BancTrad: a web interface for integrated access to annotated corpora

Carme Colominas

Universitat Pompeu Fabra
Rambla 30-32
08002 Barcelona
carme.colominas@upf.edu

Abstract

BancTRad (BT) is a web interface for access to annotated corpora developed by the Computer Linguistic Group (GLiCom) at the Universitat Pompeu Fabra in Barcelona. This contribution includes a brief presentation of both the corpora building and the search machine architecture as well as a demonstration of the whole engine. 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.

1 Overview

From the main page of BT (http://mutis.upf.es/bt/english/index.htm) users can access different monolingual corpora (BNC, Frankfurter Rundschau) as well as a parallel corpus built in our Department for Translation and Philology.. The languages included in this parallel corpus 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). Both monolingual and parallel corpora are linguistically annotated with lemma (except BNC) and POS. Our parallel corpus is furthermore annotated with extra linguistic information (original and translation references, publication date, type of text, register, subject etc.).

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; finally, section 4 describes the search possibilities of BancTrad.

2 Text collecting, extra-linguistic tagging and alignment

Texts go first through a pre-processing and mark-up stage, which is semi automatically and can be done via web (password required) by means of a small application. Through the wizard-like user interface BancTrad Manager (available as a Java Web Start application under http://mutis.upf.es/~textosbt/) the user provides information about the text(s) to be inserted into BancTrad; this mark-up takes the following parameters into account:

- Name of the person who introduced the aligned texts
- 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)

The selected options are marked in the text in SGML format and a script tags the paragraph structure of the document. Afterwards files in different formats are converted to plain text, aligned at sentence level with an align tool and uploaded. Once the texts are in the server, they undergo two further steps: linguistic tagging and corpus formatting. Both steps are completely automatic.

3 Linguistic Processing and Corpus Formatting

Each language follows a different tagging process. On the one hand, Catalan texts are parsed with CATCG (Badia *et al.* 2001), a Catalan shallow morphosyntactic parser based on a constraint grammar developed by the Computational Linguistics group at UPF. 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. 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.

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[i]). 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. It has also been crucial to us the special module that lets CQP interacting with the web.

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.

4 Search possibilities

The web interface of BancTrad enables the users to access the corpora without having to be experts neither on linguistics nor on regular expressions. Therefore, BancTrad offers two 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) expert mode: allows searching for sequences of five quadruples (form, lemma, morphosyntactic tag, and syntactic function (only for Catalan), including the iteration of identical elements and the negation. For example, the user can search for Spanish translations of the English lemma *light* as an Adjective, of sequences of Adjective-Adjective - Adjective-Noun other of sequences of *have to be* followed by some other category except an Adjectiv. 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 while formatting the corpora.

As for the presentation of the results, they are shown by default as aligned full sentences, although the user can switch to other presentation forms: a full paragraph or 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.

References

(Badia et al. 2000) T. Badia, G. Boleda, M. Quixal, E. Bofias, A modular architecture for the processing of free text. *Proceedings of the Workshop on Modular Programming applied to Natural Language Pro cessing at EUROLAN 2001*, Iasi, 2001.

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

(Christ et al. 1999) O. Christ, B. Schulze and E. König *Corpus Query Processor (CQP). User's Manual*, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart, 1999.

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

(Schmid 1997) H. Schmidt, 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, 1997.