BancTrad: a web interface for integrated access to parallel annotated corpora

Toni Badia, Gemma Boleda, Carme Colominas, Agnès González, Mireia Garmendia, Martí Quixal

Universitat Pompeu Fabra
Rambla 30-32 ,
E-08002 Barcelona
{toni.badia,carme.colominas,marti.quixal}@trad.upf.es, gemma.boleda@iula.upf.es

Abstract

The goal of BancTrad is to offer the possibility to access and search through (parallel) annotated corpora via the Internet. This paper presents the design of the whole process: from text compilation and processing to actually performing queries via the web, while it describes as well its technical architecture.

The languages we work with are Catalan, Spanish, English, German and French. Queries are possible from any of these languages to Spanish and Catalan and vice versa (but not between the language pairs formed by French, German and English). The texts go first through a pre-processing and mark-up stage, then through linguistic analysis and are finally formatted, indexed and made ready to be consulted. The web interface has been created through the integration some *ad hoc* applications and some ready-to-use ones. It provides three different levels of query expertise: basic, intermediate and expert.

The paper is structured as follows: section 1 gives an overview of the project; section 2 describes the text compilation process; section 3 explains the corpora building and parsing stages; section 4 details the search machine architecture; finally, section 5 describes foreseen applications of BancTrad.

1. Overview

The original idea of BancTrad[1] was to obtain a tool with pedagogic applications (see work done e.g. by Gaspari, Hansen, S.) especially thinking of translation and interpreting courses held at the Translation and Interpretation Faculty (FTI) of the University Pompeu Fabra (UPF). It was meant to be a translation databank that could serve both teachers and students to search for prototypical translations or texts containing special features that would make them interesting from the translator's point of view. Afterwards, the target user of BancTrad was broadened to e.g. professional translators and linguists (see section 5), through the creation of different search modes and the expansion of the expressiveness of the queries, in order to adapt to the user needs or knowledge.

As an annotated translation databank, BancTrad offers the possibility to work with Catalan, Spanish, English, German and French. Queries are possible from any of these languages to Spanish and Catalan and vice versa (but no queries are possible between the language pairs formed by French, German and English), as well as between Catalan and Spanish in both directions. The web page of the project can be accessed from http://glotis.upf.es/bt/index.html

2. Text collecting, extra-linguistic tagging and alignment

The corpora in BancTrad aim at being representative for translated texts. In other words, they don't have a normative character but a descriptive one. Therefore we have chosen to collect documents from very different sources, representing a variety of text types, subjects and registers.

The main sources we have focussed on are faculty professors, work done in translation courses, publishing houses and the Internet. Many faculty professors work also as freelance translators, which constitutes a good source of high quality translations. Besides, the fact that we include (supervised) work done in translation courses can have many advantages regarding academic self-evaluation. Specially, because they give evidence of the text types, subjects, etc., which have been worked on with pedagogical purposes. As for translations from the Internet, some supervision is done on them before they are selected to be introduced in BancTrad (for the sake of quality).

Selected texts are semi-automatically processed to be marked up with SGML tags and aligned with their respective original texts. Both the originals and the translations are marked up with some extra-linguistic information by means of a special MS Word form coded in Visual Basic (see Fig. 1).

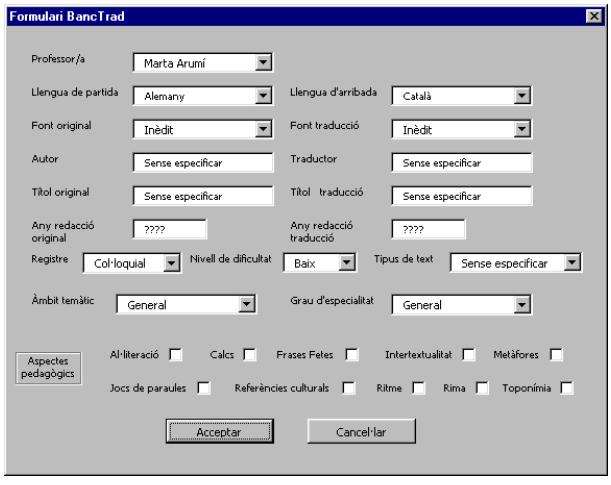


Figure 1: MS Word form used for the mark-up of extralinguistic features of the texts

[1] This project is running under the auspices of the "Programa d'Innovació Docent" (Educational Innovation Program) sponsored by our university (Universitat Pompeu Fabra) and has also been partially financed by the Spanish Government and by the 2001FI 00582 grant from the autonomous Government of Catalonia.

This mark-up takes the following parameters into account:

- name of the person who introduced the aligned texts (i. a., in order to track translation quality)
- source and target languages
- original and translation references
- publication date (for both the original and the translation)
- register (colloquial, standard, learned, etc.)
- type of text (normative, descriptive, literary, etc.)
- subject matter (economy, science, politics, etc.)
- degree of specialisation (low, middle, high).

Besides these parameters, and bearing in mind that BancTrad was originally conceived as a tool with pedagogic applications, we include information on certain aspects such as idioms, metaphors, puns, degree of difficulty, etc. All of these parameters, as well as the information coded within them, were consensuated with the teachers and researchers of the FTI. It is relevant to note that this mark-up allows us not to make a rigid classification of the texts in the corpus (see section 3).

By clicking on the *Acceptar* ("Accept") button, the options selected in the form are marked in the text in SGML format and a script tags the paragraph structure of the document. Otherwise, this very valuable piece of information on the text structure would be lost in the alignment step.

Texts are aligned at a sentence level with the align tool of the DéjàVu Database Maintenance, software by Atril (http://www.atril.com). DéjàVu aligns texts and allows editing in quite a user-friendly way.

The tasks described so far, although only semi-automatic, require neither special skills in computing nor much time (the time to go through them for a 400 word-long text -both source and target texts- is 5 to 10 minutes). We could have chosen to tackle the alignment task fully automatically instead, but the error rate of automatic aligners (notably errors in sentence identification) would have increased too much the error rate in the subsequent linguistic analysis. However, it should be kept in mind that, according to our architecture, the use of a particular tool for the mark-up and alignment independent of the rest of the process, so that other tools could be used in the future.

Finally, the texts are transferred to our Linux server to proceed with the text processing, which from this moment on will be completely automatic.

3. Linguistic Processing and Corpus Building

Once the texts are in the server, they undergo two further steps: linguistic tagging and corpus formatting. Both steps are completely automatic.

3.1. Linguistic Processing

Each language follows a different tagging process. On the one hand, Catalan texts are parsed with CATCG (Badia *et al.* 2000), a Catalan shallow morphosyntactic parser based on a constraint grammar developed by the Computational Linguistics group at UPF. Spanish texts will be handled with a Spanish version of it in a year's time. On the other hand, the linguistic analysis for English, German and French texts is made with TreeTager, a part-of-speech tagger developed at the IMS (see Schmid 1995, 1997). Both CATCG and TreeTager are shallow parsers.

La	**noia**	**de**	**el**	**port**	**de**	**Barcelona**	**dorm**
the	***girl***	***of***	***the***	***harbour***	***of***	***Barcelona***	***sleeps***

```
<s id="1">
La          el          Det         AFS         DN>
noia        noi         Nom         N5-FS       Subj
<contrac forma="del">
de          de          Prep        P           <NA
el          el          Det         AMS         DN>
</contrac>
port        port        Nom         N5-MS       <P
de          de          Prep        P           <NA
<enty>
Barcelona               Barcelona               Nom         N4G6S  <P
</enty>
dorm                    dormir      Verb        VRR2S-      VPrin
.           .           .           .           PT
</s>
```

Figure 2: Input and output of CATCG

It is important to note that, despite the use of different tagging tools for exploiting the linguistic information of our texts, all languages receive a minimum of uniform kind of information: lemma and POS tag (syntactic function is only there for Catalan). Thus, all the languages can be processed and made queries upon in the same fashion, independently of the tagging tool used. This favours modularity, for the linguistic processing of a certain language can be modified without changing any of neither the other linguistic processes nor the interface. We now proceed to roughly characterize CATCG and TreeTagger.

3.1.1. CATCG

CATCG is a linguistic-based parser that assigns each word a lemma, a POS tag and a syntactic function. It uses three major devices:

a) a Perl module for the preprocessing
b) a morphological tag mapping tool that uses a word-form dictionary created with a morphological generator developed at UPF (Badia *et al.* 1997)
c) three grammars using the Constraint Grammar formalism developed at the University of Helsinki (Karlsson et al. 1995, Tapanainen 1996), which perform the morphosyntactic disambiguation task and the partial syntactic analysis.

Fig. 2 gives an example of the input and output of our system. The SGML tags are the result of the preprocessing, and in the example they mark a contracted form, an entity and the sentence boundaries. The columns list the linguistic information: word form, lemma, part of speech tag, complete morphological information in an compressed tag and syntactic function (in order of appearance). The last piece of information is shallow and partial in the sense that it doesn't fully indicate dependency: note that the

preposition *de* (“from”) in the PP *de Barcelona* gets a tag indicating that it modifies a noun to its left (<*NA*, left adjoining Nominal Adjunct); however, no clue is given about whether it modifies *Barcelona* or *port*.

3.1.2. TreeTager

TreeTager is a probabilistic tagger that uses decision trees. It provides each word with a lemma and a POS tag (at the moment, no syntactic information is given).

3.2. Corpus formatting

After being annotated, the text files are eventually formatted and processed with the Corpus WorkBench (CWB) tools, a set of linguistic information exploitation tools developed at the IMS in Stuttgart (Christ 1994; Christ *et al*. 1999[2]). Thus we build the actual corpora making them ready to be consulted with CQP, the Corpus Query Processor, a tool from the CWB. This tool allows very flexible and expressive queries for any of the pieces of information encoded (be it the word form, lemma, POS tag or syntactic function). In fact, as a far as one gives corpora the adequate structure, one can have as a many attributes as one pleases.

One of the most significant (to us) features of the CWB is the fact that it can process aligned corpora. Not only is it possible to view the aligned sentences, but it is also possible to place restrictions both on the source and on the target language in a query (see section 5). It has also been crucial to us the special module that lets CQP interacting with the web (see next section).

4. The search machine and the web Interface

Technically speaking, the novelty of BancTrad is the integration of several tools that make available parallel annotated corpora via the Internet. This entails that the system has to be able to (1) interpret the query made by the user, (2) search for the query, (3) present the results. For this purpose, two devices were needed: a graphical user interface (GUI) with a fill-in form and an external program interface (to allow browser/server communication)

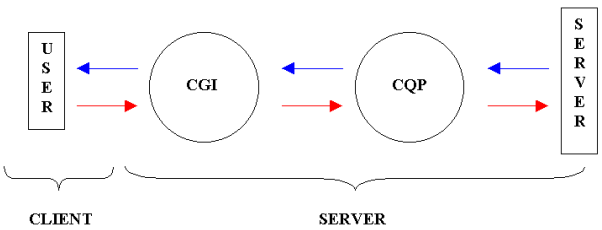


Figure 3: Query routing through the client/server architecture (query from left to right, results the other way round)

a) The GUI for query input

The GUI is intended to be adaptable to the user expertise, to have open access and to be platform independent. For our GUI to accomplish the two last features, an HTML-based interface seemed to be the best option. To qualify for the first one, the interface had to offer at least three search possibilities: common, intermediate and expert mode (see next section for details).

b) The external program interface

This is the module of the architecture that actually makes the query processing. It interprets the user's query, it searches for it in the corpora and gives the result back. The program that does the work is commonly called a cgi (Common Gateway Interface, term whose original sense has been extended to mean “external program interface”). Our cgi is composed of the following packages:

i) Common Gateway Interface (CGI)

The CGI (properly so named) is a standard device to interface with information servers (such as HTTP servers). It passes a web user's request on to an application program and gives the resulting data back to the user. Herewith the server interprets the user’s query.

ii) HTML::Entities

This formatting package ensures that special characters (tildes, cedillas, etc.) are properly transferred during the client/server session.

iii) WebCqp::Query, a web adapted version of the CQP

This package was designed by the creators of the CWB (see above) to let it interact with the web. It can perform the same kind of queries that CQP performs in its PC-Linux version. It thus allows a powerful query setting through regular expressions, access to linguistic tags (through the defined number of features in the corpora) and aligned corpus querying.

5. Exploiting BancTrad

This section outlines different ways in which to exploit BancTrad, from two different but related perspectives regarding its potential users. It describes the search possibilities that BancTrad offers (section 5.1), which relates to the user's level of expertise. Besides, it sketches some possible applications for which BancTrad is indicated (section 5.2), which relates to the user's professional or academic profile.

5.1. Search possibilities

5.1.1. Three levels of expertise

The web interface of BancTrad had to enable the users to access the corpora without having to be experts neither on linguistics nor on regular expressions. Moreover it had to offer the possibility of exploiting the full-fledged regular expressions that CQP allows, as well as the chance of profiting from the quite detailed linguistic annotation of the corpora. Therefore, BancTrad offers three different search modes (corresponding to levels of query expertise):

a) **basic mode:** allows searching for sequences of specific word forms (with possibly their equivalence in a target language).

b) **intermediate mode:** allows searching for sequences of five

[2] See also the web page of the CWB: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

quadruples (form, lemma, morphosyntactic tag, and syntactic function), including the iteration of identical elements

Fig. 4 is a screenshot of a search in this mode: it searches for causative constructions from Catalan into English, that is, for the causative verb *fer* followed by any verb (see next section for the results).

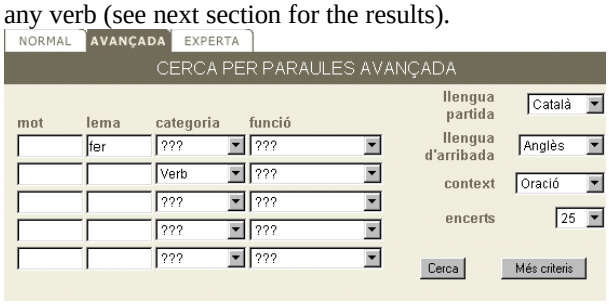


Figure 4: Screen shot of the intermediate query mode of BancTrad

expert mode: to set queries expressed in the full regular language provided by CQP.

5.1.2. Restrictions on extralinguistic features

Additionally to the word units searched for, the user can place restrictions on extra-linguistic features of the texts containing them. This is possible through the initial mark-up stage (see section 2) while formatting the corpora. Thus, through an extended web-form, the user can restrict the occurrences of e.g. the word "bank" to appear in economic texts.

This kind of mark-up gives rise to a different search possibility, planned for the original purpose of BancTrad (which was being useful for teaching purposes at the FTI): the full text query, which allows the user to search for complete texts and their translation, restricting them by the extra-linguistic features mentioned above. Fig. 5 shows a text query in which the user wants to retrieve essays (*Assaig*) on Arts originally written in German (*Alemany*) and translated into Spanish (*Castellà*).

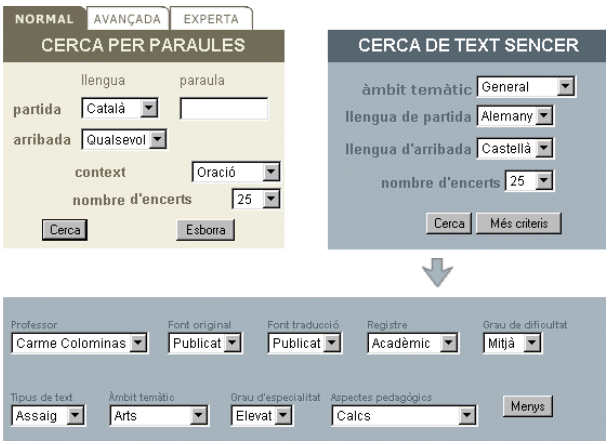
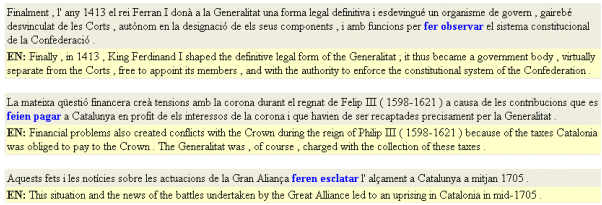


Figure 5: Screenshot of the text query mode of BancTrad

5.1.3. Showing the results

As for the presentation of the results, they are shown by default as aligned full sentences, although it is foreseen that the user can switch to other presentation forms: a full paragraph or just some words to the left and/or right sides of the query target. Of course all the capabilities listed so far are indebted to the Corpus Query Processor that we use as a searching engine.

Fig. 6 shows some of the results for the query on causative constructions made on section 5.1.1:


Finalment , l' any 1413 el rei Ferran I donà a la Generalitat una forma legal definitiva i esdevingué un organisme de govern , gairebé desvinculat de les Corts , autònom en la designació de els seus components , i amb funcions per **fer observar** el sistema constitucional de la Confederació .
EN: Finally , in 1413 , King Ferdinand I shaped the definitive legal form of the Generalitat ; it thus became a government body , virtually separate from the Corts , free to appoint its members , and with the authority to enforce the constitutional system of the Confederation .

La mateixa qüestió financera creà tensions amb la corona durant el regnat de Felip III (1598-1621) a causa de les contribucions que es **feien pagar** a Catalunya en profit de els interessos de la corona i que havien de ser recaptades precisament per la Generalitat .
EN: Financial problems also created conflicts with the Crown during the reign of Philip III (1598-1621) because of the taxes Catalonia was obliged to pay to the Crown . The Generalitat was , of course , charged with the collection of these taxes .

Aquests fets i les notícies sobre les actuacions de la Gran Aliança **feren esclatar** l' alçament a Catalunya a mitjan 1705 .
EN: This situation and the news of the battles undertaken by the Great Alliance led to an uprising in Catalonia in mid-1705 .

Figure 4: Screen shot of the intermediate query mode of BancTrad

5.2. Applications of BancTrad

There are several uses one can think of for BancTrad. Of course, the most direct and obvious one is the one for which the parallel databank was thought: educational use. But there are at least two other kinds of applications that were held in mind while developing the project: research and professional applications. The three of them are outlined, with some examples, in this section.

5.2.1. Teaching

For educational purposes, all of the search modes (be it string or text queries) outlined in the previous subsection are relevant. However, as the full text query has already been exemplified, we will concentrate on the first one. The string equivalence query, which we foresee to be the most significant application for the corpora included in BancTrad, is the search of bilingual equivalences among language pairs. This includes the search of word equivalence, restricted by its form in one of the languages, by its lemma, or by its form or lemma and its morphosyntactic tag. Thus typical searches (which demand different levels of expertise in the search mode) could be:

a) translation of the English form 'stores' into Catalan. Result: *botigues* (noun), *guarda* (verb).
b) translation of the English lemma 'store' into Catalan. Result: *botiga*, *botigues* (noun), and the whole paradigm of the verb *guardar*.
c) translation of the lemma 'store' with part-of-speech 'verb' into Catalan. Result: the whole paradigm of the verb *guardar*.

Note that as in standard corpus search engines, word forms and lemmata can be searched for in specific contexts, as well as particular combinations of forms, lemmata or part-of-speech tags. For example:

d) translation of the gerundive form of the verb 'indicate' right after a colon.

In addition, a specific search condition on the aligned text can be set. For example:

e) translation of the gerundive form of the verb 'indicate' just after a colon provided that in the translated sentence into Catalan no gerundive is present; alternatively, provided that the verb 'indicar' is used.

5.2.2. Professional and research applications

In fact, these kind of applications just follow from the examples described above and the characteristics of the corpora in BancTrad. On the one hand, as far as the corpora are real translated texts (see section 2), and provided the search possibilities sketched above, BancTrad appears to be a useful tool for professional translators. They could look for evidence of previous translation decisions and even have the information of the person in charge for that translation.

On the other hand, linguists and translation theorists (see work done by Baker, M. and Teubert, W.) could also take advantage of this search engine. In fact, this is something we have already been doing with the grammar-developing task we have been carrying on for the last three years. We can retrieve data such as most frequent readings, syntactic structures, etc. This helps us concentrate on problems arising when dealing with written text and develop more data-driven linguistic-based grammars. It is also interesting to note that searches can be made on a sole language, that is, they must not be bilingual.

Other possible applications for BancTrad include creating further Language Resources, such as multilingual dictionaries, chunkers, stochastic-based machine translation systems, etc.

5.2.3. An added value

Finally, it is important to note that an added value to BancTrad's web interface is the fact that it can incorporate other corpora (also monolingual ones) with little amount of work. This would enable our users to query on several corpora, not only the ones prepared at the FTI, in a user-friendly and familiar web interface. For instance, we already have the British National Corpus as part of our searchable corpora and we are planning to integrate the Frankfurter Rundschau corpus soon as well.

6. Conclusions and future work

We have presented a parallel-annotated corpora web interface that integrates several linguistic tools, both for exploiting linguistic information and for exploiting the linguistically enriched texts. It was originally thought to be a translation teaching help tool, but its possibilities have been so extended that it can be of use to both common public and professional users.

Technically speaking, BancTrad integrates tools from different techniques and fields. On the one hand, we use parsing tools developed at our centre, which have been developed with linguistic techniques. Moreover, we are planning to use parsers developed with stochastic techniques (TreeTagger, see above). On the other hand, we have been taking advantage of several ready-to-use packages for client/server interaction. Thus, we feel our project provides evidence of the necessity of academic co-operation to produce tools for the exploitation of linguistic information.

7. Acknowledgments

Thanks all teachers of the FTI for their collaboration. Feedback from the anonymous reviewers was also very useful.

8. References

Badia, T., À. Egea & T. Tuells (1997) CATMORF: Multi-two level steps for Catalan morphology. In *Demo Proceedings of the Conference on Applied Natural Language Processing*. Washington

Badia, T., Boleda, G., Bofias, E. & Quixal, M. (2001) A modular architecture for the processing of free text. *Proceedings of the Workshop on 'Modular Programming applied to Natural Language Processing'* at *EUROLAN 2001*. Iasi, Romania.

Christ, Oliver (1994) "A modular and flexible architecture for an integrated corpus query system", *COMPLEX'94*, Budapest

Christ, Oliver, Schulze, Bruno M. and König, Esther (1999) *Corpus Query Processor (CQP). User's Manual*, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart

Karlsson, F. et al. (1995) *Constraint Grammar: a Language-Independent Formalism for Parsing Unrestricted Text*, Mouton De Gruyter: Berlin/New York

Schmid, Helmut (1995) Improvements in Part-of-Speech Tagging with an Application to German, in *Proceedings of the ACL SIGDAT-Workshop*, pp. 47-50

Schmid, Helmut (1997) Probabilistic Part-of-Speech Tagging Using Decision Trees, in Daniel Jones and Harold Somers, editors, *New Methods in Language Processing Studies in Computational Linguistics*, UCL Press, London, pp. 154-164

Tapanainen, P. (1996) *The Constraint Grammar Parser CG-2*, Department of General Linguistics, University of Helsinki, Helsinki, Publications, number 27.