

## IULA's LSP Multilingual Corpus: compilation and processing T. Badia, M. Pujol, A. Tuells, J. Vivaldi, L. de Yzaguirre, T. Cabré\*

Universitat Pompeu Fabra  
Institut Universitari de Lingüística Aplicada  
Rambla Santa Mònica 32, 08002 Barcelona, Spain

### 1. Introduction

The Institut Universitari de Lingüística Aplicada (Institute for Applied Linguistics), known as IULA, is a public institution that belongs to the Universitat Pompeu Fabra (UPF) in Barcelona and is closely connected to the Translators and Interpreters Faculty of UPF. IULA is devoted to both postgraduate teaching and research in applied linguistics, covering basically the following fields: lexicology, lexicography and terminology; computational linguistics and linguistic engineering; and language variation. Although within each of these areas different, specific projects are being carried out, there is a single unifying project to which every researcher in IULA contributes to some extent, namely the creation of computer based resources for linguistic research in the fields covered by IULA's main research areas.

This common project consists in the creation of a language for specific purposes (LSP), multilingual corpus<sup>1</sup>. It is primarily intended to be useful for researchers in the areas of applied linguistics that are under focus in IULA and for the teaching activities in IULA and in the translators' faculty. But we also foresee collaboration with other research centres, both public and private: IULA is already heavily involved in a joint project on computational resources for Catalan with other research institutions in Catalunya.<sup>2</sup>

The creation of the LSP, multilingual corpus implies three basic steps: the compilation of the corpus, its processing, and its exploration. In section 1 we discuss the main decisions taken on the compilation of the corpus and the way it is internally organised. In section 2, we present the current state of development and integration of corpus processing tools. The integration of processing tools follows naturally the compilation of the corpus itself and we develop them only in case we have not found already existing tools which are available and easily adapted and integrated to our environment. The various subsections in section 2 cover the different processing levels that we foresee for Catalan, Spanish and English: structural tagging, text handling, morphological analysis, morphological disambiguation and syntactic analysis.

---

\* e-mail: {tbadia, mpujol, tuells, vivaldi, de\_yza,cabre}@upf.es

<sup>1</sup> The project "Llenguatges d'especialitat. Corpus Multilingüe" is supported by the Catalan research agency CIRIT (project number: CS93-4.009). The main researcher is M. Teresa Cabré.

<sup>2</sup> The "Centre de recerca en enginyeria lingüística" (CREL) has been created by the Catalan research agency in 1996 and is composed by the main linguistic research units in Catalunya.

Section 3 is devoted to explain the current exploration tool, which has been implemented according to the internal organisation of the corpus. Since this is a project under development, we have included a last section 4 in which we present the work that is planned for the near future.

### 2. Corpus Compilation<sup>3</sup>

The main objectives of the LSP, multilingual corpus is to support the research and teaching activities of the members of IULA. To this end, the corpus and the associated processing tools should provide the computational basis for a number of researches in both monolingual and multi-lingual frameworks, such as concordances based on morphosyntactic information, term detection, text alignment, syntactic analysis... According to these objectives, a number of design criteria have been set up and a specific internal organisation has been adopted. To these two questions the following two subsections are devoted respectively.

#### 2.1 Design criteria

According to the above mentioned objectives a number of design criteria have been adopted:

- Selected texts use a specialised written language. Given the thematic areas that are covered in the translators' faculty, we have decided to select texts from five such areas: law, economics, medicine, computer science and environment.
- The corpus is multilingual. Whenever possible it is intended to contain the same document in different languages. The languages involved are those that are intensively studied at the translators' faculty: Catalan, Spanish, English, French and German. Only contemporary language is included.
- The corpus organisation has to be flexible enough to be easily adapted to the different needs of IULA as well as to meet our potential users' requirements.
- In order to facilitate text interchange between different working environments we have adopted the standard ISO 8879 (SGML). Particularly, the project follows the recommendations issued by the Corpus Encoding Standard (CES) of the EAGLES initiative.
- The corpus has to be as representative as possible. To reach such an objective we decided to have a classification of each domain from two viewpoints: the own of the domain and a simple text typology. This

---

<sup>3</sup> For a detailed description see (Bach C et al. 1997) and IULA's LSP Corpus home page <http://www.iula.upf.es/corpus/corpus.htm>

task is the responsibility of experts of each domain that collaborate with our project. Such experts also provide a selection of documents relevant to the domain in relation to its taxonomies.

- We have extended the concept of representativity to each document. For such a reason we have decided to select an average of 10 samples of each document. They should be non consecutive and the number of words of each sample should be between 3 and 5 Kwords.

At present the corpus includes about 500 documents totalling more 11 million words. 90% of this material is in Catalan and Spanish; texts in English are increasingly being introduced in the corpus.

## 2.2 Corpus organisation

In devising the corpus internal organisation we have taken into account the main objectives of the corpus (as summarised in the previous subsection). In particular, we have adopted some internal organisation criteria which are expected to favour an optimal exploitation of the corpus. The basic decision has been to use the SGML standard; this decision has already forced an internal organisation of the documents. Other, complementary, decisions help in keeping the idea of the corpus as an organised collection of text. To the discussion of these aspects is devoted this subsection.

The adoption of the SGML standard allows us to have a common internal format for all documents; and makes easier for IULA to share this resource with other similar organisations.

According to this standard all the documents must belong to a type. Document type is formally defined by means of a text file (or DTD) where the designer declares the internal organization of the documents of such type. In other words it defines how the text elements may be legally combined. As with linguistic grammars, a parser must check the correctness of the document from the structural point of view. The CES has developed a DTD that should be used in the corpus compiling process. IULA has basically adopted this specification with some minor modifications in order to facilitate the mark-up with the available, limited resources.<sup>4</sup>

Every document in our corpus is divided in three main parts: the document initialisation, the header and the text itself. The document initialisation part declares the DTD to be used in the document and some other resources called entities (which mainly are auxiliary files). The header contains all the necessary information to identify the document: bibliographic information (title, author, publisher, date, ISBN, ...) and the corpus internal information (internal text classification, text typology, language information with indication if it is a translation, size in words and Kbytes, pointers to the samples, type of samples, ...).

The third part is the text itself, which is not inserted directly but included by means of pointers that

refer to the file containing the text. Such pointers have been defined

All this information is organised in three different files related to each other. The first one just includes the skeleton of the document, the second file defines the pointers to the samples of each document and the last one includes the header and the logical insertion of the samples.

It should be noted that the internal organisation of the documents mentioned so far is independent from the language of the texts. All the documents have a common structure: the same information is always placed under the same SGML element. This organisation makes it easy to locate any piece of data.

Thus the first processes that a document suffers after it has been chosen refer to its form and to the identification of its formal (non linguistic) properties. In this way it is ready for the linguistically oriented processing described in the following section.

## 3. Corpus Processing

### 3.1 Introduction

In a corpus like the one we are describing, the documents to be processed contain free text. As is well known, the processing of free text has to cope with a number of difficulties, which do not only come from the intrinsic difficulties in processing natural language. Very often they come from misspelling or unknown words, a myriad of punctuation signs, numbers, labels, dates in various formats, multi-word units, proper nouns, foreign words, etc. Some of these items have specific conventions for every language, like decimal signs or dates. All of these items have to be taken into account if the target is to process actual text for producing material ready for linguistic research.

The basic strategy to obtain free text analysed in a reasonable way is to divide the whole process in different modules, each one with a specific task. This is even more essential if the whole process has to cope with texts of different languages, given that the results of the processing have to be comparable (in terms of both the linguistic information obtained and of the formal means to represent it).

Figure 1 shows the full process to which every text is submitted before it is included in the textual database for exploration by the end users. The basic pipeline procedure includes the following main tasks :

- document selection, search and recovering
- structural tagging
- text handling
- morphological analysis
- morphological disambiguation
- syntactic analysis

These tasks are organised in independent modules, in a such a way that any one of them can be easily modified, enlarged or replaced without affecting the whole process.

---

<sup>4</sup> For a detailed description see (Vivaldi J. et al. 1996)

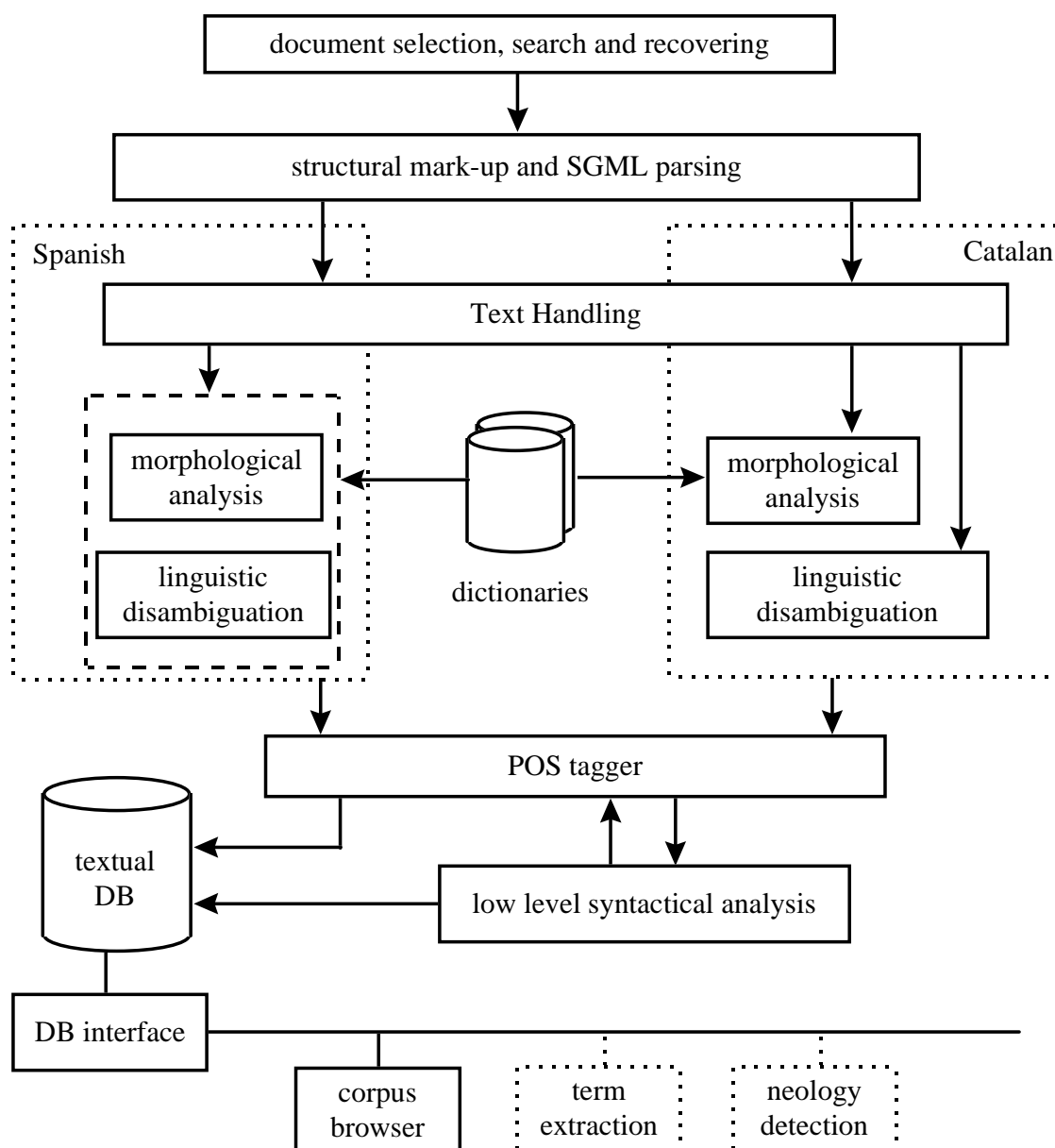


Figure 1. Processing flow of the corpus documents

Every task requires one or more specific pieces of software. Since IULA is not in the position to start from scratch the programming corresponding to every task, we use any available tool that can be easily adapted to our working environment (taking into account its multilingual character). We concentrate on the developing of those parts for which no specific tool exists or that are very specific to the internal organisation of our corpus (such as the morphological analyser for Catalan or the corpus exploration tool). Whenever possible we adapt already existing tools (like the POS tagger). It has no sense to try to build a new tool for processing English texts, so we are in the process of adopting the well known Constraint Grammar for such a purpose and integrating it in our environment. The same policy will apply for French and German.

In the following subsections we discuss each of the stages of the processing and we briefly explain the

main characteristics of the tools that we currently use. These refer basically to Catalan and Spanish, since they are the languages that are being dealt with almost completely.<sup>5</sup>

### 3.2 Structural tagging

The input text is tagged according to the SGML standard. However we do not use the whole expressivity of SGML for two very specific reasons: the main objective of the mark-up is linguistic research (and not other areas of document processing) and the resources available for this task are limited. Therefore we limit the structural information to: main divisions, paragraph and sentence

<sup>5</sup> Note that all the tools described in this paper can naturally be used for other purposes than the one explained here. In what follows we do not detail these uses at all but refer the reader to the corresponding descriptions by the original authors.

identification, lists, notes, rendering information and sequences in a foreign language.

To perform this tagging we take advantage of any kind of information associated with the input text: surface distribution of the text (legal documents), typographical codes (text coming from the publisher), HTML tags (texts captured from Internet), etc. Very often specific routines are required to deal with a certain text; as a matter of fact, any resource is taken into consideration in order to avoid manual tagging. Finally, the rest of the tagging is performed with a set of macros that help the coder to add the structural information.

The automatisisation degree strongly depends on the text origin (internet, publisher house, author, scanner, etc.), since the problems encountered are quite different. In practice, the text never is added automatically to the corpus, each type of text has its own peculiarities, that need to be solved, or alleviated, by using specific procedures.

The paragraph and sentence information is automatically added at the end of the structural tagging. This routine has to cope mainly with the dot ambiguity (decimal separator, abbreviations, person names initials, acronyms, etc.) and the accuracy rate that we achieve is not 100%. Some of the remaining mistakes are detected later by the SGML parser and others persist during the linguistic processing stage.

The text obtained at the end of this stage is parsed against an SGML parser<sup>6</sup> in order to guarantee that it is free of SGML syntactic errors.

### 3.3 Text handling

The analysis of natural language is commonly accepted as a difficult task but it becomes much more difficult when the text to be processed is actual text not just laboratory adjusted sentences. Free texts contain elements that can be considered as trivial for human processing but that create real difficulties when the text has to be processed by a computer.

Punctuation signs, dates, locutions and proper nouns, are just a few examples of units that dramatically increase the difficulty of the processing. An early detection of these phenomena will help to lighten the task of later processes. For example, the early detection of proper nouns and their mark-up as a single lexical unit will avoid the need to cope with the problem of possible unknown words in morphological analysis and with the consequences (not easily foreseen) that this may have in the following disambiguation stage. Also the syntactic analyser can take advantage of such a treatment, thus avoiding the generation of bizarre phrases.

The text handling stage has the crucial mission to tag any linguistic unit that can be detected by surface analysis of the text: dates, numbers, proper nouns, abbreviations, labels, as well as to manage the punctuation signs found in the text. In a sense this stage can be

considered as the continuation of the structural mark-up since its basic function is to facilitate further processing. In our working environment it is essential that the text handler takes care of the differences in the use of some of these items in the languages involved in our corpus.

The strings that will be processed by the text handler, and some examples, are the following:<sup>7</sup>

- proper nouns: Cambra de Comerç (**ca**) [Chamber of Commerce], Ministerio de Educación (**sp**) (Ministry of Education), OEA (**ca/sp**) [American States Organisation ], ...
- Dates: 25 de mayo de 1810 (**sp**) [25 of May of 1810], 25/5/1810, 25-V-1810, May 25<sup>th</sup> 1810, ...
- Locutions: a conseqüència de (**ca**) [as a consequence of], en definitiva (**ca/sp**) [thus], ...
- Cardinals: 3,14, 3.14, 3'14, XII, twelve, dotze (**ca**), ...
- Ordinals: 1r, 1er, 1.er, 1<sup>o</sup>, 1<sup>a</sup>, 1ro, ...
- Measure units: m/s, mt/seg, ...
- Labels: a), a., 1), ...
- Abbreviations: adj., art., v. art., ...
- Punctuation signs: , ; : - [ ] ( ) « » " ' ^ ° ...

Many of the items just mentioned have a different behaviour in each language of the Corpus. The paradigmatic example is proper nouns. The algorithm for detecting them is pretty simple but effective. We consider a proper noun to be any sequence of words starting with a capitalised letter plus some joining item like *Cambra de Comerç* (**ca**) [Chamber of Commerce] where *Cambra* and *Comerç* are the capitalised words and *de* (**ca**) [of] is the joining item. Joining items are a set of predefined lexical units specific for each language. Note that some of them have a special behaviour; the letter *i* (**ca**) [and], for example, can be a joining item when processing Spanish text ([**sp**]... *por decisión de la* [**ca**] *Conselleria de Educació i Ciència* [... as was decided by the Department of Education and Science] but not in Catalan text (...*a continuació, Espanya i França van signar l'acord* [...afterwards, Spain and France signed the agreement].

Punctuation signs are also difficult to deal with because of their ambiguity. Consider a small sample of clear examples. The slash ("/") may be found inside of a unit of measure (*m/s*) or be a conjunction (*l'objectiu és obrir/desregularitzar el mercat* (**ca**) [the goal is to open/deregulate the market]; and the closing bracket can belong to a label or be an independent punctuation sign; for example, *c) Introduir la tarja magnètica (vegeu les indicacions del capítol X)* (**ca**) [c] Introduce the magnetic card (find the instructions in chapter X)].

In some cases in order to speed up the process, there is an additional module that contains the most very common words preanalysed. At the same time some of the ambiguities are eliminated based in their very low frequency in our corpus (e.g. interjections) or orthographically complex words are assigned their tags directly (*dóna'ls-ho* (**ca**) [give it to them].

<sup>6</sup> We are using the SGML parser "nsgmls", developed by James Clark. For more information, see <http://www.jclark.com>.

<sup>7</sup> Here and in the rest of the paper, we use the following language codes: **ca** (Catalan), **sp** (Spanish).

Thus the text handler tries to find an appropriate solution to most of these noisy problems. It is built as a basic module common to all languages and with some additional modules which cope with the specificities of each language.

After the application of the text handler the document is ready for the specific, language dependent processing performed by the following tools: morphological analysis, morphological disambiguation and syntactic analysis.

### 3.4 Morphological analysis

The tools described in the previous sections have all been developed at IULA: they perform quite low level tasks which are very much dependent on the design criteria of the corpus, and on the basic decisions taken concerning its internal organisation. In this and the following sections, we describe the different linguistic modules that operate, or are under construction/adaptation. As already mentioned above, we only develop the tools that are not available in a form that is suitable to our purposes. Whenever a tool can be adopted which performs the task foreseen by our overall structure, we try and integrate it to the processing flow (see figure 2). Thus in the following we describe both the tools that we have developed and the ones that we have integrated (or are in the process of integration).

In all morphological modules we guarantee that the dictionary management is flexible enough to allow new items to be introduced easily. This, of course, occurs quite often, since we are processing LSP texts which contain many lexical items that do not appear in general lexicons (such as the ones originally included in the tools we are using).

The morphological analyzers provide, in a pretty standard way, all the possible analysis for each word form. In the following we briefly describe the main characteristics of this tool for Catalan and Spanish. English morphological analysis is going to be performed by the two-level morphological grammar included in the ENGCG package: the ENGTWOL tool.<sup>8</sup>

Catalan morphological analysis is performed by CATMORF, a module developed at IULA, which is the first wide-coverage two-level morphological analyser for Catalan. CATMORF models morphotactics in a (DCG-like) word grammar and morphographemics in SEGMORF, a two-level morphological formalism written at our institute, which is an extension of the Alep morphographemic segmentation formalism.<sup>9</sup>

As for the construction of the lexicon of CATMORF, the items in our lexicon contain information

on the word form and lemma; the inflection paradigm of verbs, nouns and adjectives (needed for both the word grammar and the two-level rules components); and the blocking of rules by several classes of stems. All this information has been obtained semi-automatically from a machine-readable dictionary (a conventional dictionary available in electronic form): the IEC dictionary (DIEC, 1995), which is a recent normative dictionary for Catalan. More recently, entries from the DLC95 dictionary (DLC, 1995) which were not present in IEC dictionary have been added to the system. Our Catalan lexicon contains more than 70000 entries.

The morphological analysis of Spanish is performed by the module PALIC. It assigns to each word form one or several pairs of lemma and tag. We use a word form dictionary which was originally derived from an electronic version of a human-user general purpose dictionary (DALE 1995).

This original dictionary has recently been extended with the addition of new lemmas (provided by the same dictionary publisher) from which all word forms have been derived. The resulting lexicon contains the word forms corresponding to about 90000 lemmas, the 70 % of which have been imported automatically.

### 3.5 Morphosyntactic disambiguation

After morphological analysis has been performed our system combines both linguistically based knowledge and a statistical tagger to obtain a text which is accurately disambiguated. The main goal is to select the correct morphosyntactic tag for each ambiguous word in context from a set of possible tags. This is the first step to facilitate high level analysis, such as recognising basic structures and other useful patterns. In the following sections we describe the two approaches in the order in which the two modules are applied.

The ambiguity rate for Catalan and Spanish morphologically tagged texts is about 1.7 tags per word. To this input we apply a linguistically based disambiguator. Its operation relies on the fact that quite often the linguistic context of an ambiguous word form allows to disambiguate it with absolute certainty. This occurs, for example, when a word presents an ambiguity between determiner and pronoun (as, e.g., both Catalan and Spanish "el"): in these cases the pronoun alternative can be safely excluded if "el" is followed by an unambiguous noun (or proper name, date, or number).

In a very standard way, the rules that perform this kind of disambiguation apply when a particular lemma or tag is encountered for which a rule has been written and when the context in which it appears matches the context specified in the rule. Additionally some of these rules can be deactivated at will according to some external criteria such as frequency, dialectal variation, ...

At the moment, the linguistic disambiguation is performed with independent formalisms for Spanish and Catalan, and their results are not comparable: Spanish disambiguator deals with more cases than the Catalan one

---

<sup>8</sup> The ENGCG package is thoroughly described in (Karlsson K. et al. 1995); you can also see the URL address: <http://www.lingsoft.fi/doc/engcg/>.

<sup>9</sup> The Alep system is described in (Alshawi H. et al. 1991) A fuller description of IULA's SEGMORF and CATMORF can be seen in Badia; Tuells (1997) and Badia; Egea; Tuells (1997:1998).

and thus reduces to a greater degree the ambiguity rate to be dealt with by the statistical tagger.

Since at this point of the process grammatical tokens of the text can still have more than one morphological category, the next step is to fully disambiguate the remaining ambiguities.

For such a purpose we integrated the tagger developed at ISSCO<sup>10</sup> under the MULTEXT project. It has been necessary to add some additional processing in order to keep the SGML tags because they were not foreseen in the design of the original tool. The main characteristic of this tool is its flexibility as explained at (Armstrong et al. 96). We successfully apply it for both Catalan and Spanish.

This POS tagger is based on a Hidden Markov Model, a well known probabilistic technique used in a variety of tasks on natural language processing. The process is accomplished in two steps: training and tagging. During the first one, the language model is built that will be applied in the tagging phase.

In order to build the language model we compiled a training corpus of about 90 K words per language, 40 % of which was fully disambiguated by hand. Additionally we prepared a set of biasing rules to improve the quality of the disambiguation, favouring some tag transitions against others. Of course, these rules are language dependant.

The full tagset developed at IULA includes about 350 tags (Morel J. *et al.* 1997), but the actual tagset used for the tagger is reduced to 23 tags. This reduction is made by mapping any complex category to the base category with some exceptions. For example, the tag for masculine singular common nouns is N5-MS but it is mapped to just N for tagging and expanded again after disambiguation. With some categories the mapping is a bit more complex; e.g., the whole verbal paradigm is reduced to 4 tags V (personal form), VC (participle), VG (gerund) and VI (infinitival). Using a larger tagset makes it difficult to compile the training corpus, because it should be much bigger in order to find enough examples of each ambiguity class; consequently the training time increases considerably.

According to our preliminary results, the error rate obtained is between 4 and 6%. Errors concentrate on ambiguity classes that require to consider a wider context (like a word following a punctuation sign or the ambiguity of some pronouns), or in which the variety of the remaining errors is so big that they become intractable at this level (adjective vs. noun).

A previous step in this stage consists in solving some critical points deriving from the fact that we are processing actual text. The input text may contain unknown words due to a variety of reasons (mistakes from the original, misrecognition by the OCR, not inclusion in the dictionary, being units of measure, ...). In these cases the only information provided by the

morphological analyser is that the word is unknown. Since this situation is unacceptable for the tagger, a process is triggered that makes a guess about the category of any word without morphological analysis.

### 3.6 Syntactic analysis

The step after morphological disambiguation is syntactic tagging. Very standardly we assume that syntactic processing of unrestricted texts is only possible if a restricted approach is adopted, that is to say, if not a full, deep syntactic analysis is intended. There are two aspects which make the only possible syntactic analysis restricted. Firstly the analyser, as in every previous stage, has to be able to deal with any kind of input, including incorrect text or non disambiguatable text. And secondly, a syntactic parser can only be built in a reasonable amount of time and effort for unrestricted text if it does not need very elaborate lexicons.

A syntactic analyser for corpora makes only sense if it can deal with any kind of input. It is a grammar that has to be used with the texts that are collected by independent criteria, without taking into account the linguistic phenomena that occur in them. It cannot therefore be a grammar built in the traditional way: by determining a linguistic coverage and setting up a test text according to the previously fixed coverage. In a corpus driven analyser the linguistic phenomena to be covered cannot be chosen: they are given.

On the other hand, ordinary syntactic grammars tend to use lexicons which are very rich in information. This cannot be the case either for a corpus driven parser. It has to be able to deal with all types of texts, including all possible sorts of words (even words that are not in any dictionary). It is not possible to wait to have a full lexicon for parsing texts: it is never going to be complete, new words are always going to occur. In addition the task of constructing a complete dictionary is a very effort consuming one, which is not available to most research centres (certainly not to ours).

With these two restrictions in mind we have adopted the Constraint Grammar (CG) formalism to build the Catalan and Spanish syntactic parsers. After having evaluated other possible formalisms, we have chosen the CG formalism because it is going to integrate very easily into our working environment and allows an easy interaction with the lexicon.

## 4. Corpus Exploration Tool

The main purpose in building a corpus is to observe the behaviour of the lexical units included in it. The whole tagging process is oriented towards the increase of the information associated to lexical units (lemma calculation, morphological disambiguation, syntactic analysis, ...), so that such information can be afterwards selectively recovered for linguistic research. The observation of this linguistic information can range from the internal parts of a word to its combinations in forming phrases, sentences or even paragraphs. The tool/s devoted to such corpus

<sup>10</sup>For more details please visit the following URL: <http://issco-www.unige.ch/tools>

exploration have a crucial role in the profit obtained in compiling the corpus.

In order to satisfy the above mentioned goals, and taking into consideration the needs of our researchers, we have adopted some design criteria about our browser. It should:

- have an user friendly interface
- be flexible enough to be useful to as many research areas as possible
- take profit from the SGML mark-up
- be multilingual
- keep linguistic knowledge separate from the process of obtaining it
- be accessible to as many users as possible
- be easy to expand to specific exploration software
- be reasonably fast
- be able to run on any platform of our working environment (mainly PCs and UNIX workstations)

The basic units considered by our tool are those resulting from the output of the text handler. They may be single (words, labels, numbers, ...), multiple (dates, proper nouns, locutions, ... ) or grammatical words (contractions, verbal constructions, ...). Each unit will have associated three basic pieces of information: form, lemma and morphological tag.<sup>11</sup> The user is able to do searches based on these three types of information freely combined. The following example shows a simple search:

Word form		
Lemma	asma	
Tag		JQ--66

In this case the browser is looking for a sequence of two grammatical units, the first one has the string *asma* (**ca**) [asthma] as lemma (the form and the morphosyntactic tag do not matter here) and is followed by a unit that must be qualificative adjective. Sequences like *asma bronquial* (**ca**) [bronchitic asthma], *asma al·lèrgica* (**ca**) [allergic asthma] and *asma crònica* (**ca**) [chronic asthma] are going to satisfy this pattern.

It is also possible to introduce an optional unit, so that in the resulting sequence one or more components may or not be present in the sequence.

Word form				
Lemma		de		
Tag	N5-66	P	N5-66	(JQ--66)1

In this case the sequence is noun-preposition-noun (where the preposition is forced to be *de* (**es**) [of]) and the second noun may be modified by an adjective. Sequences like: *factor de necrosi tumoral* (**ca**) [tumor necrosis factor], and *cas d'intolerancia digestiva* (**ca**) [case of digestive intolerance] are going to satisfy this pattern.

The expressivity of the searches is increased by allowing three other means: the multiplicity of the

<sup>11</sup> When the syntactic analysis is fully in operation the syntactic information is going to be also associated to each unit.

category (n occurrences of a certain category in the sequence), the negation of a category, lemma and/or form (a unit must not be a certain category) and a limited form of regular expressions in the word form (forms starting or ending with certain letters). So we can introduce complex queries like the following :

Word form	(pre* post*)		
Lemma			
Tag	V?????	^(Z)10	P

In this hypothetical query, the browser looks for a sequence of any verb form whose word form starts with the prefixes *pre* or *post*, followed by a maximum of ten units that must not be punctuation signs and ending with any preposition.

All queries must be done over some documents, which can be selected either by listing them individually or by defining a filter. Such a filter is based on the data contained in the header (see section 2.2) (e.g. area, text typology, language status, ...), and/or on numerical figures (e.g. number of documents, number of words, ...).

The tool also allows a simple frequency calculation based on the word form, word lemma or morphosyntactic tag. There is also the possibility to calculate such frequencies relative to a unit chosen by the user.

The last operation of the browsing process is the presentation of the results. To obtain them, the user can choose from a variety of options regarding the context, the level of information presented, and the possibility to sort the results. Finally, the user may save the result in an ASCII or RTF file.

An important aspect to be considered is that different languages usually have different tagsets. This means that the differences in the tagset for each language has to be taken into consideration. This was accounted for by allowing the definition of the tagset in an external text file.

Our browser has been developed using a standard textual database that accepts any document with a valid SGML mark-up. In particular we used the Dynatext Electronic Publishing System<sup>12</sup> together with its SIT library (Systems Integrator Toolkit), that allows an easy access to the information saved in DynaText books. In such a way we create a data retrieval engine for a fully custom SGML browser.

In order to enable the graphical interface to run on our working environment we used WxWindows class library<sup>13</sup>, that allows to compile graphical interfaces on a range of different platforms.

<sup>12</sup> We developed our browser under the DynaText Grant Program. For more details please visit the following URL: <http://www.inso.com>

<sup>13</sup> This is a public domain software developed at the University of Edimburg by John Smart. For more details please visit the following URL: <http://www.aiai.ed.ac.uk/~jacs/wxwin.html>

Although it is still under development this browser is currently used at IULA by many students and teachers, for their research purposes and for testing the efficiency of the whole process.

## 5. Further Work

Since the work reported here is work in process our first aim in the immediate future is to enhance the system, both in the performance of every module and of the system as a whole.

In addition, there are various levels at which the work reported here is foreseen to be continued. Firstly, there are some modules that we have described that have not been finished yet or have not been fully integrated in our environment: for example, the syntactic analysis for Catalan and Spanish and the browser have not been finished yet. Also we plan to integrate a general language corpora whenever possible to allow contrastive studies.

Secondly, the integration of the three other languages foreseen in our corpus (English, French and German) is going to imply that some tools for them are also integrated in the current working environment.

Thirdly, there are some levels of processing that can benefit from the work being currently done. This is the case, e.g., of morphological disambiguation, which we will probably be able to improve when a full Constraint Grammar for morphological disambiguation is available, that is to say, when we will be able to interact the CG disambiguation with the stochastic tagger currently used. Another example is the further development of the browser adding new features like: to include syntactic information in the queries, to allow queries that takes as input the result of previous queries, remote access to our corpora, batch searches,...

And finally, there are the aspects that we are going to be able to address when the major modules currently under construction have been finished. In this sense we plan to address questions like information extraction (term extraction, neology detection...) or comparison of the tagging of parallel texts. When addressing these aspects we are going to exploit, and take advantage of, the work devoted to the setting up of the processing environment just described.

## 6. Conclusion

In this paper we have described the processing environment of the LSP, multilingual corpus developed at IULA. After a brief mention to the criteria used in designing the corpus, we described the way in which the documents selected are coded and maintained in the corpus. In the main section of the paper we have described the processing environment. In its current state, the system covers fully the text handling, the morphological analysis and disambiguation for Catalan and Spanish. The low-level syntactic analyser for both these languages is under construction, and the English processing as provided by the EngCG tool is being integrated into the system. In addition we have described

the corpus exploration tool by means of which we intend our students and researchers to access the information coded in the corpus. Finally in a brief section we list the work that we foresee in the immediate future.

## REFERENCES

- Alshawi, H. *et al.* (1991): *EUROTRA ET6/1: Rule Formalism and Virtual Machine Design Study. Final Report*. Commission of the European Communities.
- Armstrong, S.; Robert G.; Bouillon P.: Building a language model for POS tagging. (<http://isssco-www.unige.ch/tools>).
- Bach, C.; Saurí, R.; Vivaldi, J.; Cabré, T. (1997): *El Corpus de l'IULA: descripció*, Papers de l'IULA, Sèrie Informes n 17, Institut Universitari de Lingüística Aplicada, Barcelona, Universitat Pompeu Fabra.
- Badia, T.; Egea, À.; Tuells, A. (1997): "CATMORF: Multi two-level steps for Catalan morphology", in *Description of Systems Fifth Conference on Applied Natural Language Processing*, p. 25-26, Washington.
- Badia, T.; Tuells, A. (1997): "SEGMORF, An extension of the Alep morphographemic segmentation formalism", in *Proc. Third ALEP User Group Workshop*, p 1-8, IAI, Saarbrücken, Universität des Saarlandes.
- Badia, T.; Egea, À.; Tuells, A. (1998): "CATMORF: un analizador morfológico de dos niveles de cobertura amplia para el catalán", in *Procesamiento del Lenguaje Natural*, n 22, Barcelona, Sociedad Española para el Procesamiento del Lenguaje Natural.
- DALE (1995), *Diccionario Actual de la Lengua Española*, Bibliograf, Barcelona.
- DIEC (1995), *Diccionari de la Llengua Catalana*, Institut d'Estudis Catalans, Barcelona.
- DLC (1995), *Diccionari de la Llengua Catalana*, Enciclopèdia Catalana, Barcelona.
- F. Karlsson; A. Voutilainen; J. Heikkilä; A. Anttila (ed.) (1995): *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*, Berlin-New York, Mouton de Gruyter.
- Morel, J.; Torner, S.; Vivaldi, J.; De Yzaguirre, L.; Cabré, T (1997): *El Corpus de l'IULA: etiquetaris*, Papers de l'IULA, Sèrie Informes n 18, Institut Universitari de Lingüística Aplicada, Barcelona, Universitat Pompeu Fabra.
- Vivaldi, J.; De Yzaguirre, L.; Solé, X.; Cabré, T. (1996): *Marcatge Estructural i morfosintàctic del Corpus Tècnic amb l'estàndard SGML*, Papers de l'IULA, Sèrie Informes n 1, Institut Universitari de Lingüística Aplicada, Barcelona, Universitat Pompeu Fabra.
- Voutilanen, A. (1995): "Morphological disambiguation", In Karlsson et al. (Karlsson F. *et al.* 1995), p. 165-283.