The Corpógrafo – an Experiment in Designing a Research and Study Environment for Comparable Corpora Compilation and Terminology Extraction

Belinda Maia & Luís Sarmento
University of Porto & PoloCLUP of Linguateca
Via Panorâmica s/n
4150-564 Porto
Portugal
bmaia@mail.telepac.pt & las@letras.up.pt

Abstract

This paper will describe how restrictions of commercial software led us to develop a suite of on-line tools for special domain corpora compilation and analysis, term extraction and term database management - the Corpógrafo. It has been developed by computer engineers from Linguateca - http://www.linguateca.pt – in cooperation with the Master's in Terminology and Translation at the University of Porto. The overall framework is designed to be flexible and extendible to other research purposes, including lexicography, corpus linguistics and information retrieval. Each research project uses the Corpógrafo tools online in an independent environment on our server.

The corpora collection tools are designed for the collection, categorization and preparation of texts that can then be combined and re-combined to form different corpora. We focus on comparable rather than parallel corpora for reasons that will be explained. Terminology is extracted semi-automatically from raw text, using an n-gram tool that imposes lexical restrictions on both the morphology and the context of possible terms. The user can view the term and related information using concordances before transferring the information to a related terminology database. There are also tools for finding in-context definitions and semantic relations. The results of ongoing research will be demonstrated.

Introduction

The nature of commercial software, its limitations and the restrictions of its use to the university campus led us to find other ways of providing tools for the study of corpora and terminology for our researchers and students. This paper will describe the resulting suite of on-line tools for special domain corpora compilation and analysis, term extraction and term database management – the Corpógrafo.

The Background

The Corpógrafo has been developed by computer engineers working with the Linguateca project (see http://www.linguateca.pt) which is a distributed resource center for Portuguese whose main aim is improve and foster R&D in the processing of Portuguese language. The project is responsible for making a wide variety of NLP tools and both monolingual and parallel corpora freely available online. For example, 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 word CETEMPúblico, Santos & Rocha, 2001, the Floresta Sintá(c)tica treebank, Afonso et al., 2002) and the COMPARA corpus, a parallel corpus in Portuguese and English consisting of about 1 million words in each language.

The Porto node of Linguateca, PoloCLUP, came into existence in October 2002 and has worked since the outset in close cooperation with the teachers and students of a Master's degree in Terminology and Translation at the University of Porto. The two areas that we have explored with the help of Linguateca expertise are the evaluation of machine translation and the construction of corpora and terminology databases. This paper will concentrate on this latter activity, and the existing resources of the Corpógrafo focus on our needs in these areas. However, the overall framework is designed to be flexible and extendible to other research purposes, including general lexicography, corpus linguistics and tools for information retrieval and categorization.

Underlying Assumptions

We work on the principle that the building of corpora and databases should be semi-automatic and that the machine should accelerate human work rather than substitute it. This principle applies both to the collection and categorization of the texts used to build the corpora and terminology entered in the databases as well as the use of this data by the final user. However, every effort is made to design and develop tools that will facilitate the tasks of compilation as much as possible.

Comparable corpora, or original texts of a similar genre or register in the languages involved, are preferred to parallel corpora, largely for the theoretical reason that the language used is more likely to be representative than that of parallel corpora, or originals aligned with translations (Maia, 2002). This has implications for the study of genre and register, but also for terminology usage which, we hope, will be the product of experts rather than translators. There is also the practical reason that comparable corpora are more available than parallel corpora.

The environment in which we work has led us to emphasize domain specific corpora rather than general language corpora, and this objective has revealed the varying demands of different domains, genres and registers. Text harvesting is done largely from the internet, on the assumption that those who publish there will not object to being used as experts, provided we register the details of all the texts and cross-reference any data extracted to the original authors. The resulting corpora are developed for private projects and the interest in special domain corpora tends to be limited to the project for which they are designed. We should like to point out that each researcher, student or working group uses the Corpógrafo tools online and works directly in an independent environment on our server. However, should there be an unexpected demand for a particular corpus, the registration of all the texts will provide the contacts necessary to make it public.

The theoretical approach to terminology research is descriptive and the objective in designing tools for semi-automatic extraction of terminology, definitions and semantic relations from texts for the creation of terminology databases is to allow the description of how the terms are used in context, rather than in an idealized vacuum.

The Origins of the Corpógrafo

When Linguateca established a new node in FLUP it soon became clear that local corpora users (researchers, and translation students) found that the existing tools for making and analyzing corpora, when available, were limited to use on Faculty computers. This severely restricted the research of Master level students who are usually in full employment, who do their research in their leisure time and who sometimes live at some distance from the university.

One factor that deterred people from compiling quality text corpora was that many 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.

These problems led us to start by 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 normal 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 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 in searchable or exchangeable formats;
- **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.

The Process of Corpora Compilation

The corpora compilation tools were designed for the collection, categorization and preparation of texts for analysis. Texts are uploaded from the usual sources of any computer to the individual's working area. It is also possible to download all importable documents from a specific web site, using Corpógrafo's small webcrawler. The webcrawler becomes especially useful when users find domain specific sites containing many possibly interesting documents.

Documents loaded in Corpógrafo are automatically converted to plain text format, although it will be possible to save the original version as well for reference purposes in the new version. Since conversion from other text formats is sometimes problematic, the Corpógrafo has a simple editing tool that may be used to 'clean' the excess material from the extracted text. For most of tasks, however, there is no need to perform any editing after conversion. In fact, for the purpose of terminology extraction, the text obtained directly from the file format conversion is usually enough.

These texts are manually categorized according to domain, sub-domain, genre and register, and the sources and their authors registered so that this information can be automatically transferred to other parts of the Corpógrafo, such as the different corpora and the terminology databases. The texts can then be combined and re-combined by the researcher to form whatever type of corpus is needed. This allows for a variety of analyses, such as the examination of levels of terminology or research into the study of text types and registers.

Until very recently, Corpógrafo stored text in the regular file system. Text file were encoded using the IMS-CWB tools, which also allowed very flexible queries. However, since the whole system is built around a database, used to store terminology and related information, we have decided to also store the entire texts in database tables, which are then searched by the database engine (MySQL). Search performance is still very good and the overall system design became more simple and easy to maintain.

Terminology Extraction

The Corpógrafo allows users to extract terminology from the corpus chosen using a semi-automatic method, which asks for user validation. Extraction is done directly from raw text (stored in the database) by collecting n-grams selectively and imposing lexical restrictions on both the morphology and the context of possible terms. After this first completely automatic step, the user can check the validity of the term and its sources by viewing the related concordance. If a valid term is found, which happens most of the time because the process has a reasonable degree of precision, it can be transferred directly to a dedicated terminology database, together with the references to its source texts.

There are also functions for finding possible in-context definitions and semantic relations and we are working on the possibilities of using the latter to create coherent ontologies, also by using semi-automatic methods. This information, together with the related meta-information can be transferred automatically to the terminology database. The database can also include images and sound files.

Tool Development with Experimental Corpora

Although we are working on larger projects with the cooperation of domain experts, tools are being developed using small but very specific corpora. Medical texts are an excellent source of material for this and a small multi-lingual comparable corpus on Neurons has been used to considerable effect. Although small, it has yielded a satisfying amount of terms, definitions and semantic relations. The domain is clearly circumscribed and the related academic culture seems to have an interest in supplying the right type of pedagogical texts online in several languages. Texts in 5 different languages were collected and used to extract terms, definitions and semantic relations. The terms are easily detected because they are clearly domain restricted, as in *neuron* or *glia* and, since the texts are often pedagogical in nature, definitions and indicators of semantic relationships between terms can be found.

Having discovered that terms are often found in the context of secondary or general terms related to the wider area, the next step was to compile lists of these terms and phrases that would automatically extract the information required. The results were then tested on a corpus of texts on Fibromyalgia. This subject was chosen because a lot is being written about this at the moment and it was possible to download a reasonable sized corpus from the internet in a few days for testing purposes. Further experiments of this kind are proposed using other techniques that will help retrieve the information required.

Ongoing Projects

The reason why the Corpógrafo focuses on terminology is 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, 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.

At present we are working in several special domains, including composite materials in mechanical engineering, population geography and natural hazards, genetics and neurons. The compilation of corpora for special domains has shown us that, whereas there are certain general rules to follow, each domain offers particular challenges.
We have learnt that every project should go through an initial phase of finding pedagogical texts. Encyclopaedia articles, introductory textbooks and similar material are an essential starting point. Such texts, apart from providing the terminologist, or non-specialist, with an introduction to the area, usually provide the key terminology, good definitions and clues to the semantic relationships between terms. Once this has been done, one can move on to more sophisticated academic texts, with state-of-the-art scientific articles written by specialists for specialists being the type of text from which one may retrieve the latest developments.

Each domain, however, has its peculiarities and challenges. The Population Geography project at FLUP started well before the Corpógrafo was functioning and, although every effort was made to use a corpus and a database, the process was hampered by the academic vision of the need for a prescriptive printed dictionary as the overall result. The texts collected were taken from the internet and many of them came from official sources like the European Commission. Since several of those who collected the texts were translators, who were influenced by the possibilities of translation memories and other software, there are several parallel texts. However, the type of text chosen was not particularly helpful. Many of the texts were of a legal genre that is somewhat repetitive and short on explanations. The domain itself relies heavily on general language for its terminology and the extraction of terms needs to be closely accompanied by an expert. However, the lessons learnt from the exercise were considerable.

We have now embarked on a new project with the Geography Department in the area of Natural Hazards with a better idea of what constitutes the 'right' kind of text for the corpora, and the aim is an online terminology database rather than a dictionary. We have learnt that the text collection and basic terminology extraction phases can be more productive if done by domain experts – or at least trainees in the area. This leaves the management and fine tuning of the corpora and database to the terminologists.

It has never been difficult to persuade engineers of the need for descriptive terminology, the possibilities of special domain corpora and the relevance of dynamic databases, rather than static glossaries or dictionaries. Our main problem here is persuading Portuguese engineers to write in Portuguese, rather than English, and to provide us with reliable texts in their own language. So far, the work has been restricted to small areas, like GPS – Geographical Positioning System, Alternative Energies and, more specifically, Wind Energy. However, we are planning a much bigger project in the area of Composite Materials with the full collaboration of the Department of Mechanical Engineering and their research units.

Distributed Building of Terminology Resources

More and more users from various domains are using the Corpógrafo or have stated their interest in using it, since they are realizing the importance of terminology for their everyday life. For instance, the academic community in general has now acknowledged the fact that good terminological resources (especially multi-lingual ones) are very useful from a pedagogical point of view, and also for producing higher quality technical documentation.

This increasing interest has led us to consider the possibilities of expanding the Corpógrafo to a distributed scenario, where multiple Corpógrafo systems are installed in several institutions to address their specific terminological needs. In fact, as new and more complex features are added to the Corpógrafo (such as some heavy distributional statistic measures currently being tested) we believe that evolving to a distributed scenario will be the most natural way to deal with the computational bottleneck that will surely be created. The distributed scenario involves installing the Corpógrafo in several organizations, which may then work on a specific knowledge domain on their own, according to their main competences. Individual contributions may be collected in a central database available to a larger group of users (for instance a University).

Additionally, international institutions may also establish specific collaborations in order to be able to produce multi-lingual resources for a given knowledge domain, by merging and aligning their individual results. We are currently open to establishing such partnerships for future projects.
During the last year the Corpógrafo has evolved from a highly machine-dependent system that required some manual-tuning, to an easy-to-install and stable platform. We are hoping to release the third major version of the Corpógrafo in the near future. It will have more flexible project management capabilities as well as the possibility of exporting terminological resources to a centralized public database searchable via web interfaces.

Conclusions

The Corpógrafo is the result of the implementation of ideas that came from the three leading members in the project with their combined interests in terminology, corpora, natural language processing, information retrieval and artificial intelligence. The development of the various tools, the usability of everything from the structure and tools to the interfaces has benefited from the feedback of the growing number of users.

The on-line nature of the work done has resulted in people from several countries being able to participate and try out the tools and we have about 60 regular users at present. A new, and more sophisticated, version is in preparation and this version will be implemented in teaching at other universities, such as the University of Pompeu Fabra in Barcelona next year. We also expect to use it to start terminology projects in other departments and faculties at our own university and create on-line virtual research projects with those who are interested in developing multi-lingual terminology and special domain text resources.

It has not always been easy to persuade the more conservative staff and students in the humanities dominated world of translation of the possibilities offered by involvement in information technology related research in the areas of translation, linguistics and related subjects in which we teach. However, our experience now is that students at both an undergraduate and graduate level are becoming increasingly curious about what we are doing, and even the teachers are no longer as dismissive as they used to be.

Acknowledgements

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 Luis Miguel Cabral for his valuable contribution to the construction of the Corpógrafo and to Ana Sofia Pinto and Débora Oliveira for their contributions to the project.

Bibliography

Afonso, S., E.Bick, R. Haber & D. Santos (2002). "Floresta sintá(c)tica": a treebank for Portuguese. In M. G. Rodríguez & C.P.S. Araujo (eds.), Proceedings of LREC 2002, the Third International Conference 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).

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

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.

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

Sarmento, L., Maia, B. & Santos, D. *"The Corpógrafo - a Web-based environment for corpora research"*. In *Proceedings of LREC 2004*. Lisboa, Portugal, 25 May 2004.

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