de droit comparé*, Milano, 1998.
8. J. Kerby, "La traduction juridique , un cas d'espèce", in Jean Claude Gémar (Ed.), *Langage du droit et traduction, Essais de jurilinguistique*, Montréal 1982.
9. N. Bel, C. Koster, M. Villegas, "Cross-Lingual Text Categorization" in *Proc. of ECDL 2003*, pp. 126-139.
10. Swish-e, Simple Web Indexing System for Humans – Enhanced, Retrieved July 7, 2004, http://swish-e.org.
11. W. Lee, S. Sugimoto, M. Nagamori, T. Sakaguchi, K. Tabata, "A Subject Gateway in Multiple Languages: a Prototype Development and Lessons Learned", in *Proc. of DC Conference* 2003, pp. 59-66.
12. F. Sebastiani, "Interactive Query Expansion with Automatically Generated Category-Specific Thesauri", in Amita G. Chin (ed.), *Text Databases and Document Management: Theory and Practice*, Idea Group Publishing, Hershey, US, 2001, pp. 103-117.
13. D. Yarowsky, "Word sense disambiguation using statistical models of Roget's categories trained on large corpora", *in COLING 14,* 1992, pp. 454-460.
14. M. Lensk, "Automatic sense disambiguation", *in Proc. of the 1986 SIGDOC Conference*, pp. 24-26.
15. I. Dagan, A. Itai, "Word sense disambiguation using a second language monolingual corpus", *Computational Linguistic*, n. 20, 1994, pp. 563-569.
16. W. A. Gale, W. K. Church, D. Yarowsky, "A method for disambiguating word sense in a large corpus", *Computer and Humanities* n. 26, 5, 1993, pp. 415-439.