CORPUS AND EXPLOITATION TOOL: IULACT AND *BWANANET*

Jorge Vivaldi Palatresi
Universitat Pompeu Fabra

ABSTRACT:
Over the last decades corpus linguistics methods have found ever increasing use in almost all linguistics related studies, mainly due to their usefulness to get and validate results.

The IULACT is a project from the Applied Linguistics Institute intended to compile a corpus of LSP texts. This corpus includes documents from a variety of domains, registers and languages. In contrast to other resources captured directly from the web, the texts of the IULACT have been selected in a supervised way, transformed to a clean SGML format and processed with a set of specific tools. The overall goal of the project is to provide not only an infrastructure to study LSP but also a resource for computational linguistic research.

This article describes the IULACT corpus, its architecture as well its processing tools. It starts from its design criteria and analyzes all the necessary processing stages and related software tools. Finally *bwanaNet*, the corpus browser tool, is described.

KEYWORDS:
corpus, corpus linguistics, corpus processing tools, corpus browser

I. INTRODUCTION

Since the 70's the study of language through corpus linguistics techniques has become more and more important. The result is that, today, corpora are considered a default resource for almost any research in linguistics; in the sense that none research may claim credibility without being verified through actual language data. Corpora are important linguistic resources that allow obtaining a great amount of knowledge about language behaviour in real use. In addition, corpora have allowed to go forward in areas where these resources are used (lexicography, language teaching, etc.) as well as issues related to the creation and exploitation of such resources (natural language processing —NLP— from a myriad of perspectives, large textual databases, statistical analysis, corpus based text mining tools, etc.).[i]

The goal of this paper is to introduce the IULA's[ii] language for specific purposes (LSP) Corpora (IULACT) as well as *bwanaNet*, its corresponding exploitation tool. We will describe the steps involved in designing, building, processing and browsing the corpus.

Following this opening, the concept of corpus and its classification according different points of view is introduced in Section II. This allows explaining, in Section III, the main decisions taken in the stage of designing such corpus. This section describes the process to create the corpus, the development of a part-of-speech (POS) annotation scheme, and the corpus annotation. Section IV will show *bwanaNet*, a corpus exploitation tool specifically designed for this corpus. Section V will give an overview of the results obtained in using IULACT in several theses and research projects. Section VI outlines some future lines of research involving future extensions of the corpus, as well as its browser and its processing tools. Finally, Section VII draws some conclusions.

II. LINGUISTIC CORPORA: BASIC PRINCIPLES

Prof. J. Sinclair, a pioneer in the field of corpus linguistics, has defined a corpus as "a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" (Sinclair, 1996). Nowadays, it is hard to conceive a corpus without a computer support and a codification to allow individuals and other computers looking up the corpus. Also, they are usually annotated with some linguistic information.

A key point in the corpus design and compilation is the need to guarantee its internal representativeness. Such representativeness does not only depend on the corpus size, but also on the statistically equilibrate presence of all the parameters that define the corpus (such as domains/sub-domains, language usages, registers, etc.); see Biber, 1993. Another important point is that texts to be included in the corpus should be selected according to external criteria. In other words, regardless of the language they contain, but according to their communicative function.

A corpus may be seen from different points of view. The following are some of the criteria normally used to classify corpora, with some examples included:

- domain covered: general (CREA[iii], CTILC[iv], BNC[v], etc.) *vs* specialised language (IULACT, CIBLSP);
- register: popular, journalistic, literary, academic, etc.;
- language of the documents: monolingual *vs* multilingual.
  Monolingual corpus may be further sorted according to the language varieties it covers while multilingual ones according to:
  - correspondence between texts from different language: unrelated, comparable or parallel;
  - alignment level (if texts are parallel): document, paragraph, sentence or word;
- language usage: written, oral or mixed;
- date of publication;
- compilation method: planned *vs* opportunistic;
- linguistic information:
  - flat (just plain text without any linguistic/structural information);
  - tagged with morphological information;
  - syntactically segmented (also known as chunked or parenthesized corpora);
  - syntactic information (or treebanks): like PennTreebank (Marcus et. al, 1993), TIGER (Brants et. al, 2004)[vi];
  - semantic information: like PropBank (Kingsbury et. al., 2002), SALSA (Erk et. al, 2003);
  - pragmatic information: like Penn Discourse TreeBank (Miltsakaki et. al, 2004), RST Discourse Treebank (Carlson et. al, 2003).
- mark-up language: none, proprietary, standard (SGML, XML, etc.);
- available metadata.

The above classification criteria are not mutually exclusive: a corpus may be classified according to several criteria as will be shown in the next section for the corpus IULACT.

In spite of the advantages mentioned in the introduction, corpus compilation shows several problematic issues such as high resource consuming (therefore expensive), difficulty to control representativeness and internal coherence, among others.

III. IULA LSP'S CORPORA

The main objective of the IULACT is to support research and teaching activities from IULA's researchers. To this end, the corpus and its corresponding 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, etc.

According to these aims, a number of design criteria have been set and a specific internal organisation has been adopted. To these issues the following two subsections are devoted respectively. Finally, the last section will show IULACT size figures.

III.1. Corpus design

A corpus is a significant resource that acquires most of its properties only if it is well-designed and carefully compiled and processed. Therefore, special attention must be paid to its design. For these reasons, in the design of IULACT the following decisions have been taken:

- language of the documents: multilingual. The languages involved are: Catalan, Spanish, English, French and German. In this sense, the IULACT is a corpus of comparable documents. However, whenever it is possible, each time a document is added, it is intended to obtain the same document in another language in order to obtain a parallel sub-corpus within the main corpus[vii]. In other words, there is a sub-corpus of parallel documents within the main corpus.
- domain: specialised language. The selected texts belong to five areas: law, economics, medicine, computer science and environment. It was also decided to compile separately a corpus of general language that which could be used as a contrastive element. For practical reasons, this general language corpus mainly comes from Spanish and Catalan newspapers.
- compilation method: planned. Domain experts provided a selection of documents relevant to the domain.
- register: it includes a range of vertical variation ranging from popular science articles to research papers.
- language usage: written;
- date of publication: only contemporary language is included;
- linguistic information: from the very beginning it has been decided that the corpus would be linguistically processed so that, all the words would be fully morphologically tagged. This allows on one side the corpus to be useful for a number of linguistic/terminological studies and at the same time to develop a number of software tools in order to improve the linguistic information included in the corpus;
- mark-up language: it was decided to adopt a standard mark-up language like the ISO 8879, best known as SGML (Goldfarb et. al, 1990). As SGML is just a metalanguage, it was decided to use the recommendations issued by the Corpus Encoding Standard (CES) as a concrete mark-up language[viii]. Nevertheless, the application of this standard has been limited in order to make the mark-up possible according to the limited available re-

sources. In order to simplify process, both the text and its tagging information (structural and linguistic) were saved in the same file.[ix]

Regarding the representativeness, it was decided that the IULACT had to be as representative of the language/s it aims to represent as possible. In order to achieve this goal, each domain is classified from two viewpoints: the own structure of the domain and a simple text typology. This first task is responsibility of experts from each domain who collaborate with the project. As mentioned above, experts provide a selection of documents relevant to the domain and the domain taxonomy. However, the exploitation tool would provide researchers with a mechanism for the selection of a sub-corpus representative to their own investigation topic (see section III).

All the design decisions mentioned above have made the IULACT a flexible tool for asking a number of research questions in the full range of linguistic studies and allowed the development of a number of tools for improving its characteristics.

III.2. Corpus organisation and building process

The CES standard adoption allows having a common internal format for all documents; and makes it easier for IULA to share the IULACT with other similar organisations. According to this standard each document must belong to a type that is formally defined by means of a Document Type Definition (DTD) where the designer declares the internal organization of the documents. In other words, the DTD defines how the text elements must be combined.

The documents were obtained from several sources such as agreement with publishing houses, direct contact with the authors, document scanning and Internet among others. Document scanning was kept to a minimum mainly due to the high amount of errors introduced by the optical character recognition software.

Every document in the 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 auxiliary resources.

The header contains all the necessary information to identify the document: bibliographic information (title, author, publisher, date, ISBN, etc.) and the corpus internal information (internal text classification, text typology, language information with indication that it is a translation, size in number of words and Kbytes, pointers to the samples, etc.).

The third part is the text itself, which is not inserted directly but included by means of pointers that refer to the file/s containing the text.

All this information is structured in three support files related to each other which organise the document electronic version. The first file only 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 different samples.

It should be noted that the document internal organisation mentioned so far is independent from the texts. Therefore, this organisation makes it easy to locate any piece of data.

In brief, after a document has been chosen and converted to an electronic format, the first process is to build the support files and carry out the identification of its formal non-linguistic properties. Only after this stage is developed the document is ready for the linguistically oriented processing.

All documents included in IULACT contain free text. As it is well known, the text processing implies coping with a number of practical issues which do not only derive from the inherent difficulties of NLP. These difficulties usually 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 (decimal signs, dates or proper nouns, among others). But all of them have to be taken into account for producing material useful for linguistic research.

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

The basic pipe lined procedure includes the following tasks (see figure 1):

a) document selection, search, recovering and conversion;
b) integration of the document according to CES recommendations;
c) structural tagging and text handling;
d) morphological analysis and disambiguation;
e) integration in the textual database and
f) exploitation tools.

Such tasks are organised as independent modules with a defined interface, allowing any of them to be easily modified, enlarged or replaced enhancing the results and without affecting the process. Even the text handling task is internally organised in this way to make the improvement of its internal sub-modules easier.

In addition, all maintenance tasks required by the language dictionaries related to the POS tagger are handled by a dictionary management module.

In the following subsections, we briefly discuss each of the processing stages and the main characteristics of the tools we currently use.

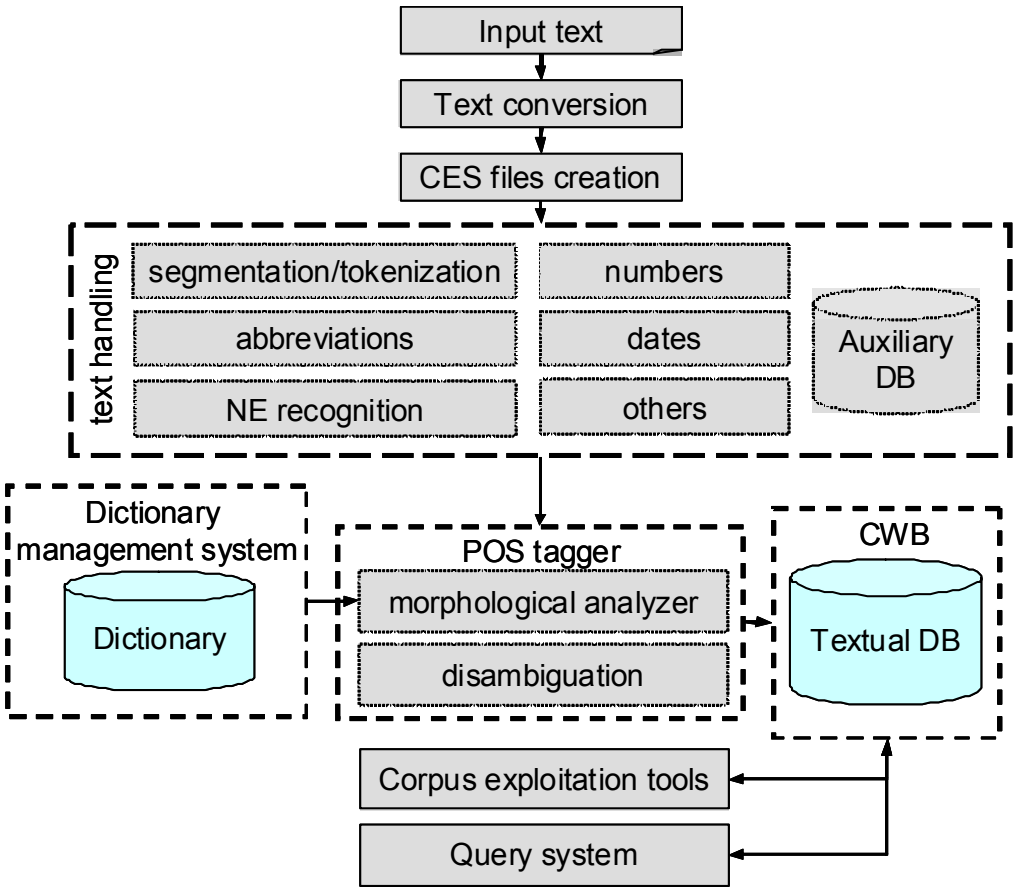

Figure 1. IULACT processing/exploitation flow

III.2.1 Structural tagging and text handling

As mentioned above, the IULACT input text is tagged according to CES standard. However, the whole expressivity of such standard is not used for two main reasons: the mark-up main objective is linguistic research (and not other areas of document processing) and the resources available for this task are limited. Therefore, the structural information is limited to main divisions, head, paragraph and sentence identification. Eventually, we may also take care (though in a limited way) of lists, notes, rendering information and sequences in a foreign language.

NLP is normally accepted as a difficult task but it becomes even more difficult when the text to be processed is actual text and not just laboratory adjusted sentences. A free text contains elements usually considered trivial by humans, but which create difficulties when such text has to be processed by a computer.

Punctuation marks, dates, locutions and proper nouns are just a few examples of the units that increase processing difficulty. 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 single lexical units will prevent the need to cope with the problem of possible unknown words in morphological analysis and with its consequences (not always easily foreseen) in the following processing stage. Also the syntactic analyser can take advantage of such a treatment, thus avoiding the generation of bizarre phrases.

The text handling stage involves the crucial mission to tag any linguistic unit that can be detected by a surface analysis of the text: dates, numbers, proper nouns, abbreviations, labels, etc., as well as to manage the punctuation marks found in the text. In a sense this stage can be considered as the second part of the structural mark-up, as 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 those items in the languages involved in our corpus.

The strings, and some examples, that are processed by the text handler are the following:

- 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 25th 1810*, ...
- Locutions: *a conseqüència de* (ca) [as a consequence of], *en definitiva* (ca/sp) [thus], ...
- Cardinals: *3,14*, *3.14*, *3´14*, *XII*, *twelve*, …
- Ordinals: *1r*, *1er*, *1.er*, *1º*, *1ª*, *1ro*, ...
- Units of measurement: *m/s*, *mt2*, ...
- Labels: a), a., 1), ...
- Abbreviations: art., v. art., ...
- Punctuation marks: , ; : - [ ] ( ) « » " '
- Other: %, €, x2, (5x+2), ...

Many of these items have a different behaviour in each language supported by the IULACT. The paradigmatic example is proper nouns. The algorithm for detecting them is pretty simple but effective (at least for Spanish and Catalan texts). 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 are ambiguous; the letter *i* (ca)

[and], for example, can be a joining item as in ... *per decisió de la Conselleria de Educació i Ciència* ["... as was decided by the Department of Education and Science"] but not in ...*a continuació, Espanya i França van signar l'acord* ["...afterwards, Spain and France signed the agreement"].

Punctuation marks are also difficult to deal with due to their ambiguity. Consider the following set of examples:

- the slash ("/"): it may be found inside a unit of measure ("m/s") or be a conjunction ("the goal is to open/deregulate the market");
- the closing bracket (")"): it can belong to a label or be an independent punctuation mark; for example: "c) Introduce the magnetic card (find the instructions in chapter X)" and
- the dot ("."): it may seem that sentence splitting is a trivial task, but actually it is no so. A full stop may act like a sentence boundary but also be part of a number (as in *10.5*), an abbreviation (*fig.*), an initial (*J. F. Kennedy*) or an item indicator (*1.*).

In order to speed up the process, there is in some cases an additional module that contains the most very common words pre-analyzed. 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].

The text obtained at the end of this stage is parsed against an SGML parser[x] in order to guarantee that the resulting text is free of SGML syntactic errors.

III.2.2 Morphological analysis and disambiguation

The minimal linguistic processing of a given text consists of assigning to each word its lemma and the corresponding POS tag. Typically, such processing consists of two concatenated stages:

a) Morphological analysis[xi]. It means to obtain all possible POS/lemma pairs for every word in the text to be processed.

b) Disambiguation or POS tagging[xii]. This task consists in assigning to each word of a text a single POS tag, which indicates the function of that word in that specific context.

Since recently such tasks were accomplished by two different tools: first a morphological analyzer and then a POS tagger. Due to efficiency reasons, the analysis is simply a dictionary look-up to a database containing the fully-fledged surface forms; therefore, there is a tendency to integrate both processes in just one tool. This is the case of the TreeTagger tool (Schmid, 1994); currently used for tagging the Spanish and Catalan texts. English texts have been processed using the two-level morphological grammar included in the ENGCG package: the ENGTWOL tool (Karlsson et. al, 1995)[xiii].

The full-form dictionary used by the morphological analyzer (see Fig. 1) has been obtained using PALIC (de Yzaguirre et. al, 2001). This tool uses a paradigm based computational morphology system. The basic information for building the dictionaries has been obtained in different ways according to the language.

For Catalan, the dictionary has been obtained semi-automatically from a machine-readable one (i.e. a conventional one available in electronic form): the IEC dictionary (DIEC, 1995), which is the normative dictionary for Catalan. More recently, entries from the DLC95

dictionary (DLC, 1995) which were not in the IEC dictionary have been added to the system. Our Catalan lexicon contains more than 70000 lemmas. The Spanish dictionary has been derived from an electronic version of a paper format general purpose dictionary (DALE 1995). The ambiguity rate for Catalan and Spanish morphologically tagged texts is about 1.7 tags per word.

In all the morphological modules it is guaranteed that the dictionary management is flexible enough to allow new items to be introduced easily. This, of course, occurs quite often, as 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 TreeTagger tool is a POS tagger based on decision trees (a well known tool used in a variety of tasks in computer science). The process is accomplished in two steps: training and testing. During the first stage, the language model to be applied in the tagging phase is built by means of compiling a training corpus of about 500 K words per language. An additional set of 100 K words has been compiled for testing the training resulting model. Both the training/testing corpus contain a sample of all the domains of IULACT that has been obtained starting from the application of a (legacy) tagger used in the past heavily corrected by hand.

The tagset developed at IULA includes about 350 tags (Morel J. et al., 1997)[xiv], but the actual tagset used for the tagger is reduced by 40 % approximately.

The error rate resulting from the tagging process of the testing corpus is about 1% for Spanish texts and 1.2% for Catalan ones. Error analysis shows that, in both cases, errors concentrate on ambiguity classes that require considering a wider context:

- a word following a punctuation mark. Consider the following Spanish sentence: *el sentido común, guía y ayuda a interpretar la vida* ("common sense, guides and helps to understand life"). The word *guía* happens to be ambiguous among verb and noun. The comma avoids the right disambiguation and the word is tagged as noun instead of verb.
- the ambiguity of some pronouns. Consider the following Spanish sentence: *la tasa de cambio no la fija el gobierno* ("exchange rate is not fixed by the government"). In this case the word *la* is ambiguous among determiner and pronoun. This and similar contexts are usually tagged as determiner instead of pronoun because correct disambiguation requires a wider left context.

It may also be necessary to consider the context to the right[xv]. Consider the case of an ambiguity among adjective and past participle followed by a preposition as in the following Spanish sentence: *La lista de personas acusadas de ...* ("The list of persons accused of …"). The word *acusadas* is wrongly tagged as adjective in spite of being a verb form as indicated by the preposition following it.

There is also a number of remaining errors whose variety is such that they become intractable at this level (adjective *vs.* noun or conjunction *vs.* relative pronoun).

III.2.3 Integration in the textual database

Following the linguistic processing the texts are formatted to produce the files which are indexed using the Corpus Workbench Tools (CWB, Christ, 1994). It is a software package designed to process large text corpora of 100 million words and more[xvi] largely used by several similar corpus browse tools. A key point of this software is that it has been specifically designed to be used with a corpus including one or more layers of linguistic annotations. Corpora indexed in this way are stored in a compact binary format that allows efficient searches and data retrieval.

The Corpus Query Processor (CQP) is a main component of the CWB. Its query language allows sophisticated searches for both words (individual or defined by regular expressions) and data saved in the linguistic layers.

The task developed in the framework of IULACT is to provide the necessary glue software to use CWB to efficiently index the corpus and facilitate the task of retrieving fragments of the corpus according to the requirements of the IULA's researchers. The result of the latter objective is *bwanaNet*, a tool for querying the IULACT that is described in section IV.

III.3 Current status

At present the LSP part of the IULACT includes 1,753 documents while the general language module comprises 1,523 documents (21,488 K words). Table 1 shows the distribution of these figures among languages and domains.

From the observation of the figures shown in Table 1, it may be surprising the volume difference among LSP and general corpus. The only reason for such difference is that general corpus is built by automatic downloading of Spanish newspapers.

	Catalan		Spanish		English	
	Docs.	Words	Docs.	Words	Docs.	Words
Law	153	1,685	124	2,085	65	431
Economics	72	1,777	47	1,091	18	275
Environment	78	1,506	55	1,083	86	600
Medicine	236	2,625	402	4,410	284	1,700
Computer science	39	655	67	1,227	27	338
Total	578	8,248	695	9,896	480	3,344
General	769	30,147	752	23,248	2	14

Table 1. Size of the IULACT monolingual corpus and its distribution among languages and domains[xvii]

As mentioned before there is a section of IULACT that may be considered as parallel corpus because its documents are a translation of each other. This section comprises 221 documents (2,858 K words). Table 2 shows the distribution of these figures among domains and languages pairs.

	CA-ES		CA-EN		ES-EN	
	Docs.	Words	Docs.	Words	Docs.	Words
Law	63	412	1	12	1	12
Economics	16	403	5	714	9	146
Environment	2	34	1	9	11	109
Medicine	5	129	0	0	85	560
Computer science	1	28	0	0	21	290
Total	87	1006	7	735	127	1117

Table 2. Size of the IULACT parallel corpus and its distribution among languages and domains

IV. *BWANANET*: THE IULA LSP'S CORPUS BROWSER

One of the main aims of building a corpus is to observe the behaviour of the lexical units included in it. The whole linguistic process is oriented towards the increase of the information associated to lexical units (lemma calculation, morphological disambiguation, etc.), so that such information can be afterwards selectively recovered for linguistic research. Observation of this linguistic information may range from the internal parts of a word to its combinations to create phrases, sentences or even paragraphs. The tools devoted to such corpus exploration have a crucial role in the profit obtained in compiling the corpus.

In order to reach the above mentioned goals, and taking into consideration our researchers' needs, *bwanaNet* has been developed. This corpus browser may be queried from Internet at the following address: bwananet.iula.upf.edu[xviii].

bwanaNet has been designed according to the following criteria:

- have a user friendly interface
- keep usage complexity as low as possible
- be flexible enough to be useful to as many research areas as possible
- be accessible to as many users as possible
- take profit from SGML mark-up
- be multilingual
- keep linguistic knowledge apart from the process of obtaining it
- include facilities to allow easy creation of sub-corpus
- be able to query both monolingual and parallel corpus
- be easy to expand to specific exploration software
- be reasonably fast

The basic units considered by *bwanaNet* are those resulting from the output of the text handler. They may be single units (words, labels, numbers ...), multiple units (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.

As a result of the previously mentioned design criteria, *bwanaNet* allows to easily select either the whole corpus or a sub-corpus. In case a sub-corpus is chosen, it may be defined in several ways:

a) One or more individual documents
b) One or more domains/sub-domains. In addition, the user may optionally add more filtering choosing only documents of a given type and or language status (original, translated text, etc.)
c) Reaching a user specified amount of words. Here there are three different options: maximum number of documents, minimum number of documents or random selection.

Needless to say, the user may save his/her selection for later usage.

After selecting the sub-corpus the user has the possibility to choose the type of query. Basically, such query may refer to single units or, more interesting, to a concordance. Such concordance may be by three different types: simple, standard or complex.

In the case of single units, it is possible to retrieve a list of lemmas, word forms or POS tags. Such list may be raw text or HTML formatted.

In the case of simple concordances, the user has to indicate simply if the query unit is a lemma (default option) or a word form. Optionally the context size may also be established: full (indicating the full textual unit) or the number of units to the left and right of the query unit. In order to speed up the query, the user may ask just for a limited number of concordance lines.

The standard concordance, perhaps the most useful one, allows performing queries according to a user defined pattern. Such pattern may freely combine the basic elements from

corpus: lemmas, word form and POS tags. This kind of querying allows a large number of possibilities, some of them are shown in Figure 2. The meaning of such queries is the following:

a) A standard query consisting of two elements. The first one requiring that the word form be *bronquitis* ("bronchitis") followed by another element that has been tagged as adjective. This may be useful to do a terminological search: in this case for looking up different types of such disease (*bronquitis aguda, bronquitis crónica*, *bronquitis asmática*, etc.)
b) A query may introduce one or more optional units. In this case the sequence is a typical noun-preposition-noun (where the preposition is forced to be *de* (es) [of]) and the second noun may be optionally modified by an adjective. Sequences like: *factor de necrosi tumoral* (ca) [tumoral necrosis factor], and *cas d'intolerancia digestiva* (ca) [case of digestive intolerance] satisfy this pattern.
c) A standard query where some elements are negated and others are optional. The browser looks for a sequence of any verb whose surface form starts with the prefix *pre-* followed by a maximum of ten units that must not be punctuation marks and end with any preposition.

Word form	Bronchitis						pre*		
Lemma			factor	de					
POS		JQ			N5	JQ*	V	^Z(10)	P

a) simple query b) optional elements usage c) using word starting

Figure 2. Examples of standard queries

In addition to the possibilities shown in Figure 2, the user may also:

a) add a multiplicity factor to the POS;
b) sort alphabetically the concordance lines;
c) force the pattern to be starting/ending of the textual element;
d) limit the search to certain textual elements (headers, sentences, etc.).

Finally the user may opt for the complex search. As its name suggests, it is the most complex way to query the IULACT as some knowledge of the CQP query language and the corpus tagset is required. The web page includes the necessary links to obtain some help regarding both topics. In spite of its complexity this search is the most flexible onexix. It allows taking profit from all the data encoded in the tagset (like a POS including a specific value of gender/number or some specific type of verb among many others) and doing some minimal statistics on the query pattern.

Although only in the complex query the CQP query language has been explicitly mentioned, all the queries to the IULACT benefit from the CQP. The only difference is that in the simple and standard queries the web interface takes cares of the query building while the in the complex query is the user who is responsible of the query building.

The parallel corpus section of the IULACT can also be queried by using bwanaNet. In this case, all the selection possibilities (domain, sub-domain, concordances, etc.) are quite the same but the resulting concordance lines include not only the information of the monolingual queries but also the parallel sentence in the other languagexx. Figure 3 shows a typical query to the Spanish-English parallel sub-corpora and the corresponding result.

The query, as shown in Figure 3.a, looks for a word whose lemma is ley ("law") followed by any token tagged as a qualificative adjective. Sequences like ley orgánica ("organic law") or ley general ("general law") should be found with this query. Figure 3.b shows one of the concordance lines issued by bwanaNet. As it is observed the desired pattern is shown for

Spanish while in the English parallel sentence there is no indication about the corresponding words. This is due to the fact that the aligner tool used works at sentence level.

Word form		
Lemma	ley	
POS		JQ*

a) query expression

<s>Cualquier alteración de los límites provinciales habrá de ser aprobada por las Cortes Generales mediante *ley orgánica* .</s>
<s>Any alteration of the provincial boundaries must be approved by the Cortes Generales by means of an organic law .</s>

b) sample of the results

Figure 3. Querying the parallel section of IULACT

V. RESULTS

The IULACT is a resource that has directly or indirectly supported a large number of activities at IULA. A number of PhD theses have been completed in a variety of areas:

- language variation,
- phraseology,
- specialized texts and vertical variation,
- terminology and conceptual relations,
- abbreviation in specialised discourse,
- linguistic model for text summarization,
- neology,
- discourse analysis lexicography,
- etc.

A number of software tools have also been developed or are currently under development:

- morphological analysis (Catalan, Spanish)
- linguistic based tagging (Catalan, Spanish)
- text alignment
- term extraction and detection (Catalan, Spanish)
- neologism detection (Catalan, Spanish)
- treebank creation (Spanish)
- syntactic parsing (Spanish)
- dictionaries management tool
- etc.

Finally, IULACT has been intensively used in several current and past public founded projects. The full list of projects may be found at www.iula.upf.edu/iulaterm/tprojuk.htm. The following are some of the more relevant projects:

- Automatic Acquisition of Lexical information (AAILE2),
- Ontology Enlargement for Information Extraction from Specialised Discourses (RICOTERM3),
- Basis, strategies and tools for automatic extraction and processing of specialized information (TEXTERM3),
- Bases, strategies and tools for the processing and automatic extraction/retrieval of specialized information (TEXTERM2),

VI. FUTURE WORK

Even though the work reported here may be considered completed, IULA is willing to start a new stage with two main purposes: enlarge the corpus and enhance the system performance.

On the one hand, IULA wishes to extend the corpus not only in the already mentioned domains but also embrace to other domains where researchers have shown some interest. On the other hand, there are some tools that nowadays are out-of-date; so a first step should be to improve the performance of such tools as well as introduce more linguistic processes. The main topics to be dealt with are the following:

a) Corpus text handling. To improve the current text handling by making it more robust to foreign formats (PDF, PS, etc.)
b) Corpus management tool. To make easier acquisition, validation from the specialist and support the browsing tool
c) Update English processing tools
d) Syntactical analysis
e) Stand-off markup. Evaluate the use of this architecture and its impact in corpus processing tools.
f) Enhance sub-corpus selection. To allow the *bwanaNet* user to choose documents according to all the available metadata.
g) Speed up consultation. Update the CWB interface to take profit of the latest improvements of this tool.
h) Eliminate the time out web server barrier in order to allow more complex queries to the browser.
i) Include some statistical analysis to the browser results.

Some of these topics are currently under active development while others will be afforded in the near future.

VII. CONCLUSIONS

This paper has described the IULA's LSP corpus and its processing environment. After a brief introduction about the basic principles of corpus linguistics the criteria used in the design of the IULACT have been presented as well as a short description about its internal organisation. This has been followed by a description of the processing environment which included some information about the text handling, the tagger and the integration of the results in a textual database. After that, bwanaNet, the tool that allows querying the corpus from Internet has been described.

It was also shown that the IULACT and bwanaNet have been successfully applied to a number of research issues regarding both linguistics and NLP tools.

Finally, a list with the foreseen work for the near future in order to enlarge the existing corpus and improve both the linguistic processing and the corpus browser has been presented.

VIII. AKNOWLEDGMENTS

The full IULACT project , directed by M. T. Cabré, has been supported by the Catalan research agency CIRIT (project number: CS93-4.009).

References

Biber D. (1993): "Representativeness in corpus design". *Literary and Linguistic Computing*, 8(4), 243-257.

Brants, S.; Dipper, S.; Eisenberg, P.; Hansen, S.; König, E.; Lezius, W.; Rohrer, C.; Smith, G. & H. Uszkoreit (2004): "TIGER: Linguistic Interpretation of a German Corpus". *Research on Language and Computation*, 2(4), 597–620.

Carletta, J.; Kilgour J.; O'Donnell T.; Evert S. & H. Voormann (2003): "The NITE object model library for handling structured linguistic annotation on multimodal data sets". Proceedings of the *EACL Workshop on Language Technology and the Semantic Web*, Budapest.

Carlson L.;, Marcu, D. & M. E. Okurowski (2003): "Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory". In Jan van Kuppevelt and Ronnie Smith (eds.), *Current and New Directions in Discourse and Dialogue*. Kluwer Academic Publishers.

Christ, O. (1994): "A modular and flexible architecture for an integrated corpus query system". Proceedings of the *COMPLEX'94*, Budapest. (http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/)

DALE (1995): *Diccionario Actual de la Lengua Española*. Biblograf, Barcelona.

De Yzaguirre, L.; Matamala, A. & M. T. Cabré (2001): "El lematizador PALIC del IULA". In Muñoz, C. (editor), Conference of the Spanish Association of Applied Linguistics (AESLA), 481-485. Barcelona.

De Yzaguirre, L.; Ribas, M.; Vivaldi, J. & M. T. Cabré. (2000): "Alineación automática de traducciones: descripción y usos en los ámbitos de la profesión, de la docencia y de la investigación traductológica". Proceedings of *IV Encuentros de Traducción: las nuevas tecnologías y el traductor*, Alcalá. Downloadable from http://terminotica.upf.es/CREL/alcala.ps

DIEC (1995): *Diccionari de la Llengua Catalana*. Institut d'Estudis Catalans, Barcelona.
Dipper, S. (2005): "XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation". In Proceedings of *Berliner XML Tage 2005*: 39-50. Berlin,

DLC (1995): *Diccionari de la Llengua Catalana*. Enciclopèdia Catalana, Barcelona.

Erk, K.; Kowalski, A.; Padó, S. & M. Pinkal (2003): "Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation". In *Proceedings of ACL 2003*, Sapporo, Japan.

Goldfarb, C. F. & Y. Rubinsky (1990): *The SGML Handbook*. Oxford University Press.

Karlsson F.; Voutilainen, A.; Heikkilä, J. and A. Anttila (ed.) (1995): *Constraint Grammar. A Language-Independen System for Parsing Unrestricted Text*. Berlin-New York, Mouton de Gruyter.

Kingsbury P. & M. Palmer (2002): "From TreeBank to PropBank". In *Proceedings of LREC 2002*, Las Palmas, Spain.

Schmid H. (1994): "Probabilistic Part-of-Speech Tagging Using Decision Trees". In *Proceedings of the International Conference on New Methods in Language Processing*: 44–49. Manchester, UK.

Sinclair J. (1996): "EAGLES Preliminary recommendations on Corpus Typology". Document EAG--TCWG--CTYP/P. Downloadable from http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Marcus, M.; Santorini, B. & M. A. Marcinkiewicz (1993): "Building a large annotated corpus of English: the Penn Treebank". *Computational Linguistics*, 19(2):313–330.

Miltsakaki E.; Prasad, R.; Joshi, A. & B. Webber (2004): "The Penn Discourse TreeBank". In Proceedings of LREC 2004, Lisbon, Portugal.

Morel, J.; Torner, S.; Vivaldi, J.; De Yzaguirre, L. & M. T. Cabré (1997): "El Corpus de l'IULA: etiquetaris". *Papers de l'IULA*, Sèrie Informes n 18, Institut Universitari de Lingüística Aplicada, Barcelona, Universitat Pompeu Fabra.

---

i Many of these resources may be obtained from repositories like ELRA (www.elra.info) or LDC (www.ldc.upenn.edu).

ii The Institute for Applied Linguistics (IULA, www.iula.upf.edu), is a public institution that belongs to the Universitat Pompeu Fabra (UPF, www.upf.edu) in Barcelona and is closely connected to the School of Translation and Interpretation at UPF. IULA is devoted to both postgraduate teaching and research in applied linguistics, covering basically the following fields: lexicology, lexicography, terminology, computational linguistics, linguistic engineering and language variation.

iii The CREA (Corpus de Referencia del Español Actual) may be queried at: corpus.rae.es/creanet.html

iv The CTILC (Corpus textual informatitzat de la llengua catalana) may be queried at: ctilc.iec.cat

v The BNC (British National Corpus) may be queried at: www.natcorp.ox.ac.uk/

vi TIGER includes also TIGERSearch, a tool for searching in this treebank. Its query language allows to refer to one tree node, two or more tree nodes related by some syntactic relation and the definition of boolean expressions among the previous mentioned relations.

vii This decision has allowed the development of an alignment tool. See deYzaguirre et. al, 2000.

viii See http://www.cs.vassar.edu/CES/CES1-0.html for details.

ix An alternative decision would have been to use stand-off mark-up. This tagging architecture keeps text and tagging information in separate files. It allows to keep the original text untouched and at same time such text may be tagged in different ways (even by using different tools for the same task) and also allow to solve the problem of conflicting hierarchies due to overlapping segments. The main drawback is a greater complexity in the software tools, human visualization becomes more difficult and it is necessary to keep annotations synchronised. See Dipper (2005).

x We are using SGML parser "nsgmls" developed by James Clark. For more information see www.jclark.com/sp/nsgmls.htm.

xi The main models for computational morphology are: paradigm based (valid only for languages with a relatively poor morphology) and finite-state morphology (useful for language having a highly productive morphology).

xii There is a number of different algorithms useful to solve this task like transformation-based error-driven, hidden Markov models, relaxation labelling and decision trees among others.

xiii We plan to move soon to TreeTagger tool for tagging English texts too in order to achieve maximum flexibility and uniformity at the processing stage.

xiv At www.iula.upf.edu/corpus/corpusuk.htm it is possible to see IULA's morphological tagsets details.

xv Most POS taggers take into account just the left context, usually limited to one or two words.

xvi A version of this tool has been recently released as open-source software under GPL license. It can be downloaded from cwb.sourceforge.net.

xvii The word sizes are expressed in thousands.

xviii The number of concordance lines is limited to 2,000 (50 from outside our university). This limit can only be eliminated by using some specific script from command line. Such script may also concatenate an unlimited number of queries.

xix However, the complexity of this query is limited by the timeout of the web server. Also some of CQP characteristics are not available for *bwanaNet* (macros, sub-queries, etc.).

xx This is possible because, in the corpus indexing phase, parallel documents were aligned using an aligner tool (we used the one developed at the University of Stuttgart).