The Corpógrafo – a Web-based environment for corpora research

Luís Sarmento
Linguateca, Porto, FLUP
Via Panorâmica, s/n, 4150-564
Porto, Portugal
las@letras.up.pt

Belinda Maia
Fac. de Letras de Univ. do Porto
Via Panorâmica, s/n, 4150-564
Porto, Portugal
bmaia@mail.telepac.pt

Diana Santos
Linguateca, Oslo, SINTEF ICT
Pb 124 Blindern, 0314
Oslo, Norway
Diana.Santos@sintef.no

Abstract

In this paper we present the Corpógrafo, an integrated web-based environment for corpus linguistics and knowledge engineering that is being developed at the Porto node of Linguateca. The Corpógrafo aims to provide an integrated corpora research environment by making freely available on the web a comprehensive set of text and language tools (http://www.linguateca.pt/corpografo/). We describe the current stage of development of the Corpógrafo, discuss its current limitations and propose possible developments.

Introduction

The use of corpora in linguistics, natural language processing, translation and teaching has been progressively gaining acceptance since the 90's. Large general language or newspaper corpora (100 million words or more) have been publicly available for years, but recently attention is being increasingly focused on small special domain corpora. Small corpora are usually tailored with a specific task in mind: translating a document, building a domain glossary, or studying a particular linguistic phenomenon. Most of the time these corpora are only used during the time taken to execute the task and are disposed of afterwards. Therefore, it is impossible to provide users with ready-made corpora: users have to build their own corpora to fulfill their needs ("Do it yourself Corpora", Maia, 1997). However, most users lack the technical expertise and software tools to go through the whole process of building and doing research on corpora by themselves. Although some people have been creating personal corpora, there is no systematic framework for the collection and registration of texts in a way that allows for the re-use of these resources.

This is the rationale behind Corpógrafo, which has been developed by Linguateca's Porto node, at Faculdade de Letras Universidade do Porto (FLUP). Linguateca is a distributed resource center for Portuguese whose main aim is improve and foster R&D in the processing of Portuguese language. In the corpora domain, it has made available through the Web a large number of Portuguese (syntactically annotated) corpora (the AC/DC project, Santos & Bick, 2000), and created other resources from scratch (such as the 200-million words CETEMPúblico, Santos & Rocha, 2001, or the Floresta Sintá(c)tica treebank, Afonso et al., 2002). Neither of these, however, addressed the needs of terminologists, translation teachers or teachers of language for specific purposes. The Porto team has thus started the development of the Corpógrafo, which is an integrated web-based tool for corpus linguistics and knowledge engineering.

Corpógrafo, which is publicly available at http://www.linguateca.pt/corpografo/, is a free-of-charge computational environment on the Web (anyone working with Portuguese can use it, it is enough to register) that allows users to build and research personal corpora without the need for specialized technical skills or additional software, apart from a normal Internet browser. It includes tools that range from text converting utilities to concordance engines and semi-automatic semantic relations extractors. Users are also able to create dedicated databases to store and share their work.

Our development approach has been that of starting by providing simple (sometimes naive) functionalities and then letting Corpógrafo users give us their feedback and explain their needs. We have thus been able to make a deep requirement analysis and understand what users really need or expect of such a tool. Furthermore, the modular architecture of the Corpógrafo enables the easy inclusion of new functionalities or the improvement of current ones. In the following sections we will focus on the design and development details of the Corpógrafo.

The Birth of the Corpógrafo

When Linguateca established a new node in FLUP it soon became clear that local corpora users (researchers, and translation students) had to face too many barriers. Corpora software, when available, was limited to use on Faculty computers. Interesting text documents were usually available in PDF or in other structured file formats that text analysis software would not handle. Converting these files to plain text format was not easy and sometimes users would perform the conversion manually. Moreover, results gathered from corpus analysis would usually be stored in a proprietary format file.

This situation led us to start developing a set of simple web tools to help local users work with corpora. Making the tools available on the web seemed a very convenient option because it avoided all the problems related with installation: all that was required was a common Internet browser. At the same time, web based applications are convenient for the developer because they are easier to update and they also motivate users to send their feedback almost immediately. After developing some isolated tools, we decided to integrate them in a common environment (Maia & Sarmento, 2003) where users would, in the future, be able to perform several of the most frequent tasks related to specific domain corpora:

- **Text collection:** text extraction from structured files (PDF, HTML, MS-Word, PS), downloading of new texts from the Web
- **Text pre-processing:** "cleaning" text, segmentation, text annotation, text encoding searchable or exchangeable format;

- **Corpus search:** regular expression concordances, collocation extraction, frequency-based statistics (N-grams count);
- **Information extraction:** terminology, semantic relations, conceptual maps
- **Knowledge-resource building:** specific-domain glossaries, thesauri, terminological databases and ontologies; categorized word-lists;
- **Comparable corpora studies:** compilation and search in comparable corpora (same domain, genre, language pairs, etc.).
- **Exporting of results to other formats and applications:** to standard terminological databases, translation memories, etc.

This environment was baptized the Corpógrafo.

Development Strategy

The Corpógrafo has some very particular software requirements that make it different from other NLP tools: the Corpógrafo is a tool intended for ordinary users who are not required to have specialized knowledge about computers. Additionally, since the global philosophy of Linguateca is to develop and make available tools that can satisfy users and therefore promote use, we have adopted a user-centered software development strategy. During this first year of development, we have tried to make available new functionalities as soon as possible, even if they are not very sophisticated from an NLP point of view. This has been essential in achieving good software requirements specifications because we have been able to get an almost immediate feedback from users. In this way, users feel engaged in the process of developing the Corpógrafo and constantly share their ideas about new functionalities to be implemented. Even users from fields that are not traditionally associated with corpora, such as sociology, have been interested in making suggestions of how the Corpógrafo could be extended to help them in their research activities.

Such an open attitude towards users' opinions has also created some difficulties. A lot of time has been spent in developing and reformulating user interfaces. Graphical user interfaces are crucial to the success of an application but they consume a very significant percentage of development time. Unfortunately, web interfaces are even more time consuming. However, the global balance is definitely positive. We have developed the application skeleton, which already provides simple functionalities at different levels of corpora usage. We have also developed a set of NLP software modules that allows us to quickly implement new and more complex functionalities, while spending less time on user interfaces.

Current Functionalities

Currently, the Corpógrafo offers a significant set of functionalities, ranging from assisting in the preparation of corpora to the creation of terminological databases.

Text Compilation, Pre-Processing and Indexing

The most basic step in preparing a corpus is text compilation. The Corpógrafo allows users to upload text files in various formats (PDF, HTML, DOC, PS, RTF and plain text) to a private personal area in the web server. Text data is extracted from uploaded files and stored in the user's private area up to a given quota limit (10Mb for standard users).

The Corpographer also allows users to directly download data from the web using three different options. The simplest one allows the user to indicate the URL of a resource to be directly downloaded by the Corpógrafo, which is especially helpful for users with slow dial-up access. The second option enables users to inspect an entire site under a given URL. Users may then choose which texts from that site should be added to their personal area without having to go through the trouble of visiting the entire site. We are also working on a more informed third option, intended to help users in gathering larger amounts of similar texts quickly, that uses the Google search engine to try to find texts similar to other texts already in the user's personal area (something like a "more like this" option). We are planning to achieve this by building search queries enriched with the meta-information about texts, provided by the user, and the terminological entries already stored in the Corpographer, whenever possible. Uploaded texts may be cleaned and segmented in sentences using a special editing interface. The Corpógrafo also allows users to store meta-information about the file in a specific database, which will not only help the organization of texts and corpora but may be important in subsequent searches. We are also considering ways to protect the authors from abuse of copyright, e.g. by enforcing registration of all texts that are to be used for terminology production. Users are thus requested to fill in a form containing fields for:

- **generic file information:** document title, language, year of publication, authorization of other Corpógrafo users to access the file;
- **source information:** authors' names, source organization, URL (if applicable);
- **content categorization:** domain (from a 3 level taxonomy tree), genre (from a pick-list), source medium (from a pick-list), Universal Decimal Classification Code (if available);
- **description:** description provided by the user.

This meta-information may be used or back referenced in subsequent searches. For example, terminology entries extracted will automatically keep the reference of the authors of the text in which they were found.

Building Searchable Corpora

After compiling and preparing the text files uploaded to their private area in the server, users may finally create searchable corpora. The Corpógrafo allows users to choose text files and group them in a *Selection*, which is the actual searchable corpus. *Selections* may be composed of an arbitrary number of text files that, in turn, may also be used in different *Selections*. Users may update *Selections* by adding or removing new files at any moment. Each *Selection* is encoded in two different formats that are suited for searching: XML and IMS-CWB (Christ et al., 1999).

Searching and Analyzing Corpora

The Corpógrafo has several different search procedures available. The most essential one is the concordance search using regular expressions. The Corpógrafo uses the

IMS-CWB system as one of the engines for corpora search. The IMS-CWB system provides a very efficient way of performing a wide variety of searches on corpora, especially those based on regular expression queries. The Corpógrafo allows users to choose one of the Selections previously built and introduce arbitrarily complex regular expression queries to search that Selection. The expression is handed to the IMS-CWB engine and search results are presented to the user. Users may choose how to visualize the result from a set of options (context, grouping and ordering; collocations). Another analysis procedure available through the Corpógrafo is the N-Gram distribution of a corpus. Users may obtain absolute and relative counts about the occurrences of N-grams from 1 to 6 words. Results may then be saved on a text file in CSV (comma-separated values) format that is readable in most spreadsheet software. This enables users to perform subsequent analysis using custom statistical software.

Terminology and Terminological Databases

During the development of the Corpógrafo we felt that it was important to be able to use it as soon as possible to produce palpable linguistic resources. More than just a tool for corpora analysis, the Corpógrafo should also enable users to produce resources that could be shared with other researchers or usable in other applications.

Therefore, we began focusing the development of the Corpógrafo on terminology because this field has an interesting cost-benefit relation. Terminological resources are extremely useful to a wide variety of applications, ranging from human translation to automated document retrieval. They may be reused easily if stored in the appropriate formats. On the other hand, they may be produced without the need for very sophisticated algorithms as long as some human intervention is provided. Semi-automated methods for terminology extraction are reasonably simple to implement and yet effective enough for most purposes. Moreover, users are usually willing to invest some time in validating the results of a semi-automatic or naïve extraction methods because the value of the resource produced is considered to justify the effort.

The Corpógrafo already has a significant set of tools for terminology extraction and management. Terminological extraction is achieved using N-Gram analysis on a corpus after imposing a set of lexical restrictions on possible terminological candidates using an electronic dictionary. For example, for Portuguese we only consider N-Grams beginning and ending by a name or an adjective, and we use the dictionary provided by LabEL (Ranchhod, 2001). Despite using such a naive algorithm, the Corpógrafo is able to produce a satisfactory list of terminological candidates ordered by frequency (in Portuguese or English). Users may inspect concordances of the selected candidates within the corpus to verify if they are valid terminological entries. Valid terminological entries may be sent to a dedicated database.

The Corpógrafo allows users to manage all their terminological work using personal dedicated terminological databases. Users may create as many databases as needed to work in terminology extraction on different knowledge domains. The Corpógrafo terminological databases were developed in order to store information for many possible uses of terminology. Therefore, the database includes a large number of fields that may not be useful for every user. In fact, these databases may be considered as an expansion of the more traditional terminology databases because they also include fields especially related to semantic analysis. At the same time we also follow the ISO standard for Terminological databases (ISO 12620) in order to enable users to develop more traditional terminological work. The Corpógrafo terminological databases may store the following information for each terminological entry (fields marked with * are automatically filled using information associated with files, or directly inferred from corpora):

- **general terminological information:** language*, type of term, administrative state, register, frequency*, linguistic origin, authors, source documents*, definition*, examples in context
- **semantic information:** list of semantically related terms (30 system-defined relations and the additional possibility of defining custom relations)
- **morphological information:** gender*, number*, animacy*, POS of each word.
- **multilingual information:** bilingual synonyms, synonym type, example of translations or bilingual usage.

The information stored in the databases may be used in the automatic production of new linguistic resources or exported to other file formats. The Corpógrafo automatically creates glossaries in HTML using the list of terms, their definitions and the information about related terms. In addition, the Corpógrafo is able to generate thesauri in HTML using the semantic information between terminological entries. The Corpógrafo also exports this information to the CharGer (Delugach, 2003) file format allowing users to visualize the corresponding semantic network with the CharGer application. We are currently trying to develop alternative ways of visualizing these semantic networks directly from the Corpógrafo interface without the need for an external application. This provides a more convenient (alternate) option because it allows users to check terminological entries and semantic relations without needing to install another application, which may not be that easy for some users.

Semantic Relations

The Corpógrafo has a simple mechanism for discovering possible semantic relations between terminological entries. The user may request the Corpógrafo to try to find possible relations between two terms stored in the database. The Corpógrafo will run a series of tests trying to find patterns in the corpus that show evidence of a given semantic relation. For instance, hyponymy could be detected by finding in a Portuguese corpus a pattern like "A é um tipo de B". The Corpógrafo also uses a very similar pattern matching mechanism to detect possible definitions of available terminological entries. Users may validate the extracted semantic information and store it in the database. Again, these naïve methods show interesting results, even taking into account that they require human intervention to validate them. Clearly, further refinements are needed but for the moment they have a satisfactory

performance and provide the information needed for the other Corpógrafo tools presently being developed, namely the editing and search tools for comparable corpora.

Administration Interface

Since we designed the Corpógrafo in order for it to be used by a large number of users, it has become indispensable to have tools to manage all user-related activities, because this tends to consume too much time and usually requires a specialized operator. For this reason, we have developed an administration interface that helps to manage user accounts and to customize several parameters of the Corpógrafo. Using a common browser, the Corpógrafo Administrator is able to create and remove user accounts. It is also possible to change personal information about users and alter settings of their accounts (e.g. disk quota).

The Administrator may also use the Administration Interface to send emails to all users and post news on Corpógrafo's main page. This helps the development team to keep users informed of the changes in the Corpógrafo and, of course, of possible problems detected, as well as of corpora related news and events. The Administration Interface allows changing some of the internal settings of the Corpógrafo. The Administrator is currently able to add, remove and change taxonomic options used to classify text (domain, genre, register, idiom, etc). We intend to make all parameters of the Corpógrafo configurable through the Administration Interface in order to adapt it according to the user's needs.

Concluding remarks

The Corpógrafo is still under development and a lot remains to be done, both in documentation (help, users' manual and more informed descriptions) and in usability. We are currently investing in both fronts. Also, we are aware that the current NLP capabilities are still of a very basic kind and we intend to endow the Corpógrafo with deeper parsing capabilities to improve knowledge extraction. Software engineering issues related to several users working concurrently have to be addressed, and we are considering three different possibilities: transferring part of the load to the Web client; using a cluster of computers to divide the load; or choosing a grid-like solution, sharing the load among several user machines. See Sarmento (2004) for further details.

The grand (and initial) aim of Corpógrafo was to provide an environment to work with comparable corpora, but we had to start by developing its monolingual capabilities. So the obvious developments we are now working on are developing tools for working with comparable bilingual corpora in the Corpógrafo. These would include a tool to allow the user to match two groups of texts in different languages and mark them as comparable. The next step would involve using the information stored in terminological databases (e.g.: bilingual synonyms) to match text segments from previously built comparable corpora. Such an approach could be complemented by other more general techniques such as using information about verb-object co-occurrences, cognate terms and general bilingual dictionary matches. Furthermore, the Corpógrafo should be able to suggest new possible bilingual matches that could be stored in the database after validation. Finally, it would also be interesting to make available through the Corpógrafo some kind of scripting language. The scripting language would enable computer-literate users to build and run small programs, making it possible for them to use current and future capabilities of the Corpógrafo without being limited by the web interface.

Up till now we have been able to obtain a deep requirement analysis of the users' needs and to provide some simple functionalities to address those needs. At the end of February 2004, over a 100 people had access to the Corpógrafo and several others had had temporary access for demonstrations purposes at the CULT BCN conference workshop in Barcelona and at a workshop at the University of Tampere in Finland.

Acknowledgments

The authors wish to thank the Fundação para a Ciência e Tecnologia for the grant POSI/PLP/43931/2001, co-financed by POSI, We are also grateful to Ana Sofia Pinto for her valuable comments and suggestions regarding both the Corpógrafo and this paper.

References

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos (2002). "Floresta sintá(c)tica": a treebank for Portuguese. In M. G. Rodríguez & C.P.S. Araujo (eds.), Proc. of LREC 2002, the 3rd Int. Conf. on Language Resources and Evaluation (pp.1698--1703). ELRA.

Christ, O., Schulze, B. M., Hofmann, A., & Koenig, E. (1999). The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. University of Stuttgart, March 8, 1999 (CQP V2.2).

Delugach H. (2003). CharGer 3.0b, A Conceptual Graph Editor, http://www.cs.uah.edu/~delugach/CharGer/

Maia, B. (2003). Using Corpora for Terminology Extraction: Pedagogical and computational approaches, in B. Lewandowska-Tomasczczyk, (ed) 2003 *PALC 2001* – Practical Applications of Language Corpora. pp. 147-164. Lódz Studies in language, Frankfurt: Peter Lang.

Maia, B. (2003). What are comparable corpora? in Proc. of pre-conference workshop Multilingual Corpora: Linguistic Requirements and Technical perspectives, at Corpus Linguistics 2003, pp. 27-34. Lancaster U.K.

Maia B., Sarmento L. (2003). Gestor de Corpora - Um ambiente Web para Linguística Computacional, presented at CP3A 2003: Corpora Paralelos, Aplicações e Algoritmos Associados.

Ranchhod, E. M., (2001). «O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais». In Ranchhod, Elisabete M. (org.), Tratamento das Línguas por Computador. Uma Introdução à Linguística Computacional e suas Aplicações, pp. 13-47, Lisboa: Caminho

Santos, Diana & Eckhard Bick (2000). Providing Internet access to Portuguese corpora: the AC/DC project. In Maria Gavrilidou et al. (eds.), Proceedings of LREC 2000 (pp.205—210). ELRA

Santos, Diana & Paulo Rocha (2001) Evaluating CETEMPúblico, a free resource for Portuguese. Proceedings of the 39th Annual Meeting of the ACL (pp.442--449). ACL.

Sarmento L. (2004) Relatório Técnico sobre o Corpógrafo, http://poloclup.linguateca.pt/docs/cg/.