Corpógrafo V.4 – Tools for Researchers and Teachers Using Comparable Corpora

Belinda Maia, Sérgio Matos
Linguateca – PoloCLUP
Faculdade de Letras da Universidade do Porto
Via Panorâmica s/n
4150-563 Porto
Portugal
E-mail: bmaia@mail.telepac.pt, sgmatos@letras.up.pt

Abstract

This paper describes the response by the NLP project Linguateca to the needs of researchers, teachers and students in the areas of terminology, translation, contrastive linguistics, and related areas, for user-friendly tools for building and using comparable corpora. It will present the latest developments of the Corpógrafo, a suite of freely available and fully integrated online tools that allow for individuals or small groups to do linguistic research, or simply study the implications of corpus and terminology research for translators. The new developments include considerable improvements to the previous corpus and terminology database tools, a parallel corpus aligner, an aligner of parallel segments in comparable corpora, the integration of the NooJ engine with dictionaries in English, French and Portuguese, and a lexical / phrasal database structure designed for both normal lexicography and for the storage and analysis of the multi-word expressions of interest to those researching genre, text or discourse analysis. The results of this research will, in turn, contribute to the enrichment of the Corpógrafo tools.

1. Introduction

The reasons for building comparable corpora vary considerably, but the call for papers for this workshop focuses on several of the computational interests involved. We shall begin by referring briefly to the way the Corpógrafo functioned in the past and describe some of the improvements made for finding terms in special domain comparable corpora. We shall then concentrate on the possibilities of our new tools for collecting and analyzing phrases in comparable corpora. It is hoped that the data thus acquired can, in turn, be used to enrich and develop the tools themselves. Our approach is based on our own experience of the symbiosis needed between developing such tools, finding a practical research usage for them, and improving them using feedback from users.

The tools are not particularly new individually, but as an integrated suite they are useful. Before we discuss the computational tools for linguistic analysis of comparable corpora, however, we shall begin by reflecting briefly on the nature of comparable corpora and how they can serve as a basis for a wide variety of research projects for which computational tools offer possibilities.

2. Reasons for Building Comparable Corpora

What is a comparable corpus? It is not that easy to either define a comparable corpus or, having done so, to find suitable texts with which to build one. However, it is clear that, more often than not, comparable corpora are seen as domain or subject specific, such as texts about composite materials, fire hazards, or pet cats. Once the domain has been chosen it is also normal to restrict the genre so that, for example, scientific texts and publicity texts are paired separately. Besides this, it is often assumed that comparable corpora are bi- or multi-lingual.

As has been said in the call for papers, comparable corpora are of increasing interest because of the scarcity of reliable parallel corpora. Most of the workshop topics contemplate comparable corpora which are bi or multi-lingual, and presume that one will build a corpus of this kind for mining information of various kinds. As comparable corpora also have the advantage that most specialized texts will have been written by domain experts, they will therefore be more reliable for terminology extraction than translations that, despite all the recommendations of the European Norm EN 15038, may not have been revised by an expert.

Another advantage is that texts in comparable corpora are usually written by native speakers and should be better examples of the language or languages being studied. This means that they can also serve for various kinds of genre, text and discourse analysis.

There are also several reasons for creating monolingual comparable corpora. Someone may wish to discover why one text is more successful with its audience than another as, for example, in publicity texts. Others may want to study different authors, in the attempt to find out who influenced whom, and this has applications for discovering plagiarism and for forensic linguistics. Yet others may wish to create a corpus of exemplary texts in different domains and genres in English and extract phrases that would be useful for the growing number of non-native English speakers who feel obliged to write directly in English.

3. Linguistic v. Computational Approaches

It should be clear by now that our approach will necessarily have to combine computational tools with ‘manual’ intervention by linguists, and we believe that it is essential to unite the two skills for better research. Computational approaches tend to favour acquiring large quantities of text in the hope that the number of examples of the required information, terms, or phrases will prove significant enough to allow one to safely ignore anything that appears infrequently or not at all. There are a variety of computational methods for finding texts in certain domains, an example of which is BooTCaT (Baroni & Bernardini, 2004). However, one of the problems of dredging the internet for such texts is that a lot of repeated material and noise come back with whatever it is we are looking for. Internet mirror pages and plagiarism are responsible for much of this.

On the other hand, corpora consisting of texts that have been carefully chosen by a linguist may not need to be enormous in order to provide useful information. For several years now, translation teachers have encouraged students to create small corpora for specific assignments, called 'do-it-yourself' corpora (Maia, 1997) or 'disposable' corpora (Varantola, 2003), and they have proved pedagogically useful, despite their limitations for NLP research.

The design of the Corpógrafo was based on the assumption that individuals would invest time in finding texts that suited their research needs, but needed help in converting them into plain text and combining them selectively into searchable corpora. Choosing the texts is in itself part of the pedagogical process. Cleaning up a large automatically extracted corpus may actually take much longer and the process is hardly educational for the trainee translator, terminologist or linguist. Now that the Corpógrafo is being extended to more general language analysis, the need to create carefully chosen corpora continues to be relevant.

4. Building Comparable Corpora and Related Databases

The Corpógrafo was originally designed for the building of comparable corpora in special domains for the extraction of terminology, but the tools can be used for any kind of corpus. It offers a complete framework for working with text, from extracting text from different types of files, to editing and cleaning the texts, to grouping the files selectively into separate monolingual corpora, and using simple concordance tools for studying these corpora.

When the corpora have been created it allows users to create related multilingual databases in an efficient manner, by using the system's semi-automatic methods for registering metadata on the corpora, extracting lexical and phrasal items, as well as term candidates, using n-gram tools with or without filters, and finding definitions and semantic relations between lexical items or terms using underlying list of lexical patterns (Sarmento et al., 2006). Once the initial texts, monolingual corpora and related multilingual databases are operable, statistics on the frequency of lexical items or terms and the way they occur in the texts in a corpus can be generated automatically.

A new feature is a tool to bootstrap information from the internet directly into Corpógrafo's file preparation system using a starting list of seed expressions from this statistical information. This feature follows the same idea as implemented by the BooTCaT toolkit (Baroni & Bernardini, 2004), but allows the researcher to select and process relevant texts as needed.

This general workflow in Corpógrafo and an overview of the system's architecture are illustrated in Figure 1. All data added by the users and associated metadata, are kept on the user's working area. Operations on these data are managed by Corpógrafo, and are available to the users through graphical interfaces to the system's functions.

5. Genre Specific Comparable Corpora

One of our earliest tools was a simple n-gram tool which served to help find the lists of expressions used to find definitions and semantic relations in the Corpógrafo. It also drew our attention to what people call 'lexical bundles', 'multi-word units/expressions', 'paraphrases', and similar phenomena (Maia et al., forthcoming). Silva (2006) used the tool to search for discourse phrases in information on art exhibitions in English and Portuguese and was able to show the differences in the text conventions for this genre in the two languages/cultures. He first searched his corpora using the n-gram tool, selected expressions that could be considered discourse connectors, like *in order to, at the same time, for example* and then classified these expressions in terms of discourse markers, such as 'purpose', 'inclusion' and 'exemplification', respectively, He then analysed the examples in comparable corpora of about 128,000 words for each language, quantified the results and drew certain conclusions about the cultural differences between English and Portuguese conventions when writing on the subject of art exhibitions.

This experiment led us to create the possibility of creating multilingual lexical and phrasal databases with appropriate classifications for lexical and syntactic information, as well as for lexical and semantic conceptual relations, similar to those used in the terminology databases. This will allow us to develop Silva's methodology and apply it to further research.

The new lexical/phrasal database structure also offers the possibility of classifying the word or phrase for the effect of discourse analysis. The choices of classification offered are derived from the Rhetorical Structure Theory discourse relations developed by Maite Taboada (see: http://www.sfu.ca/rst/index.html) and adapted for Portuguese by Rui Silva. It is also possible to create one's own classifications, if one wishes.

The objective here is to develop lists of expressions that will semi-automatically retrieve the discourse elements according to this classification. Since there is a growing interest at both a research and pedagogical level in raising awareness of the conventions of different genres and text types and comparing these conventions in different social and cultural situations, this development offers new opportunities for this type of analysis.

Another development is the use of the NooJ engine (see: http://www.nooj4nlp.net) to query the corpora for phrasal units, using regular expressions and grammatical (part-of-speech) tags. This now works in French, English and Portuguese. In the future, we plan to allow users to save and edit the NooJ annotation so that it becomes possible to correct the results and even add – semi-automatically - tags related to one's own discourse analysis or similar

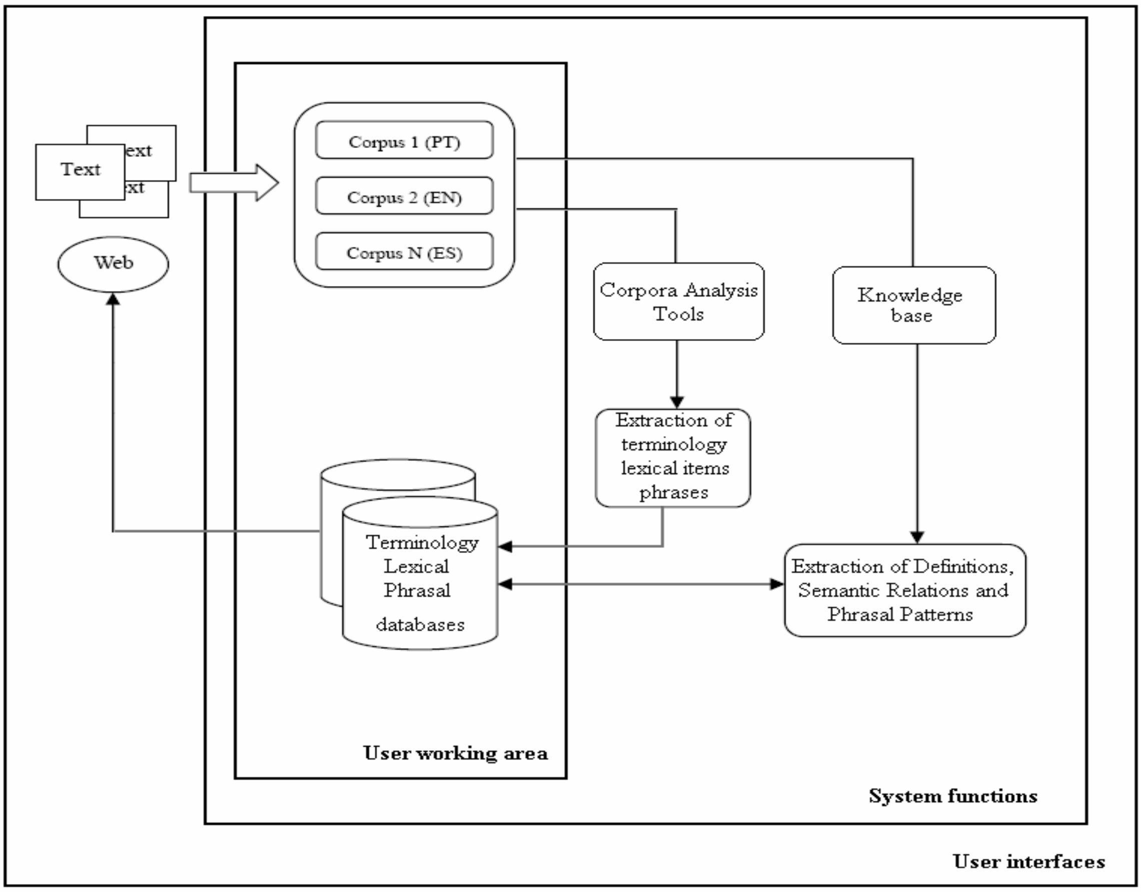
Text
ext
xt
Web
Corpus 1 (PT)
Corpus 2 (EN)
Corpus N (ES)
Corpora Analysis Tools
Knowledge base
Extraction of terminology lexical items phrases
Terminology Lexical Phrasal databases
Extraction of Definitions, Semantic Relations and Phrasal Patterns
User working area
System functions
User interfaces

Figure 1: General workflow and architecture of Corpógrafo

research. This will help with the identification and correlation between languages of phrasal, syntactic or discourse patterns, and once these patterns are entered in the multilingual databases, they can be observed using the concordancing tools for parallel and comparable corpora described below.

6. Aligning Parallel and Comparable Corpora

To address the need of some of our users, we have integrated a sentence alignment tool (from IMS-CWB) in the Corpógrafo environment. The alignment is performed without the user's interaction, allowing users to create their own parallel corpora, without the need of any knowledge of programming or on how to configure the aligner. The alignment results are presented on screen in a tabular form, each row representing an alignment unit, and can be easily edited to correct any alignment errors.

We are at present working on a tool which offers dual concordances from two monolingual comparable corpora of sentences that include the segments that the researcher has marked as equivalents in the multilingual databases of terms, lexical items or phrases. The objective will be to verify if the information is correct and to see if the apparent equivalents do actually function in the same collocational or textual circumstances as the researcher originally supposed.

7. Research and Teaching Applications

The Corpógrafo has been used for a variety of teaching and research applications for some time. Although it was originally designed for use by individuals, it is now possible for groups of people to work on the same area and distinguish the work done by the different contributors. It is available online to whoever asks for a username and password. We use it for teaching purposes and several of our masters' and doctoral dissertations depend on the system for their research. There are also many users from all over the world, particularly from Brazil, who use if for pedagogical and research work.

So far, most of our research has been in the areas of terminology and lexical analysis, and is becoming increasingly sophisticated now that the tools have been improved. However, the new tools allow for much more.

These tools can now be used to search corpora for various forms of multi-word expressions using n-grams, normal lexical concordances and concordancing using the NooJ POS analysis. Parallel texts can be aligned, and data extracted from comparable corpora can be concordanced in two languages simultaneously. The resulting databases can be used to store and categorize lexical, syntactic and textual information that can be exported for a variety of uses.

Apart from the more obvious applications to research projects, there are several ways in which practical results can be obtained for translators and others. For example, in order to facilitate the organization and translation into English - or even the writing of the original in English - of the programmes of our university courses, we are at present using the tools to find and store in our databases, useful phrases in comparable corpora built from texts from English speaking university sites on-line. This is being done using an n-gram tool and/or the NooJ POS analysis using patterns typically associated with the type of text under analysis. The results are being used to create a list of English and Portuguese "useful phrases", available on the university intranet or on a special area of our translator's page at http://web.letras.up.pt/traducao/TRAD/trad.htm The same idea can be applied to a variety of similar uses, and provide useful pedagogical tools for teaching levels of language from lexicography to text analysis.

8. Final Remarks

The Corpógrafo has always been driven by the needs of researchers in linguistics who want to take advantage of user friendly language technology. It is also useful for teachers who want to train their students to understand the possibilities of these technologies without necessarily having to beg their universities for constant upgrades of very expensive commercial translation software with which to do so.

In other words, we have always tried to foresee a use for the tools rather than simply create tools that may or may not be wanted. The tools themselves are not a novelty, but the combination and integration of several tools into one integrated system is less usual.

We must emphasize the fact that the tools have been conceived to encourage the general linguist to use and understand the possibilities of NLP tools. This means that the tools should provide the general linguist with the possibility of collecting, observing and validating data and inserting it into the Corpografo in their personal area. The results can then be used to integrate information in the Corpógrafo tools as, for example, when lists of expressions to retrieve definitions and semantic relations were retrieved for terminology processing.

The latest developments will allow us to create lists of discourse markers, lexical bundles and other linguistic phenomena that can be used in both monolingual and multilingual comparable corpora. The work-in-progress is at the level of research and individual project work being done by post-graduates in translation, terminology and general or contrastive linguistics.

Acknowledgements

This work was done in the scope of the Linguateca, contract nº339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union. We should also like to thank the colleagues and students who have contributed with ideas and opinions on how best to develop new tools.

References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa & R. Silva (eds.), *Proceedings of LREC 2004* (Lisboa, Portugal, 26-28 May 2004), pp. 1313-1316.

Maia, B. (1997). Do-it-yourself corpora ... with a little bit of help from your friends! in B. Lewandowska-Tomaszczyk and P. J. Melia (eds) *PALC '97 Practical Applications in Language Corpora.* Lodz: Lodz University Press. 403-410.

Maia, B., Sarmento, L., Santos, D., Cabral, L., & Pinto, A.S. (2005). CORPÓGRAFO - an online suite of tools for the construction and analysis of corpora, semi-automatic extraction of terminology and the construction of conceptual databases. *Proceedings from the Corpus Linguistics 2005 Conference Series* (Birmingham, UK, 14-17 July 2005), s/pp.

Maia, B. Silva, R., Barreiro, A., & Frois, C. (forthcoming). N-grams in search of theories, in B. Lewandowska-Tomaszczyk *PALC 2007 Practical Applications in Language Corpora.*

Oliveira, D., Sarmento, L., Maia, B., & Santos, D. (2005). Corpus analysis for indexing: when corpus-based terminology makes a difference. In P. Danielsson & M. Wagenmakers (eds.), *Proceedings from the Corpus Linguistics 2005 Conference Series* (Birmingham, UK, 14-17 July 2005), s/pp.

Sarmento, L., Maia, B., & Santos, D. (2004). The Corpógrafo - a Web-based environment for corpora research. In M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa & R. Silva (eds.), *Proceedings of LREC 2004* (Lisboa, Portugal, 26-28 May 2004), pp. 449-452.

Sarmento, L., Maia, B., Santos, D., Pinto, A. & Cabral, L. (2006). "Corpógrafo V3: From Terminological Aid to Semi-automatic Knowledge Engine". In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik & D. Tapias (eds.), Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006) (Genoa, Italy, 22-28 de Maio de 2006), pp. 1502-1505.

Silva, R. (2006). *Performance and Individual Act Out: The Semantics of (Re)Building and (De)Constructing in Contemporary Artistic Discourse*. Master's dissertation. Porto: FLUP.

Varantola, K. (2003). Translators and disposable corpora. in Zanettin, F., S. Bernardini and D. Stewart (eds). *Corpora in translator education*. Manchester: St Jerome. 55-70.