Corpógrafo – Applications

Belinda Maia & Luís Sarmento
PoloCLUP - Linguateca
Universidade do Porto
Faculdade de Letras
Via Panorâmica s/n
Portugal
E-mail: bmaia@mail.telepac.pt & las@letras.up.pt

Abstract

This paper will discuss how the Corpógrafo, a suite of on-line tools created by PoloCLUP of the Linguateca project (http://www.linguateca.pt) for the construction and analysis of corpora and the building of terminological databases, has been used for training professional linguists in corpora compilation, terminology extraction, terminology management and information retrieval. Reference will be made to the research which contributed to the development of the different tools that combine to make the suite usable, and examples will be given of the work possible using both the general language analysis tools and terminology and related data extraction tools.

1. Corpógrafo

The Corpógrafo is an on-line suite of tools for the creation and analysis of personal corpora and the creation of terminological databases that can be found at http://www.linguateca.pt/Corpografo. Although it was designed primarily for the study of terminology, translation and information retrieval, it also provides tools for the more general study of language. An individual or team may do their research independently in their own space on-line using these tools. The Corpógrafo tools are freely available on-line and anyone can sign in and start a personal or group project. Users receive access to the tools and a tutorial, but have to create the content of texts, corpora and databases.

The ideas for the Corpógrafo originated from a pedagogical idea in which special domain mini-corpora were created for the effect of teaching appreciation of text genre and register and the extraction of terminology (Maia, 1997). The creation of a branch of Linguateca, a project devoted to the Natural Language Processing of Portuguese, at the University of Porto led to the creation and implementation of technological tools to speed up this process, create integrated terminology databases, and permit the semi- automatic extraction of terms, definitions and semantic relations. The prototype was first described as the GC (Maia & Sarmento, 2003) and is now known as the Corpógrafo, now in its third version, (see Sarmento et al, 2006). We have always worked on the conviction that computer engineers and linguists have to work in harmony and that semi-automatic procedures, in which the computer programme accelerates the work of the human linguist is more satisfactory than either fully automatic or conventional human methods.

At present the Corpógrafo offers the following functions:

- Gestor (File Manager): the area where each individual or group can upload texts to the server, convert text formats like .doc, .html, .pdf, .ps, and .rtf texts to .txt, edit the texts, check for tokenization, chunk the text into sentences, register metadata on the text, and group texts into corpora.
- Pesquisa (Search): an area that allows for general corpus analysis, with tools for producing wordlists, n-grams and statistics, and studying words or phrases with sentence and KWIC concordancing which allows for sorting according to word position, as well as collocations and other phenomena.
- Centro de Conhecimento (Knowledge Centre): the area where terminology databases can be created and then linked to the corpora from which terms, definitions and semantic relations can be semi-automatically extracted. Term candidates are extracted automatically using an n-gram tool with filters to extract noun phrases from raw text. The terminologist then observes the list of term candidates, checks the term against the context of the underlying concordanced sentences, and clicks the term into the database. Each term automatically takes with it all the meta-data on the texts and corpora in which it appears if it has been previously registered.
- Centro de Comunicação (Documentation): the area where you can find a tutorial and news about the Corpógrafo as well as presentations and publications our group has produced.

2. Pedagogical applications

In the more specific environment of training at academic institutions, Corpógrafo has important pedagogical implications. Perhaps one of the most useful lessons students learn from all the technology we use in the using, making and analyzing of corpora is what they learn about the value of a corpus as a resource of information. They start by learning how to use large monolingual corpora like the British National Corpus and the Portuguese Linguateca corpus, CETEMPúblico, or a parallel corpus like the Linguateca Portuguese/English COMPARA, and they soon become enthusiastic about the advantages of corpora for providing solutions on usage and collocation that dictionaries do not offer.

Once the initial corpus linguistics methodology has been learnt, it is not difficult to build on this and encourage the compilation of corpora for a variety of uses, including terminology work. The various pedagogical exercises that are possible using Corpógrafo are very useful training as

they give a more rounded view of the theory underlying the commercial translation software, with translation memories, associated term databases and other tools, that they will use in the future as professional translators.

2.1 General corpus compilation and analysis

The exercise of constructing a corpus of any kind is important as a means of teaching students how to apply theories of genre and register in practice. Large, general purpose corpora are very useful for a wide variety of applications and research, particularly general lexicographical and language analysis. However, one often needs to work with small specialized corpora in order to study more specific aspects of different language varieties and lexicons. For this one needs to collect the texts in digital form and have access to concordancers and other language analysis tools. The Corpógrafo started out as a way of simplifying this process for the individual researcher.

Although the Corpógrafo has been developed primarily for work with special domain corpora, as we shall describe below, it is also possible to use it for other tasks, such as studying a specific author or genre. In these circumstances, the tools and methodology are those of normal corpus linguistic research.

The more general language analysis tools offered by the Corpógrafo effectively allow anyone to build their own corpus for their own personal project work, and we encourage people to do this and inform us of new ideas for improvement of this area. The work carried out under our supervision includes small individual projects in contrastive and corpus linguistics that have been varied and interesting. The Corpógrafo is often used for analysing specific lexical items or syntactic structures.

A typical piece of project work will take a lexical item that is difficult to translate, either due to its polysemous nature, with words such as *get*, *look*, and *issue*, or because they are closed system items like the adverbs *indeed*, *too*, and *just*, which rarely translate easily, or because they belong to lexical sets that do not easily find direct synonyms in the target language, such as the group *beautiful*, *handsome*, *pretty* and *good-looking*. The behaviour of these words are observed in monolingual and parallel corpora and small 'corpora' can be constructed out of the concordanced examples from these larger corpora for more minute and flexible analysis using the general language analysis tools in Corpógrafo. Similar work has been done with lexical bundles such as *I know that*, *I wonder if*, or any of the many examples in Biber et al (1999) as well as syntactic structures such as complex noun phrases or examples of the use of tense and aspect. The pedagogical objective of this type of work is to raise students' awareness of translation problems at a micro-linguistic level.

2.2 Corpora for terminology work

Most students in applied language or translation related courses come from a traditional language learning environment, and do not always find it easy to understand special domain texts. They tend to call the terminology 'jargon' and to consider the texts themselves boring. Restrictions on time usually mean that the translation teaching programme provides variety rather than subject depth, and 'terminology' is often little more than a short list of difficult words. As future translators, they sometimes ask why, when we can retrieve almost any information we need off the Internet, one should undertake the labour of building corpora for the extraction of terminology.

Clearly, in the everyday world of a professional translator, building corpora and terminology databases is apparently a luxury. However, in order to produce reliable terminology one needs good sources from which to extract information and, although the Internet contains a lot of good information, it also provides us with a good deal of rubbish. One of the objectives of the corpus and terminology building exercise is to teach the value of searching for and recognizing quality resources. As professional providers of language services now understand, proper investment of time and effort in reliable terminology means better quality control and results in the longer term.

Building special domain corpora with a view to extracting terminology encourages students to explore the domain in a certain depth and, in our experience, as the information becomes knowledge, curiosity to know more about the subject takes over. This type of exercise brings them closer to professional translation because it forces the student to become more familiar with the subject matter than is normal in most translation teaching. The exercise of choosing texts and analysing them in terms of genre and register is also useful for teaching them to find and imitate appropriate models in their own text writing or translation. They also learn to assess texts for their lexical quality and density, and consequent appropriateness for terminology extraction.

We recommend that beginners in the special domain start with encyclopaedia articles and then move on to pedagogical introductions to the subject, before including more complex texts like master's and doctoral dissertations, which usually include plenty of definitions and other relevant information. As the terminology database grows, keywords can be used to search for further appropriate texts, gradually leading to peer-to-peer publications for the extraction of more sophisticated or 'state-of-the-art' new terminology in the domain.

2.3 Extraction of Terminology, Definitions and Semantic Relations

Although there will always be a need for standardized terminology, for legal and simple administrative reasons, the emphasis in now on describing which terms are actually used in different contexts, as well on detecting the appearance of neologisms and/or mutation of terms. This information is essential for domain experts, translators and others who work with monolingual and multi-lingual documentation. The fast evolution of most technical and scientific knowledge makes it necessary to create more dynamic resources to cope with this phenomenon, and paper-based dictionaries and glossaries have given way to the terminology database.

It is very important to choose the texts for analysis by the Corpógrafo tools carefully. They will not find what is not in the texts that compose the corpus. However, once one

has a good corpus, the tools in the Centro de Conhecimento are of particular interest. The term extraction tool allows for n-grams to be filtered according to restrictions on the lexical items that can appear in proximity to possible term candidates. This tool functions for Portuguese, English, French, Italian, Spanish, and German, and we are working with the University Pompeu Fabra in Barcelona on Catalan. Although it produces a certain amount of noise, the recall is good and the human terminologist can select good term candidates and reject unacceptable ones very quickly, before submitting the results to the appreciation of the domain specialist for confirmation. The human labour of term extraction that could take months can thus be reduced to a few days.

The tools for extracting definitions and semantic relations depend on a bank of lexical patterns that is under constant development. The underlying theoretical approach is that of Pearson's (1998) 'terms in context', Partington's (1998) and Hunston & Francis's (1999) 'patterns', Biber et al's (1999) 'lexical bundles', and Hoey's (2005) 'lexical priming'. In practice, the task of finding the lexical patterns depends on combining computational expertise with human observation and analysis.

The terminology databases are conceived as essentially multilingual. This allows for terms to be extracted from the corpora in different languages and then linked within the database. The main database fields are typical of those used in terminology, but the pick-lists within them can be modified as and when the occasion arises. For example, the domain and sub-domain fields offered reflect the areas we are working on, but they can be added to on request. Also, although the more classical semantic relations are already part of the programme, researchers are encouraged to create their own as well. Experience has shown us that each domain reveals different types of semantic relation, as Sager (1990:30) demonstrates.

Once the more basic terminology has been extracted, it can be used to discover more specialized texts on the Internet. One can use a function that indicates the co-occurrence of terms in the different texts in the corpus, and this allows for further relevant texts to be found using normal Google-type searches. The tools are being improved on an on-going basis in order to provide further possibilities of extracting and structuring domain knowledge, creating further corpora and providing tools for more general information retrieval.

3. Research Applications

The Corpógrafo is being used for a variety of projects, many being prepared by people we do not even know. Here we shall concentrate on showing how the users have cooperated with us in its development, and refer to some of the projects with which it is being used.

The development of the Corpógrafo has resulted from working from an overall concept to the small details that make it workable. The process of trial-and-error that produced it is possibly as relevant to research methodology as the results themselves. Computer scientists and computational linguists clearly had a leading role, but the need to cooperate with general linguists, terminologists and translators forced them to contemplate the human + machine cooperation aspect. This attempt to create genuine understanding between two research groups which do not always work easily together was fundamental to the way the Corpógrafo developed and resulted in the coordination of the various tools and the user-friendly interfaces.

Much of the work done so far with the Corpógrafo has been experimental and has led to further improvement of the tools. The more general language work done in courses in contrastive and corpus orientated linguistics led to the way the concordancing tools developed, while the terminology work within master's degree projects was essential to the development of the terminology database. The compilation of the banks of lexical patterns and semantic relations has been carried out by research assistants and within the scope of masters' dissertations.

3.1 Terminology projects

The Corpógrafo has been very largely developed to deal with terminology projects, and version 3 now permits these to be carried out successfully, with the resulting databases being exportable in .xml for formatting in other programmes. We hope soon to develop tools for exporting the terminology data to a format that can be consulted on-line. It must be remembered that the existing system only allows consultation of the corpora, terms and other data by individual researchers, or by those individuals they authorise to consult their work.

For demonstration purposes there is a small project which is described in Portuguese on the site under the title 'Neurodemo'. This project started out as two small comparable corpora in English and Portuguese of about 25,000 words each on the subject of neurons created by an undergraduate student for a term paper. It now has comparable corpora on the same subject in five other languages, all of which have been used for the extraction of terms, definitions and semantic relations. The texts come largely from on-line popular science texts explaining neurons and have proved exceptionally useful for searching for information in a small well-defined area. The instructional nature of the texts provides the terms, as well as useful definitions and contexts from which the semantic relations between terms can easily be deduced by the human observer. The small size of the corpora in relation to their comparative success is proof that a well-selected corpus of texts is often more useful than a loosely constructed large corpus of only partially relevant texts.

There are several other on-going terminology projects that are not yet ready for publication, but we hope that they will be available in the near future.

3.2 Research for the improvement of the Corpógrafo tools

Most of the research done so far at dissertation level has involved the production and analysis of corpora, and the extraction of terms and other data as a method of testing and developing the Corpógrafo rather than for the production of full-scale databases. For example, the corpus analysis area has already proved useful for studying the instability of terminology in the fast developing area of GPS – Geographical Positioning

System (Brito, 2005). This is a study of concepts and how they are represented by several different terms, depending on who is using them and where. The Corpógrafo was used for the creation and observation of the corpus used, although our version of the terminology database at the time was not yet ready for the developments registered, and the terminology was created in another system. There is also a study of how concepts and their related terms have developed over decades in the field of Genetics and, although a much larger project is planned, (Fróis et al, forthcoming) shows how the concept behind one term has evolved seventy years, and how the expansion and subdivision of meaning within the concept has led to changes in usage of the original term and its expansion into various terms by the addition of adjectives to the original noun. These studies show how knowledge evolves and how terminology sometimes struggles to keep up with the pace of development and with the shifting concepts involved. They also show how a diachronic corpus can often be useful in explaining apparent inconsistencies in the evolution of the terminology of a certain area, and how different participants in the process contribute to the proliferation and confusion of terms.

Other work at dissertation level has also tested and provided incentive for further developments. One dissertation, by Almeida (to be defended) involves the testing and development of definition patterns in the domain of Natural Hazards. Having extracted definitions from corpora using the general language analysis concordancing function in the Corpógrafo, based on the ideas of Pearson (1998) and others, she tested her results against those obtained later using the bank of lexical patterns being built to support the Corpógrafo's definition extraction tool. Another dissertation by Jesus (to be defended) involves the building of networks of semantic relations in the area of Seismology. The resulting database should prove very useful in the development of the tool we are at present designing for the visualization of semantic networks. However, we cannot divulge further details until these dissertations have been defended.

4. Conclusions

The Corpógrafo is freely available online, which may partly account for its popularity. Sarmento et al (2006) supplies more details on who is using it and for what purposes. The original objective of producing pedagogical tools has been successful, with the users participating in the brainstorming over the development of the original tools, testing them and providing further ideas for improvement. Although we would not claim that the work done during this development has proved easy or perfect, we hope that the resulting Version 3 will soon prove its worth as a tool for more professional situations of terminology retrieval and management. However, the state-of-the-art of tools and resources in this area is moving fast and we recognize the need to refine the existing Corpógrafo and add to its potentialities in the future.

Acknowledgements

We should like to thank Linguateca, a distributed language resource center for Portuguese, for the opportunities offered to develop all the tools that are described here, and, more specifically, Diana Santos, Luís Cabral, and Ana Sofia Pinto, for all the work that has gone into their production. We should also like to thank the researchers who work with us for their ideas and for the research referred to here.

References

Almeida, A.S. (to be defended). *Pesquisa de Informação Terminológica: dos Marcadores Lexicais Aos Padrões Suporte: um Estudo no Dominio dos Riscos Naturais – Cheias*, Master's dissertation, Universidade do Porto.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Ltd.

Brito, M. (2005). *"Um conceito = um termo?" – Multiplicidade na relação conceito-termo numa Base de Dados Terminológica de orientação conceptual no domínio da terminologia do GPS.* Universidade do Porto: Master's dissertation.

Fróis, C., B. Maia & A. Videira (forthcoming). 'A Case of Meaning Extension', in the *Proceedings of PALC 2005 – Practical Applications in Language and Computers*, University of Lódz, April, 2005.

Hoey, M. (2005). *Lexical Priming,* London/New York: Routledge.

Hunston, S., G. Francis (1999). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English (Studies in Corpus Linguistics).* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Jesus, C. de (to be defended) *Terminologia e Representação do Conhecimento do Domínio Específico da Geodinâmica Interna: Uma Abordagem ao Subdomínio da Actividade Tectónica*. Master's dissertation, Universidade do Porto.

Maia, B. & L. Sarmento (2003). 'GC - An integrated Environment for Corpus Linguistics'. Poster at CL2003: CORPUS LINGUISTICS 2003 - Lancaster University (UK).

Maia, B. (1997). 'Do-it-yourself corpora ... with a little bit of help from your friends'. In Lewandowska-Tomaszczyk, B. & P.J. Melia, (eds.) *PALC'97: practical applications in language corpora* (pp. 403-410), Lodz. Lodz University Press.

Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching (Studies in Corpus Linguistics.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Pearson, J. (1998). *Terms in Context*. Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Sarmento, L., B. Maia, D. Santos, L. Cabral, A. Pinto. (2006). "Corpógrafo V3 - From Terminological Aid to Semi-automatic Knowledge Engineering", in Proceedings of LREC 2006.