Ontology for economics and Information Retrieval

Mercè Lorente Casafont

Citación recomendada: Mercè Lorente Casafont. *Ontology for economics and Information Retrieval* [en linea]. "Hipertext.net", num. 3, 2005. <http://www.hipertext.net> [Consulted: 12 feb. 2007]. .

1. Introduction

When talking about information retrieval (from now on IR), we may be referring to very different objectives. As with the majority of scientists, being linguists makes us privileged users of many of these techniques. Our participation in IR research and development projects however, does not tend to be as great as that of IT professionals and documentalists. There are two main reasons for this: 1) The main aim of most IR systems lies in document retrieval and the automatic classification of documents 2) Mathematical methods have proved to be highly profitable in terms of their costs to efficiency ratio. Any method that's independent of language and subject matter will be preferable to others, particularly when dealing with large volumes of data, provided it's results in terms of precision and recall are satisfactory.

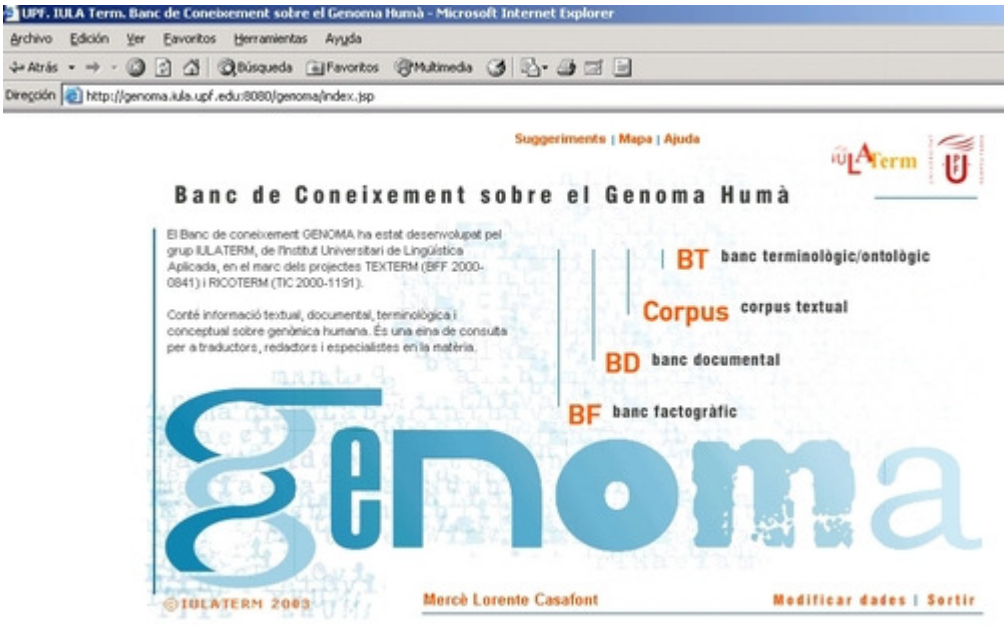
The incorporation of linguistic strategies has however been growing in certain IR projects. The primary reason for this is that some IR projects have been dealing with new objectives that lie beyond the recovery of indexed documents. Automatic abstract generation, automatic terminology extraction and data mining are examples of the new aims. These new systems search for specific information within texts or se ts of texts. This makes elements such as lexis, lexical combinations and text structure acquire a renewed importance. The second reason for the involvement of linguistic work in IR projects is related to the results brought about by these techniques. In addition to the usual evaluation criteria of precision and recall, relevance has become a criteria of the utmost importance. The meaning we attribute to the term relevance is one that implies greater precision in retrieval. While the precision criteria regulates the system's capacity for retrieving documents which are pertinent to the query that has been made, the concept of relevance refers mainly to the system's capacity for retrieving semantically and pragmatically adequate information within the pertinent documents. For example, let us suppose an inquiry about mines obtains a 70% precision level in documents on mining. From this point on, the objective is to apply a complementary measure to make sure that, within these documents on miming, the information we're recovering is roughly what we we are looking for. In other words, the relevance criteria would be the adaptation of the precision criteria when applied to the automatic extraction of terminology out of the terms found in the text, the later being treated as knowledge nodes. Various experiments have proved that the combination of statistical and linguistic techniques leads to a significant improvement in precision results in the retrieval of documents and in the retrieval of precise information within these texts.

In this article we wish to present RICOTERM-2 [1] , a project coordinated by the university of Santiago de Compostela and the Universitat Pompeu Fabra of Barcelona, in which researchers from the University of The Basque Country and Antwerp University have also participated. The main aim of this applied research project, is the design and development of a prototype queries reelaborator for Internet searches (RECCI). This application consists of an interface located in an economics specialized web portal which will transform a simple monolingual search into a group of multilingual searches expanded towards other concepts related to the initial search. For this project Catalan, Basque and Galitian, in addition to Spanish and English, are being used as working languages.

During the first phase of this project our aim is to construct a modular economics knowledge

bank, composed of a textual corpus for each of the working languages, a computational dictionary enriched with semantic and terminological information, and an ontology covering a restricted area of economics. The model to be followed for the construction of the modular bank is that developed in previous genomics projects by the IULATERM [2] group, the structure of which can be seen in the figure bellow:

Figure1: Access screen showing the different modules of the genomics knowledge bank of the IULATERM group.



The economics bank however, having a greater number of working languages, is expected to be smaller. It will not contain the documentary and factographic module, and given that its utility will arise mainly from experimentation with the reelaborator, it does not intend to be exhaustive . The content of each one of its modules will be limited both quantitatively and thematically.

Another relevant feature of the economics knowledge bank is that it is based on the idea of the reutilization of resources in such a way that new textual corpus's and lexical data bases are not been constituted. A variety of cooperation agreements for the usage of existing materials developed for other projects (economics dictionaries, dictionary and corpus in Galitian , dictionary and corpus in Basque) and for the usage of Galitian and Basque processing tools are being reached for this purpose.

2. The expansion of IR queries

Experiments in query expansion are based on methods which take a user query as their starting point for detecting that which lies at the heart of the query (the informational need of the user), and subsequently expand on it with variations which allow us to find the documents which contain the most relevant information. The first difficulty facing these experiments is that usually the query does not supply sufficient data to distinguish between relevant information and that which is thematically related to it (Strzalkowski et al. 1999:136).

Throughout the 1990's the most frequently used query expanding methods have been only-term expansion and full-text expansion. In the first case probability techniques establish a greater weighting for frequent terms within relevant documents. This is usually completed by using other automatic methods, such as the relevance feedback method, or manual methods, such as user involvement in relevance determination. In the second case queries are expanded through the use of sentences or paragraphs extracted out of documents considered relevant or not.

In both cases, the most important challenge is managing to control the semantic and pragmatic content. This is not always explicit through the forms of the texts. Natural language processing still requires much development in semantic and pragmatic representation. Although much progress has been made over the last few years, further advances will be necessary if we want to avoid expansions being based solely on the identification of lexical units and the establishment of correlations with morphological variations, interlinguistic equivalents, or lexical combinations. The aim of RICOTERM-2 is to engage the design of the economics query reelaborator taking as a starting point the two main methods: only-term expansion and full-text expansion.

2.1. Full-Text Query Expansion

Systems based on full-text query expansion tend to present satisfactory results showing around 40% precision (Strzalkowski et al. 1999:136). These results are too low for professional users searching for specific information who have very different expectations than those of ordinary users simply surfing the web. The proposed solutions for result improvement are based not only on query completion and verification through manual user intervention and automatic techniques, but also on restriction criteria applied to the texts utilized for the expansion:

- Structural restriction: a structured set of texts.
- Thematic restriction: only documents pertaining to a specific field

The MURAX project (Kupiec 1999:314) provides a good example of structural restriction applied to query expansion. It is based on a questions and answers system which is used to elicit concrete answers to concrete questions of the kind "Who killed President Kennedy?". MURAX expands the query to locate partial information in the various texts and combines it to produce an ideal answer. The system is presented as a combination of methods which includes a superficial syntax analysis for the extraction of phrases. What is most interesting to note here is that the texts used for expansion are taken from the Grolier Encyclopedia (1990). As any good l exicographica l product, the Grolier Encyclopedia presents simple and systematic definitions and controls lexical remmissions (paraphrasis, synonims, hyperonims). The system experiments which take the game of Trivial Pursuit as their point of departure would not obtain the same results if the texts for the extensions belonged to very differently structured documents such as contracts, scientific articles, conferences, laws, dialogues, etc.

We have come to considering that, in the line of combining methods, good results for query expansion in selected areas could arise from using thematically selected structurally and linguistically labelled corpus texts. These specialised corpus texts (such as the Technical Corpus of the Institut Universitari de Lingüística Aplicada (http://www.iula.upf.edu) would present limitations in terms of extension and complexity when compared to the structured concise character of definitions in an encyclopedic dictionary such as in the MURAX project. However, continuing with the same combination of techniques and resources for RICOTERM-2 we envisage two complimentary textual strategies for query expansions:

- The use of an automatic term extraction tool (relevant lexis for the restricted field) and the use of a conceptual relationship detector. This would allow for the identification and labelling of terms in context.
- The use of terminological databases and specialised dictionaries, particularly those in the field of economics and those containing definitions.

2.2. Ontology based query expansions

The query expansion systems that have managed to make significant improvements in results over the last few years are those that interact with ontologies and lexical hierarchies. Ontologies are formal constructions that represent conceptual nodes and express the way in

which these are conceptually related to one another. Given that their aim is not so much to classify and locate documents as to organise 'the concepts of the world' [3] and relate them to the linguistic expressions that carry them, their complexity and atomisation tends to be greater than is the case in thesauruses. Further on we shall be referring to the theoretical and practical problems arising from the construction of ontologies, but for the moment let's have a look at the various contributions that ontologies have made to query expansion systems.

There is a growing number of voices claiming that the use of ontologies for information retrieval is an efficient method that can be superior to others in both precision and relevance. The query expansion systems that make use of ontologies and alternatively of thesauruses are based on only-term expansion. The lexis identified as relevant in a document is used as a starting point for establishing correlations with concepts or other lexical units that represent these concepts or related concepts. From a linguistic point of view we could say that from one word, syntax unit or set of words from the query, the system would search in the ontology for other words or phasal units expressing related concepts. This proximity would correspond in linguistics (and as a general rule) to synonyms or variations, with hyperomyms or concept classes, and with cohyponims or concepts that belong to a same class; in this kind of system we may find it very hard to find other kinds of conceptual relationships such as causality or time sequence.

Wordnet is one of the most popular resources used as ontology in query expansion based IR. The way lexical hierarchy structures the lexis of languages is based on the concept of synset or synomym set. For each query vector the query expansion systems automatically asociate sets of synonyms (Voorhees 1994: 223). One of the first problems to be detected in this kind of system is that being a general use resource, non thematically restricted, simple queries seem to be subject to ambiguity (polysemy for linguists). This limitation tends to be overcome in experimentation by manual intervention on behalf of the user, who selects the relevant synset (Voorhees 1994: 223), as well as by *machine learning* techniques.

The only alternatives to generalist ontologies such as WordNet that are capable of producing better results in terms of precision and more importantly in terms of relevance are based on confining the scope of the ontology or lexical structure. There are cases of ontologies generated ad hoc from words and from the concepts used within an organisation (company, public sector body, work network). These experiments which work with controled lexis and little data are not easily exported to other organisations, languages or open IR on the internet (Stenmark 2003: 9).

Other projects on the other hand confine their reach to a specialised field. These fields generally have a consolidated tradition within the scientific community, and therefore an ample consensus on concepts, conceptual relations and standardised terminology tends to exist within them. This is the case for the exemplary achievements in IR in medicine that we ilustrate here with the reference to a query expansion project based on association to concepts in the UMLS metatheasurus (Aronson et al. 1997: 485).

The UMLS metathesaurus is part of the resources developed by the National Library of Medicine, it contains information about concepts and biomedical terminology in several languages and develops automatically from controlled vocabularies and classification systems (Feliu et al. 2002: 24). In its 2001 version it had a capacity for 800.000 concepts and 1.400.000 sayings or different words. We may venture to predict that the positive results deriving from the system when experimented with a set of selected texts, some of which (but not all) were indexed, might reoccur when applying the system to a different set of selected specialised texts and a completely different ontology corresponding to a different domain of specialisation. This project in fact concludes that the combination of UMLS-based and MEDLINE quotes based query expansion methods, which include both probabilistic and linguistic techniques, should lead to further improved results.

In accordance with the reutilisation of existing resources criteria, in RICOTERM-2, efforts are been made in the localisation of accessible and format compatible conceptual structurings about economics which might be easily incorporated into a base ontology. It has also become apparent to us that the presence of concepts in this restricted field is very high in Wordnet,

more so than for other less heavilly represented specialised fields. Another criteria guiding ontology building in RICOTERM-2 is precisely the automatic acquisition of concepts and conceptual relations from real texts. In this sense the use of the automatic term extractor and conceptual relations detector, to which we refered in the previous section, seems to be essential.

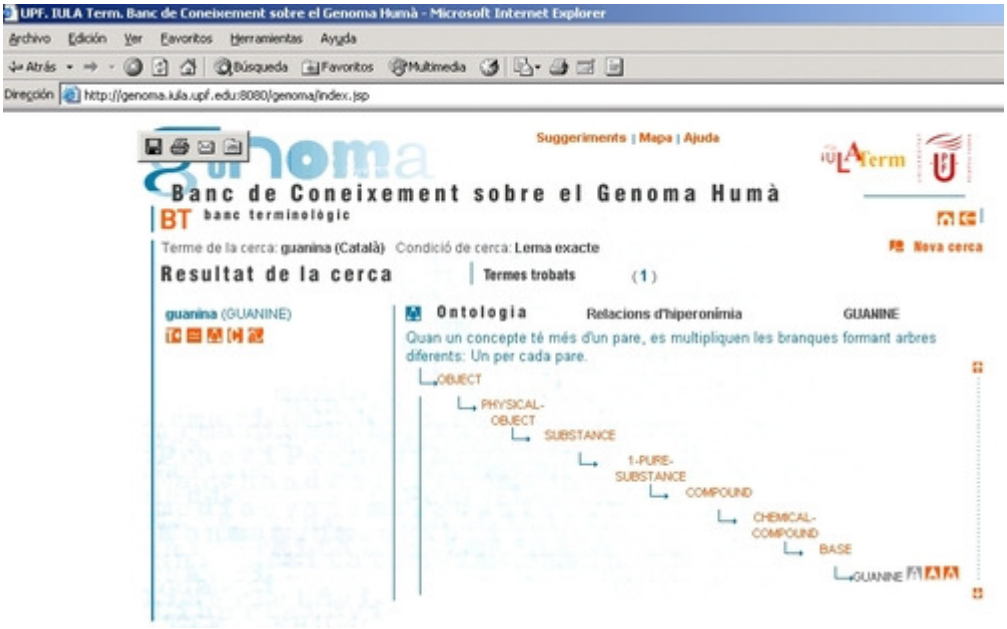
3. The construction of ontologies

We wouldn't like to leave aside an issue that often arises in these kind of applied linguistics projects. On the one hand it has become undeniable that the use of ontologies, particularly in the case of restricted and specialised domain ontologies, allows for precision and relevance improvements in both query making and in IR system responses. On the other hand, however, much reflection is taking place over the theoretical and descriptive appropriateness of this type of resource. We will now take a look at some of the advantages and problems that these bring about.

3.1. The advantages of specialised field ontologies

Specialised field ontologies present a very high degree of granularity. Classes of concepts, nodular concepts and diverse conceptual relations arising from the processes of conceptualisation and categorisation carried out in a given specialised field, are all represented. In this sense they are no doubt superior to documentary thesauruses in which generally hierarchical and associational relations predominate and the primary function of which is to establish relationships to pre-established categories. At least on a theoretical level, ontologies allow us to approach a specific subject's concepts in a more open manner, making use of a greater range of conceptual relationships, as these are not restricted from the outset by a particular practical aim. In order to ilustrate granularity we wil now look at an example of the ontology of the human genome which we reffered to earlier.

Figure2: Screen visualisation of the ontology (example of guanine hyperonymia relationships).



Ontologies belonging to specialized fields of long academic and professional tradition (and it is not by chance that we express this limitation) show a high degree of stability. Although disciplines such as medicine or biology have experimented drastic changes, developments and innovations, this does not however mean that these novelties completely invalidate earlier conceptual organizations. An ontology about pathologies or pharmacy may well be affected by scientific advances (concept location, properties assigned to concepts, creation of new concepts, fragmentation of concepts) but it is much less likely that it will be reformed in its totality. The

reason for this stabilization in a discipline's ontology, as for the stabilization of used terminology , is the high level of international consensus that some of these disciplines demonstrate. This is favored by the swift transfer of knowledge and by the control exerted by standardization bodies.

The sources used for the construction of specialized field ontologies tend to be examples of controlled communication. Be it scientific articles, encyclopedias and dictionaries, academic manuals, documentation used by professionals, conferences and scientific presentations, technical documentation or expert knowledge of specialized users, the contents and their expression tend to adapt to the formal discourse register. Moreover, written text, and oral ones which have been elaborated previously for reading or semi-improvisation, are subjected to various revision processes by their authors or by linguistic mediators (translators, editors, proof readers).

To summarize we may say that granularity in conceptual representation, stability of these concepts in the scientific community, and the formality of the overwhelming majority of documents transmitting these concepts and their relationships, make specialized field ontologies a reliable resource for the retrieval of information, as well as a more effective one than its counterparts of non restricted fields and those used for common language.

3.2. Problems with ontologies

We cannot however expect specialised field ontologies to solve all problems related to specific information retrieval. User satisfaction level results may improve with the development of new specific ontologies, and more so with the use of combined techniques, but we cannot ignore the very development of ontologies presents deep seated problems to which we'd like to refer briefly in this section. For this we will take as a starting point a couple of contradictions or paradoxes.

b) The dynamism linguistic paradox. From a linguistic perspective, concepts don't exist independently of the term or term set that carries them. As a case in point, the only truly onomasiological [4] action is the creation of terms by specialists in a specific discipline (scientists who discover an element, engineers who construct an artifact, professionals who wish to distinguish a partially new element from other existing ones, etc.) , and this is always done from within a specific language. Furthermore, the current terminology (Cabré 1999: 149) assumes that terms, as all language signs, are not independent from language and that specialized communication is conditioned by social and cultural aspects. In this sense one of the first paradoxes confrontingus as ontology builders working with terminology and sharing this socio-linguistic vision, is precisely the use of managers and interfaces that use English as if it were a lingua franca to which we can associate the rest of working languages without any problem; or even worse, as if it were simply the neutral representation of a common conceptual structuring.

Needless to say, applied linguistics, with its eminently pragmatic character, won't stop building ontologies because of such a contradiction. It will most likely look for methodological alternatives that are better at satisfying users and the builders themselves, such as the use of alphanumeric characters or considering the ontology itself as an autonomous lexical hierarchy [5] for each working language.

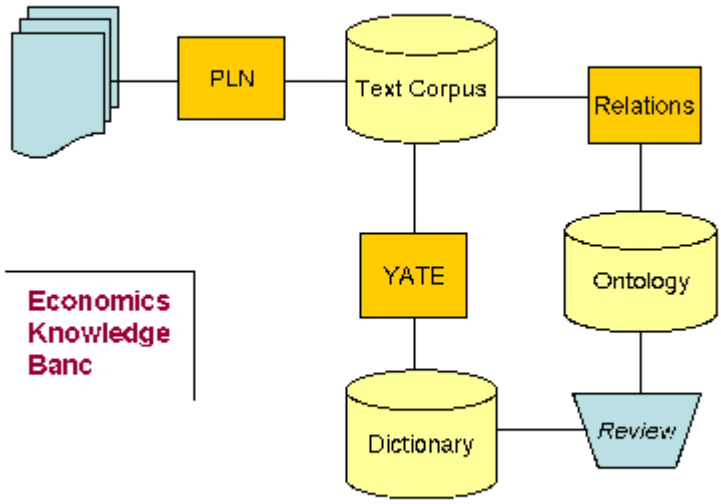
b) The dynamism of concepts. The contradiction that exists between the elaboration of an ontology as a stable representation and the view that concepts and concept structurings are dynamic entities has been brought to the fore not only by philosophers but by linguists and terminologists too. Even in informatic medicine and in some economics forums (Vromen, 2004:213) it has become apparent that there exists an open debate between advocates of the use of ontologies for the representation of information, and their critics, who perceive limitations in ontologies for representing, for example, evolutionist economics concepts. As for dealing with the changes brought about in the revision of concepts and the creation of new ones by the passage of time, applied linguistics only provides a continued defense of the need to

update resources permanently. This solution however presents a difficulty which tends to be unsurmountable: that of associated costs. For the representation of dynamic concept ontologies , responses are still in the early stage of first attempts (hyperlinks, treatment of polysemy and ambiguity, diffuse objects, etc.) and we should still wait some time before analyzing specific suggestions with real examples.

4. The first RECCI system designs

The RICOTERM-2 project, as introduced at the beginning of this paper, has as its primary focus the design and elaboration of a prototype for the RECCI (Query Reelaborator for Internet Searchers) system. We chose to implement the first version of the prototype in relation to the retrieval of economics information for three reasons: firstly because this discipline already has specialized textual and lexical resources; secondly because we have the capacity to compile textual corpuses in each of the working languages; and thirdly because we consider economics to be a prototypical subject in terms of the need for updated information on behalf of users of information retrieval systems. The agents that might be interested in this kind of system located in a specialized portal could be professional societies, media, public administrations, stockbrokers, banks and building societies, as well as other companies and groupings of professionals.

Figure 3: Project Ricoterm-2's Economics Knowledge Bank architecture.



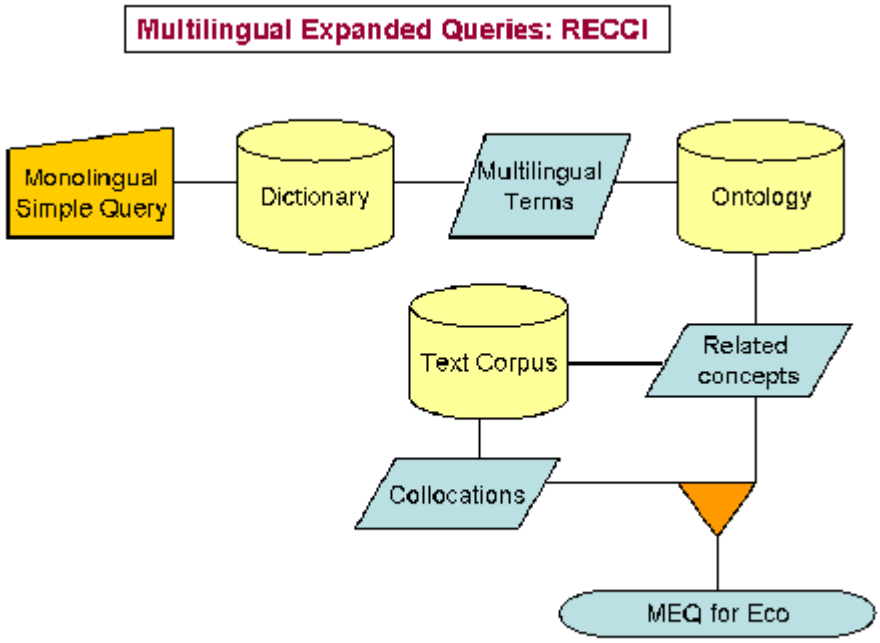
The previous illustration shows the economics knowledge bank architecture designed as a testing ground for the RECCI system. The textual corpuses for each working language are structurally labeled and linguistically processed. This will allow us to extract information by lexical unit and grammar category, as well as by broader sequences and syntax patterns.

Although the recycling of existing conceptual structurings and economics dictionaries has been foreseen , we do not rule out, as part of the construction of the knowledge bank, the exploiting of labeled textual corpuses for the automatic and semi-automatic extraction of terms and the conceptual relations linking them.

It is for this reason that one of the secondary aims of the project consists in the adaptation of YATE for economics and the different working languages. YATE is an automatic extractor of term candidates which is language independent and which combines probabilistic and linguistic

strategies (Vivaldi 2001). The integration of a dictionary of linguistic expressions carrying conceptual relations has been envisaged as a way of broadening the range of linguistic strategies available to this tool (Feliu 2004).

Figure 4: Query expansion design for the RICOTERM-2 Project.



This figure illustrates the original query expansion process design. We start from the premise that users, although they may be professionals, usually make simple queries in a single language. The system foresees the broadening of this simple monolingual query through the use of dictionary queries, with equivalent terms in other languages and with synonymic variations. This multilingual and varied query enlargement, taken as an entrance point for the ontology would allow us to detect other associated terms. By providing an entrance into the corpus query tool, it would also help us detect collocations and linguistic combinations in context.

To conclude we may say that the design of the RECCI system is one that corresponds to the expansion of multilingual queries combining different techniques. In addition to mathematical methods, to which we have made reference in this article, the linguistic techniques utilized will be both the lexical base expansion and expansion through corpus. We consider that the use of a modular knowledge bank for the field of economics which includes linguistically processed texts, dictionaries and an ontology, will allow the query reelaborator to supply good consistent results for multilingual and complex queries. These should reflect ideal questions made by professionals to launch common search engines.

The project will continue in the following phases with system experiments using real data from the economics knowledge bank and with evaluations of the degree of satisfaction attained by a group of pilot users made up of documentalists and economists.

5. Summary

In addition to presenting the first steps of the RICOTERM-2 project, this article has sought to provide answers to some of the main preocupations arising in decision-making for this kind of applied projects.

We have presented this project as one that is attempting to unite efforts in the development of linguistic resources for specialised domains and in the design of information retrieval systems capable of improving user satisfaction and increasing their expectations about the precision of retrieved information. Following this line we defend three key aspects of the aplied project: The recycling of existing resources, the project's orientation towards specialised fields of non restricted use, the integrating mutilingualism of the official languages of the Spanish state and the adaptation of resource design to real (not ideal) user profiles.

In order to complete the presentation of the project we have introduced the first models of the architecture of the economics knowledge bank and its modular components and we have presented a first draft of the functioning of the multilingual query expansion.

We have also taken advantage of the occasion to revise some of the questions which need to be addressed in order to proceed with this kind of design. These range from a typology of existing query expansions to reflections on the adequacy of using ontologies in a communicative based terminology theoretical framework.

6. Bibliography

Angelova, Galia (2000) 'Ontologies for Natural Language Processing Applications'. In: Simov, Viril; Kiryakov, Atanas (eds.) (2000) *Ontologies and Lexical Knowledge.* Sofia: OntoText Lab, Sirma AI EOOD, p. 1-15.
Aronson, Alan R.; Rindflesch, Thomas C. (1997) 'Query Expansion Using the UMLS Metathesaurus'. In: *Journal of the American Medical Informatics Association* , 1997: Supplement. Proceedings of the 1997 AMIA Annual Fall Symposium, p. 485-489.
Buitelaar, Paul (2000) "Semantic Lexicons: Between Terminology and Ontology". In: Simov, Viril; Kiryakov, Atanas (eds.) (2000) *Ontologies and Lexical Knowledge.* Sofia: OntoText Lab, Sirma AI EOOD , p. 16-24.
Cabré, M. Teresa (dir.) (2004) Banc de coneixement sobre Genoma Humà. IULATERM . Consultado: 28-feb-2005, http://genoma.iula.upf.edu:8080/genoma/index.jsp
Cabré, M. Teresa (1999) *La terminología: representación y comunicación.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. ISBN 84-477-0673-7.
Feliu, Judit (2004) *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra , 2004. ISBN: 84-89782-19-9. [CD-ROM]
F eliu, Judit; Vivaldi, Jorge; Cabré, M. Teresa (2002) *Ontologies: a review.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra , Papers de l'IULA, Sèrie Informes, 34.
Kupiec, Julian M. (1999) "MU