

CABRÉ, M. T. (2003) «State of the Art on Computer Texts: Documentation, Linguistic Analysis and Interpretation. The borders of applied linguistics». En: Veneziani, Marco (2003) *Informatica e Scienze umane. Mezzo secolo di studi e ricerche*. Serie Lessico intellettuale europeo, vol. 94. Florencia: Leo S. Olschki Editore, p. 43-56. ISBN: 88-222-5225-X (CL).

State of the Art on Computer texts: Documentation, Linguistic Analysis and Interpretation

The borders of Applied Linguistics

M. Teresa Cabré
IULA (Institute for Applied Linguistics)
Universitat Pompeu Fabra
teresa.cabre@trad.upf.es
<http://www.iula.upf.es>

Let me to thank the organizers of this meeting, and mainly Dr. Tullio Gregory, to trust me for this presentation. I am not sure I am the most adequate person for speaking about the borders of the Applied Linguistics. I am only the director of a young Institute on Applied Linguistics. We mean Applied Linguistics in broad sense: whatever knowledge domain where linguistics studies are oriented to solve social problems in communication.

And it is from my experience and about our experience I am going to speak.

Introduction

The University Institute of Applied Linguistics, belonging to the Pompeu Fabra University at Barcelona, is an institution devoted to technical and scientific research as well as to graduate training.

In the Institute's research framework we deal with a broad concept of applied linguistics including the creation of different resources, the creation of varied tools, the design of dictionaries and the design of workstations.

Our main lines of research concentrate on terminology, language engineering, lexicography, computational linguistics, linguistic variation, knowledge representation and, finally, neology.

As for the infrastructure and the human resources we are organised in research working groups which are the following:

IULATERM (directed by M. Teresa Cabré)
Language Engineering Research Group (directed by Lluís de Yzaguirre)
Lexicography group (directed by M. Paz Battaner)
Computational Linguistics Group (directed by Lluís de Yzaguirre and Núria Bel)
UVAL (Linguistic Variation Research Unit) (directed by M. Teresa Turell)
Knowledge Representation Group (directed by Lluís Codina)

Observatori de Neologia (Neology Observatory) (directed by M. Teresa Cabré).

Each of these groups is constituted by a group of people working for a particular research subject in order to obtain some particular goals depending on the public funding received by every group.

As already mentioned, one of the main goals of the Institute is to create linguistic resources. In this sense, we have been developing a multilingual corpus of specialised domains in five different languages. Moreover, we are building some other kinds of corpora: press corpus, parallel corpora, and spoken corpus.

Also in the linguistic resources creation, we are devoted to the construction of terminological, lexical and neological databases which are continuously updated.

Corpus processing and results

Let us look in detail which are the main characteristics of each type of corpus we have just mentioned. As for the design criteria of the multilingual corpus of specialised domains, the corpus is built on the basis of specialised written language texts concerning different specialised domains (Law, Economics, Medicine and Human Genome, Environment and Computer Science) in five different languages which are Catalan, Spanish, English, French and German. The corpus has a flexible organisation and the tagging process follows the standard ISO 8879 on SGML. Texts are classified according to the domain classification and the text typology we apply in order to identify the documents in the exploitation stage. See the following table for numerical information about the specialised corpus.

| Area | Catalan | Spanish | English | French | German | Total |
|--------------------|-------------|-------------|-------------|------------|------------|--------------|
| Law | 1518 | 2137 | 548 | 44 | 16 | 4263 |
| Economics | 1821 | 1000 | 285 | 78 | 27 | 3210 |
| Environment | 1487 | 1000 | 586 | 230 | 429 | 3733 |
| Medicine | 1487 | 1910 | 303 | 27 | 198 | 3925 |
| Comp. Scien. | 654 | 1024 | 363 | 194 | 83 | 2318 |
| Total . . . | 6967 | 7071 | 2085 | 573 | 753 | 17448 |

Concerning the press corpus, we treat two journals *El País* and *Avui*, which follow a full computer recording and mark-up process. The processing is done on eight daily editions per month and neologism candidates are automatically retrieved by matching against reference dictionaries. These neologism candidates must be manually validated by collaborators in the project. See the following table for information concerning this neologism extraction process.

| Área | Catalan | Spanish | Total |
|------|---------|---------|-------|
|------|---------|---------|-------|

| | | | | | | |
|---------|------|------|------|------|------|--------|
| | Doc. | Occ. | Doc. | Occ. | Doc. | Occ. |
| general | 199 | 9434 | 46 | 1977 | 245 | 11.411 |

Finally, as for corpus is concerned, we must talk about the parallel corpus. It is worth mentioning that one part of the corpus is composed by parallel texts. The most outstanding linguistic pairs are: Catalan-Spanish, Catalan-English, and Spanish-English. See the following table for the present state-of-the-art.

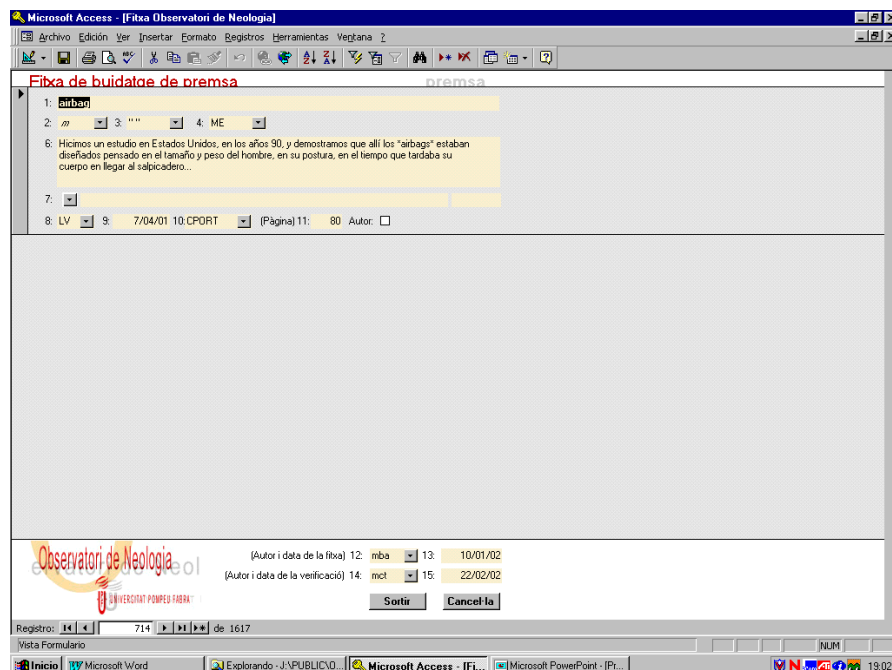
| Area | Catalan-Spanish | | Catalan-English | | Spanish-English | |
|--------------------|-----------------|--------------|-----------------|------------|-----------------|------------|
| | Doc. | Occ. | Doc. | Occ. | Doc. | Occ. |
| Law | 61 | 460 | 1 | 12 | 2 | 57 |
| Economics | 21 | 600 | 10 | 250 | 13 | 283 |
| Environment | 10 | 214 | 11 | 213 | 13 | 144 |
| Medicine | 4 | 108 | 1 | 40 | 4 | 125 |
| Comp. Scien. | 1 | 28 | - | - | 23 | 300 |
| Total . . . | 97 | 1.410 | 23 | 515 | 55 | 909 |

Terminological and Lexical Databases

A part from the corpus, and as we have already mentioned, some of our projects are also devoted to the construction of terminological and lexical databases. As for the terminological databases, we have been working in a project gathering together all terminological databases built by final-year degree students working on terminology dictionaries and glossaries. See the following example built using the terminology management application called MultiTerm:



Concerning lexical databases, we have built a neological database where all neologisms retrieved by texts are collected in an organised way. Next image shows an example of a neological entry contained in the lexical database.



Tools for corpus processing

The corpus processing concerning text retrieval and structural mark-up starts with the text recovering process both in paper and in electronic format. If the text is in paper format it has to be scanned in order to obtain its electronic version. After the spell checking, the process is the following: first, the text is tagged according to the SGML standard; second, it is delimited sentence by sentence and, finally, it is processed by the SGML parser. See next fragments of a text including structural mark-up and morphological parser:

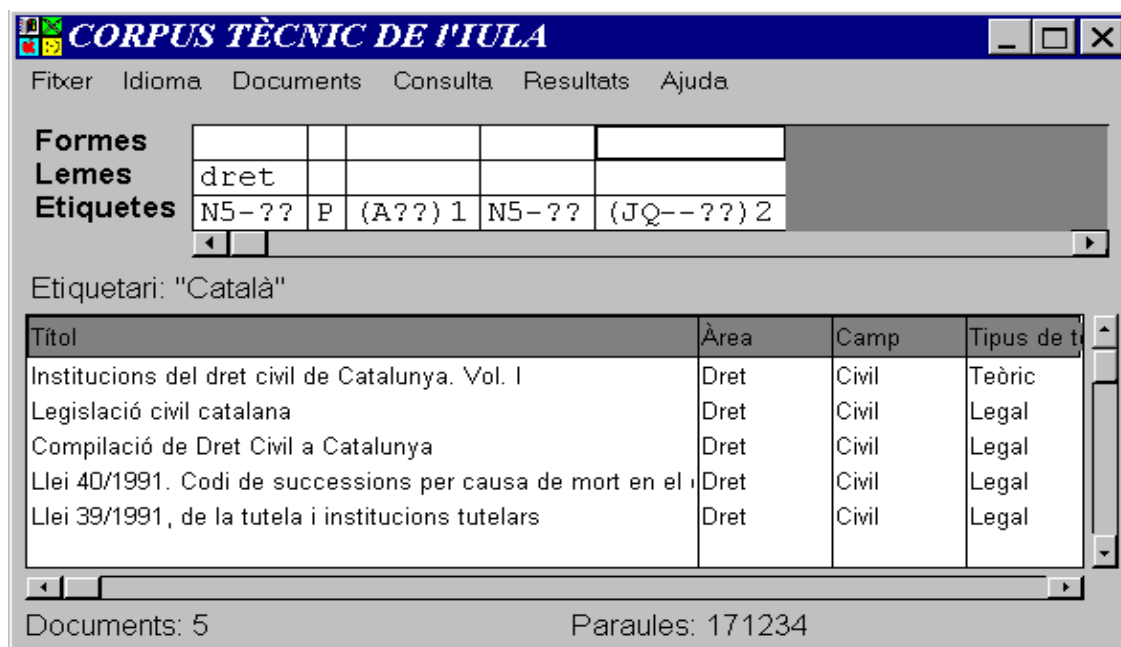
```
<div1 n=6 complete=n>
<head type=main>DRET DE LA NAVEGACI&Oacute;</head>
<div2 n=72>
<head type=main>ESTATUT JUR&iacute; DIC DEL VAIXELL I DE
L'AERONAU</head>
<div3 n=1>
<head type=main rend=il>Concepte i naturalesa jur&iacute; dica</head>
<p><s>En sentit t&egrave; cnic parlem de vaixell per referir nos a qualsevol
construcci&oacute; destinada a la navegaci&oacute; mar&iacute; tima o fluvial.</s><s>A
aquesta idea atenen tamb&eacute;., en general, els ordenaments positius moderns en
determinar, amb major o menor amplitud, la noci&oacute; jur&iacute; dica del
vaixell.</s></p>
<p><s>En el nostre ordenament legal, l'article 146 del Reglament del Registre Mercantil
de 1956, transit&ograve; riament vigent (veg. disposici&oacute; transit&ograve; ria sisena
del Reglament del Registre Mercantil de 29 de desembre de 1989), suplint la llacuna del
Codi, estableix, tamb&eacute; en aquest sentit, que "es reputaran vaixells, per als efectes
del Codi de comer&ccedil; i d'aquest Reglament, no nom&eacute; s les embarcacions
destinades a la navegaci&oacute; de cabotatge i altura, sin&oacute; tamb&eacute; els dics
flotants, pontons, dragues, g&agrave; nguils i qualsevol altre aparell flotant destinat o que
pugui destinar se a serveis de la ind&uacute; stria o comer&ccedil; mar&iacute; tim
```

| | | | | | |
|-----|-----|--------------------|-----|--|--|
| ## | TAG | <s> | | | |
| 20 | TOK | En | BOS | en\Pen\AMS pr\REE7--- | |
| 21 | TOK | sentit | | sentir\HMS sentit\N5-MS | |
| 22 | TOK | tècnic | | tècnic\JQ--MS tècnic\N5-MS | |
| 23 | TOK | parlem | | parlar\V7R1P- | |
| 24 | TOK | de | | de\P | |
| 25 | TOK | vaixell | | vaixell\N5-MS | |
| 26 | TOK | per | | per\P | |
| 27 | TOK | referir | | referir\VI---- | |
| ##- | DLI | - | | =\DELIM | |
| 28 | PGR | nos | | pr\REE616P | |
| 29 | TOK | a | | a\P | |
| 30 | TOK | qualsevol | | qualsevol\EN--6S qualsevol\N5-6S | |
| 31 | TOK | construcció | | construcció\N5-FS | |
| 32 | TOK | destinada | | destinar\VC--SF | |
| 33 | TOK | a | | a\P | |
| 34 | TOK | la | | el\AFS pr\REEC3FS | |
| 35 | TOK | navegació | | navegació\N5-FS | |
| 36 | TOK | marítim | | marítim\JQ--FS | |
| 37 | TOK | o | | o\C | |
| 38 | TOK | fluvial | | fluvial\JQ--6S | |
| --- | DLD | . | EOS | =\DELS | |
| ## | TAG | </s> | | | |

Talking about the disambiguation process, the corpus follows both a linguistic (ambiguities are solved on linguistic knowledge base) and a statistical disambiguation (ISSCO tool using a training corpus). Afterwards, a first-level syntactic parser is applied to all documents.

Tools for data retrieval

The first tool designed, built and used in the Institute framework is called Bwana. This data retrieval tool was built in order to exploit all information contained in the specialised corpus. See a brief example of its capabilities in the following image:



Bwana is a tool which has to be locally used. For this reason, we have developed an extended application called BwanaNet which works on the web site. In this sense, BwanaNet is more user friendly, faster and results can be captured in an easy way.



Another tool used to retrieve information is SEXTAN. This tool provides a register containing all information automatically drawn from the press corpora which will be later on validated by our collaborators. Sextan provides the neologism, the category, the typographical indication, the kind of neologism, the exclusion corpus or dictionary that has been consulted, the source, the date of creation of the register, the section in the journal where the neologism has been detected, the note, the author indication and the context where the neologism occurs. See the following image to visualise the information given before.



Last tool we would like to mention is YATE. YATE stands for Yet Another Term Extractor and is the final application of two previous Ph. Dissertations which are:

•*Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'extracció automàtica de candidats a unitats de significació especialitzada*, 1999 (rosa.estopa@iula.upf.es)

•*Extracción de Candidatos a Término mediante combinación de estrategias heterogéneas*, 2001 (jorge.vivaldi@info.upf.es)

YATE extracts terminological units from specialised texts by the combination of method strategies and with the use of semantic information. See the following results for the pattern noun and noun-adjective:

Extracted terms (noun)
nefritis
blefaritis
rinitis
bronquiolitis
nefrosis
neurosis
fibrosis
pletismografía
inmunólogo
tuberculosis
bronquitis
otorrinolaringólogo
odontólogo
dermatólogo
hemorragia
. . .

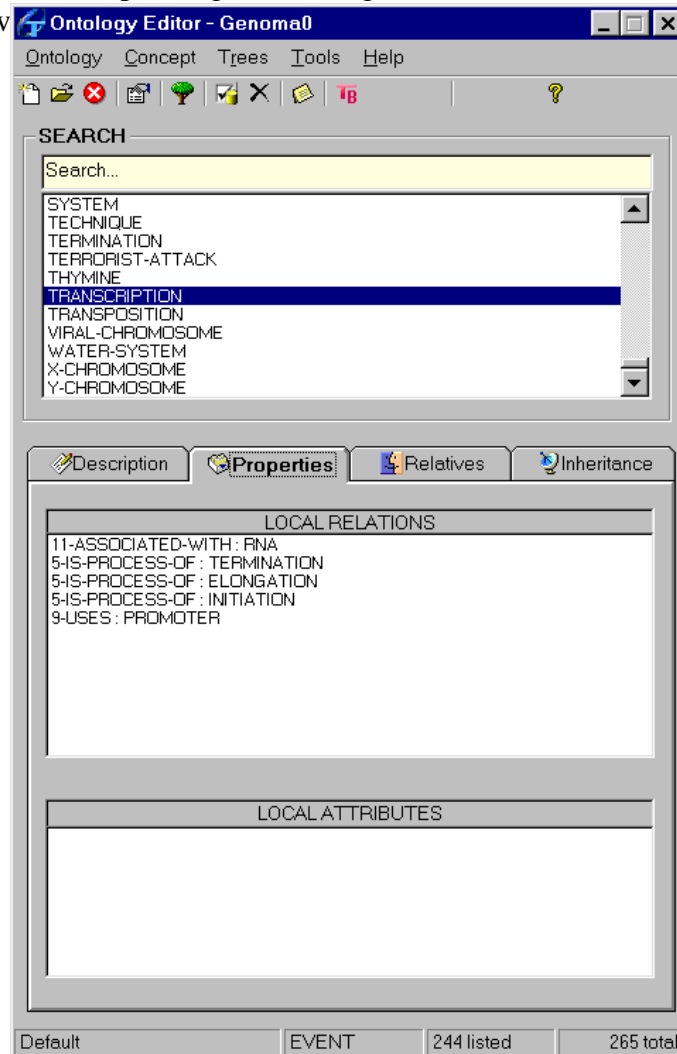
Extracted terms (noun-adjectives)
enfermedad cardiovascular
patología neurológico
glándula mamario
dengue hemorrágico
fibrosis pulmonar
tuberculosis pulmonar
glándula bronquial
enfermedad cardiopulmonar
enfermedad pulmonar
úlceras gástrico
espasmo bronquial
espasmo glótico
enfisema pulmonar
hipertensión arterial
maloclusión dental
. . .

Tools for information management

As for the information management is concerned, in the IULA we work with a dictionary management application which is used in the LSP corpus, the neology detection and the term extraction. A catalogue of main dictionaries are used in order to export data and to generate dictionary forms which are included in mirror dictionaries in order to be reused in new applications.

Moreover, we are also working in the building of a human genome ontology. In this sense, we are using OntoTerm, an ontology management system built by Dr. Antonio Moreno (Universidad de Málaga [amo@uma.es]). The architecture of this management system includes four essential modules which are: the Ontology Editor, the TermBase Editor, the Ontology Navigator and the HLML Report Generator. The system works on Windows and it is a quite user-friendly application allowing the construction of an

ontology and its corresponding terminological database. See the following example of the ontology w



On-going Projects at the Institute for Applied Linguistics

The main present projects we are carrying on are the following:

- TEXTERM
- RICOTERM
- Human Genome Knowledge Base
- DOPO
- Alignment Tool for Parallel Corpus
- Prototype of lexicographical workstation
- Prototype of linguistic planning workstation

We are going to briefly describe each one of the previous mentioned projects bearing in mind that the first two projects join together most efforts done by the collaborators of the Institute because they both are oriented to the construction of the Human Genome Knowledge Base.

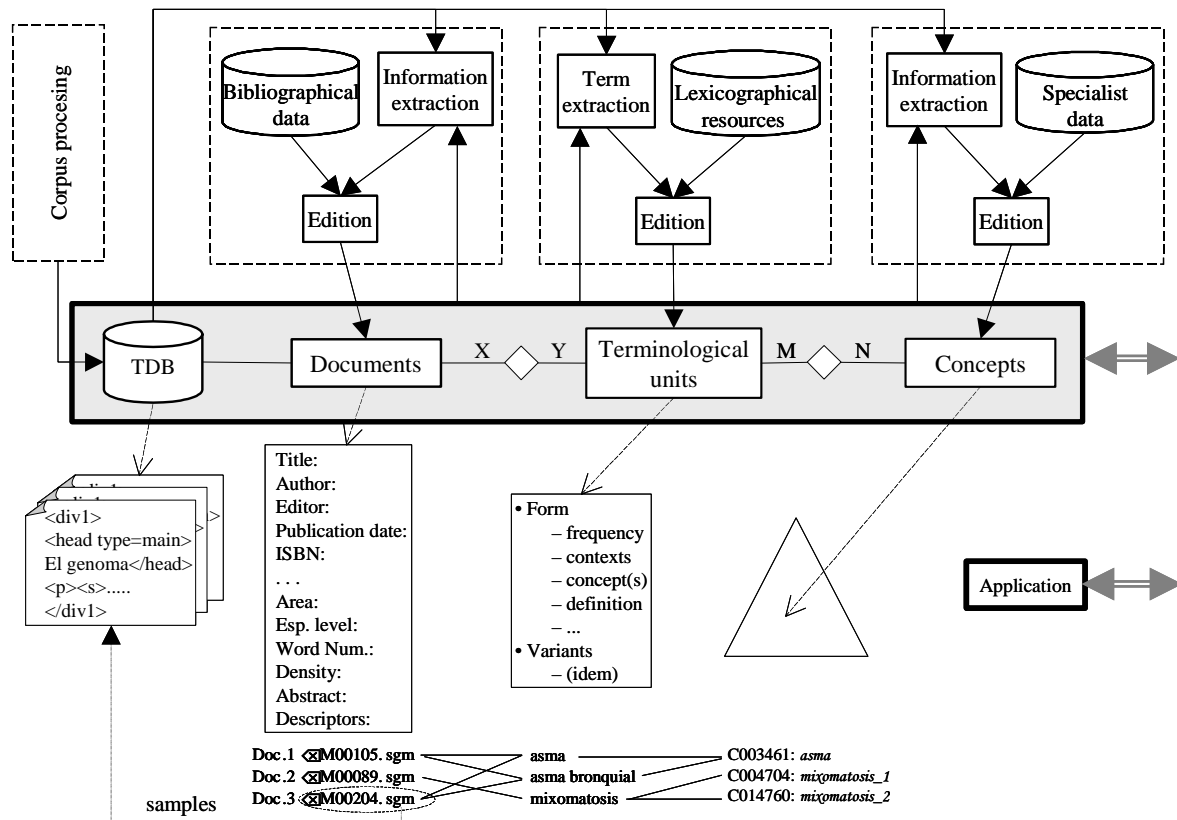
First project to be mentioned is TEXTERM: *Textos especializados y terminología: selección y recuperación automática de la información* (BFF2000-0841) / Specialised Texts and Terminology: automatic information detection and retrieval. The TEXTERM project aims to go a step forward in discourse, grammar and semantic analysis of specialised texts. It is more specifically devoted to the characterisation of the lexical (simple or complex) and phraseological units, which constitute the terminology of those domains, with the final purpose of building an automatic detection system of the cognitive underlying structures in specialised texts.

The main goal of this first project is to provide a sound theoretical basis for computer-aided unit detection, semi-automatic mapping of cognitive nodes and conceptual relations, and the algorithm and protocol designs. It is foreseen that our working methodology —oriented to improve information retrieval (IR) systems— would combine strategies from the cognitive sciences and from linguistics. We will also resort to indexation strategies and thesaurus building standards, coming from information science, and some other linguistic engineering working lines, such as natural language processing and statistical analysis.

Directly related to the TEXTERM Project we must mention RICOTERM: *Sistema de recuperación de información con control terminológico y discursivo* (TIC2000-1191) / An Information retrieval system based on terminological and discourse information control. Main objective of RICOTERM project is to build an information retrieval (IR) system, capable to improve present systems using the terminological control. This control will be achieved with the grammatical, semantic and pragmatic information associated to the units or occurrences that transfer the specialised knowledge. The system will be also improved with the discourse control, drawn from the information concerning intention, communicative purposes, and implicits in a text.

The methodology used combines a tool for natural language processing, which includes structural mark-up, morphological and syntactic analysis, disambiguation, and a terminology extraction system based on formal patterns and lexical ontologies. Ground criteria will be refined by standards for the identification and mark-up of semantic and pragmatic elements within a restricted domain. For this reason, we will refine and update some of the tools used by our group members and to create new ones to complete the modular system of information retrieval.

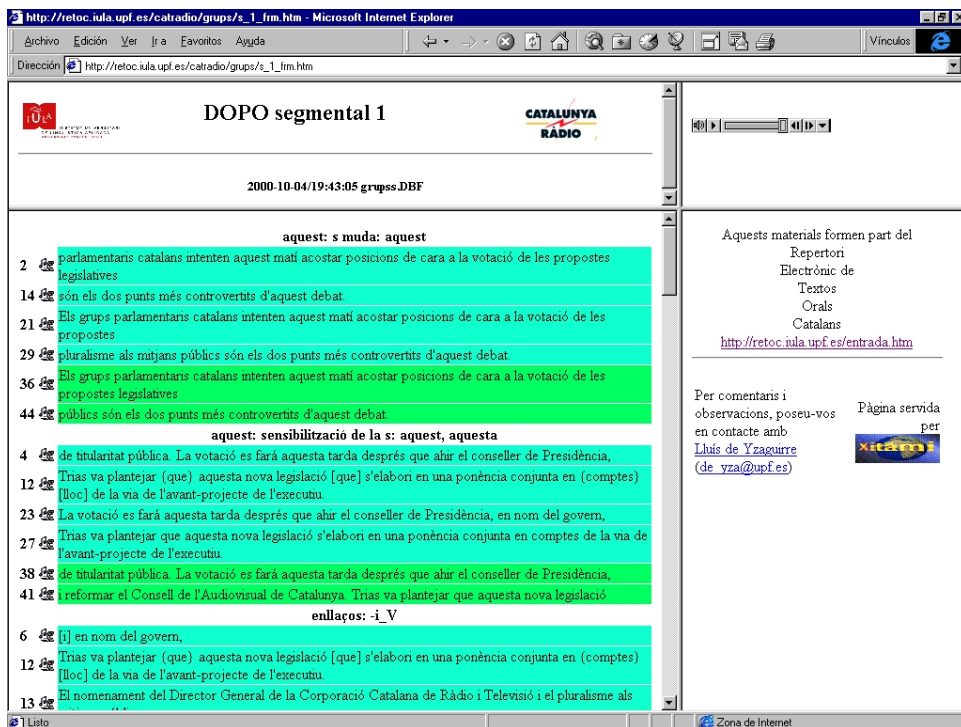
The two projects briefly described above are carried on bearing in mind one general goal: the construction of the Genome Knowledge Base, whose architecture is shown in the next figure.



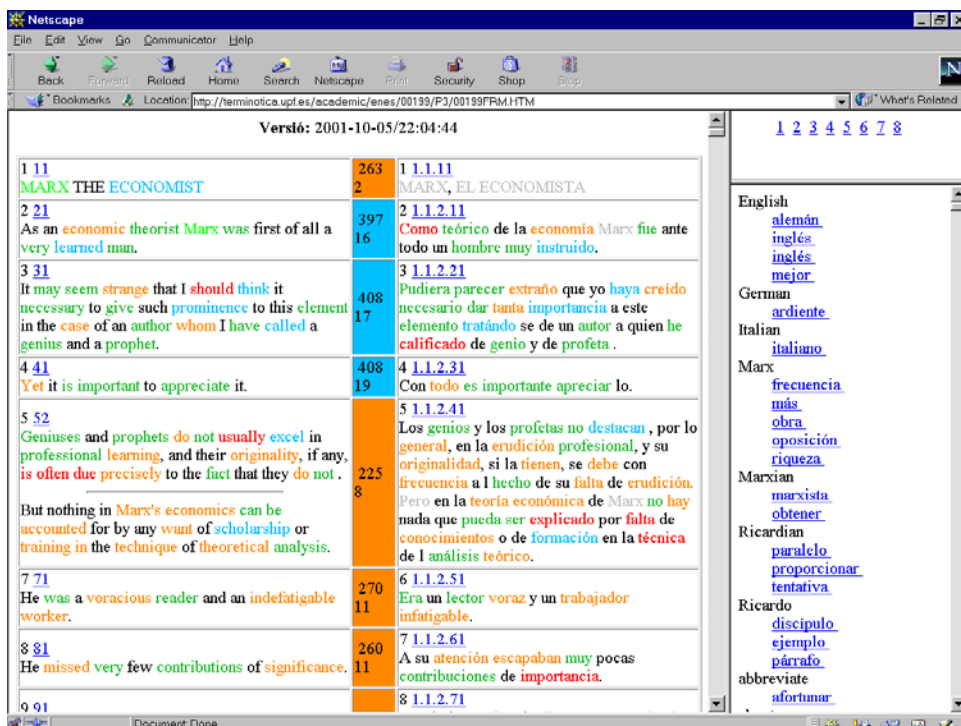
This figure shows the tight relation between the terminological database, the specialised knowledge units, the concepts and the documents related to the Human Genome domain which will constitute the core of the project. The ontology, directly related to concepts, will be used in order to classify and structure specialised knowledge drawn from the corpus. In order to ease such task, the documents included in the corpus are previously morphologically and syntactically tagged.

The terms registered in the terminological database will be linked to both the ontology and the documents from where they have been retrieved. The resulting set of knowledge will be used for different tasks, such as document indexation and summarisation, machine translation support, etc.

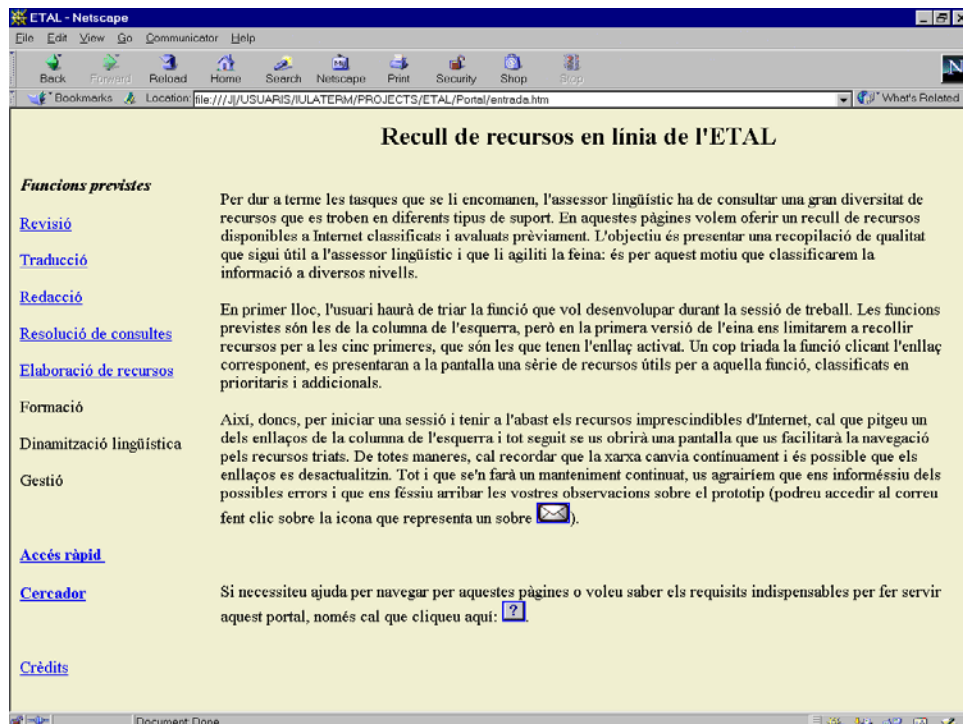
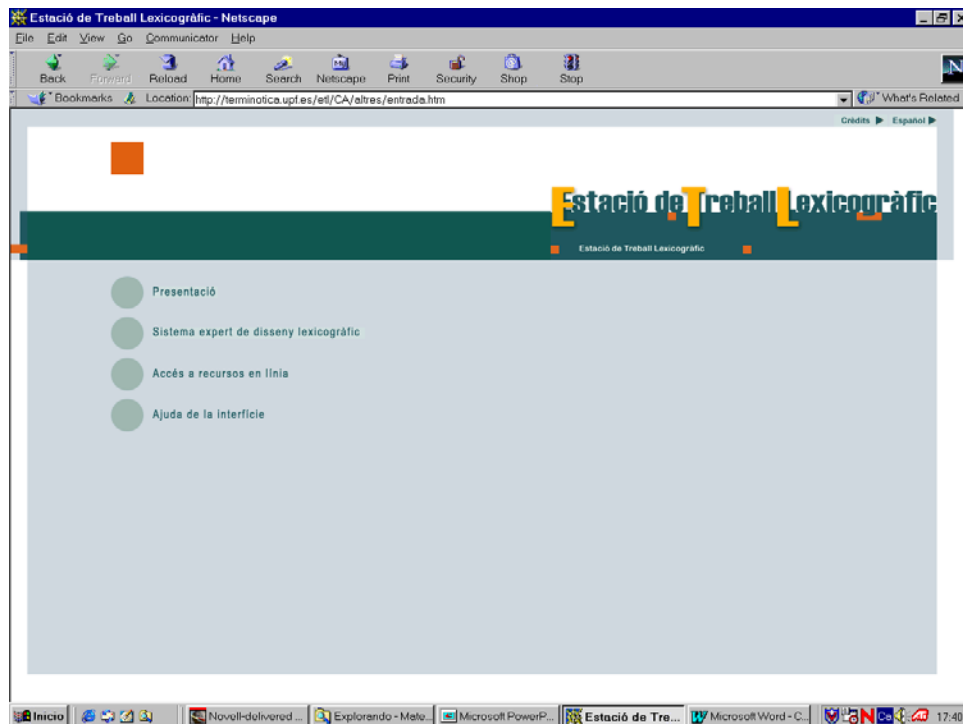
The DOPO project (Diagnòstic Ortològic per Ordinador / Computer Orthological Diagnoses) is a system for the detection of orthological indexes in a computer-assisted way. It is based on the synchronisation of voice and text files. The system is been jointly developed by the Universitat Pompeu Fabra and Catalunya Ràdio. The latter will be its main user when the system becomes efficient enough. The following image shows a brief representation of its working methodology.



The Alignment Tool for Parallel Corpus is an application for the alignment of texts at the sentence level. Precision reaches more than 95% and the alignment results are also given at the lexical level. The following image indicates which are the sentence and the words pairs resulting from the alignment process.



Finally, the Institute for Applied Linguistics has also worked in the construction of two workstations. Many efforts have been devoted to the construction of a lexicographical workstation and the prototype of a linguistic planning workstation. Both of these applications contain all resources necessary for a lexicographer and a linguistic planner aiming to work with all integrated tools in a particular workstation. See the following two images to illustrate the above mentioned.



L'Institut d'Estudis Catalans (Institute for Catalans Studies)

The Institut d'Estudis Catalans has built a web site called Portal de dades lingüístiques (Linguistic Data Website) where it is possible to reach all information concerning the dictionaries and the corpus they have available. Information can be obtained using the query page and results can be retrieved and saved in order to be reused. Next image shows the initial web page of the Portal de Dades Lingüístiques (<http://pdl.iec.es>).

