Corpógrafo V3

From Terminological Aid to Semi-automatic Knowledge Engineering

Luís Sarmento[12], Belinda Maia[12], Diana Santos[1], Ana Pinto[12] and Luís Cabral[1]
[1]Linguateca, [2]Universidade do Porto

las@letras.up.pt, bmaia@mail.telepac.pt, Diana.Santos@sintef.no, asofia@letras.up.pt, Luis.M.Cabral@sintef.no

Abstract

In this paper we will present Corpógrafo, a mature web-based environment for working with corpora, for terminology extraction, and for ontology development. We will explain Corpógrafo's workflow and describe the most important information extraction methods used, namely its term extraction, and definition / semantic relations identification procedures. We will describe current Corpógrafo users and present a brief overview of the XML format currently used to export terminology databases. Finally, we present future improvements for this tool.

1. Introduction

Corpógrafo, www.linguateca.pt/Corpografo/, now at version 3, is a project that has been evolving from a simple on-line word-concordance tool to a knowledge engineering platform. During the last three years we have followed a bottom-up approach in close connection with the users (Sarmento et al., 2004). This strategy has allowed us to develop tools that, despite their technical simplicity, address people's needs and has lead to the emergence of a technically-aware audience, now asking for more complex tools for larger projects.

Corpógrafo was initially built to help users study small specific domain corpora instead of the general scope and very large corpora that are available for most languages (see the "do-it-yourself" corpora proposed by (Maia 1997). These small technical corpora are very useful for translation purposes, both for the examples of technical language usage they contain, and for their terminological density. Given the relatively small size of these corpora, we were able to develop a centralized web environment that allowed users to upload their own texts and work on them with simple online linguistic analysis tools (concordancers and n-gram lists).

Such a web platform was, at the time, original (Maia & Sarmento, 2003) because the online tools available only allowed users to work with a set of previously collected and fixed corpora. Corpógrafo has since then evolved from a simple word-concordancing tool to a more complex platform capable of supporting the development of ontologies from text.

2. Corpógrafo's Workflow

The most significant feature of Corpógrafo is its integrated workflow. Corpógrafo provides non-technical users with a complete environment for their work from corpora preparation to terminological database management. A great deal of effort has been invested in developing an easy to use, coherent interface, to allow users with no programming skills whatsoever to develop valuable resources using Corpógrafo.

Corpógrafo users are given personal accounts in a web server, where they may upload their own very specialized (and often private) corpora and process them to produce language resources for their own use. The development workflow in Corpógrafo includes the following steps (Figure 1):

1. Text acquisition and pre-processing: Corpógrafo allows users to upload into their private accounts text files in several formats, or crawl specific web sites for text. Text is extracted from HTML, MS Word, PDF, Postscript and RTF file formats. Users are also provided with a simple text editor for "cleaning" the text whenever necessary.
2. Corpus management: users may create searchable corpora by combining texts in their account in a variety of ways. The same text may belong to several searchable corpora, allowing users to build corpora for very specific purposes.
3. Creation and management of terminology databases: users may create, edit, delete, and export databases for storing terminological information. These databases are private.
4. Terminology extraction: Users may extract terminology from text using semi-automatic terminological extraction capabilities which work robustly in five languages: Portuguese, English, Spanish, French and Italian. This information may then be stored in the databases managed by the user.
5. Definition extraction and semantic relation discovery: after extracting terms, the user may try to find their definitions in corpora. The user may also try to find semantic relations between the terms. All information obtained can be stored in the database.
6. Exporting Results: corpora and terminological databases may be exported in XML to be used in other applications.

Corpógrafo provides a complete resource-building environment that starts from raw text and is able to produce structured knowledge in a step-by-step process. Because it does not require programming skills, Corpógrafo allows terminologists and domain experts to work together autonomously in small teams. For example, we have partners within our University that are creating terminological resources in specific domains combining terminologists with experts from the fields of "telecommunications" or "composite materials", neither of them needing to program Corpógrafo or knowing low-level technical details about it. This is a first step to promoting the emergence of domain-sensitive NLP applications, such as specialized search engines and document classification applications (Gulla et al, 2004;

Cimiano & Völker, 2005).

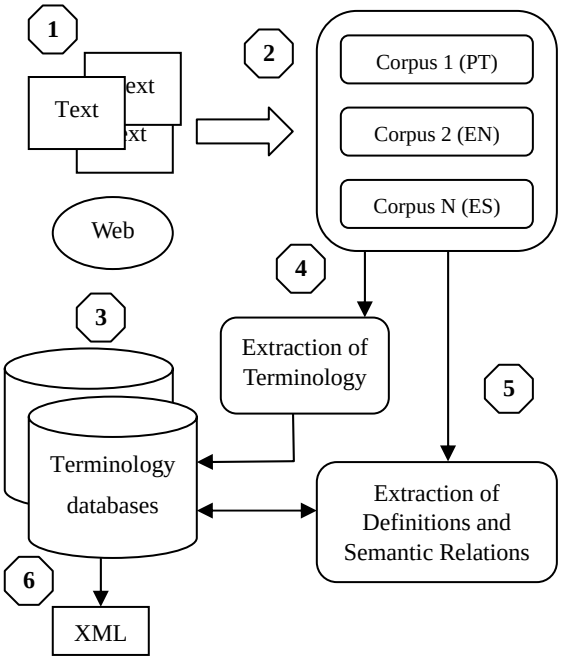

Figure 1: From text files to structured knowledge

3. Focus on Semi-automatic Methods

Although many researchers have not given hope on achieving fully automated methods for various knowledge extraction tasks, our research within Corpógrafo is evolving in a different direction: our exclusive aim is to develop semi-automatic methods to widen the practical extraction capabilities and Corpógrafo's fields of application. We have thus developed a set of simple semi-automatic methods for terminology extraction, finding definitions and for identifying possible semantic relations between terms. These methods require a small effort of user validation, but are still many times faster than the corresponding manual process of building terminological resources.

3.1 Extracting Terms

Our term extraction algorithm was designed to perform robustly on all possible technical domains, since we cannot know in advance which domains of knowledge our users will search. However, term morphology and syntactical structure vary according to the specificity of the terminology and different technical domains. For example, Sager (Sager, 1990:p80) states that "It is now realized that term formation is and can be influenced according to the subject area in which it occurs, the nature of the people involved and the origin of the stimulus for term formation'. Therefore rules and heuristics that are applied in a particular domain (such as chemical compounds) may not be transposable to others (e.g. GSM communications). It is also obvious that different languages have different syntactic means for creating noun phrases.

In order to obtain a term extractor that could be used in several domains, we based our term extraction algorithm in the empirical observation that it is easier to describe what a valid term candidate is not, than what it can be. Since it is agreed that terms are usually noun phrases, our algorithm (Sarmento, 2005) begins by compiling possible candidate noun phrases from text, and then excludes all noun phrases that cannot obviously be terms, because they are "ill-formed". The resulting (filtered) list of noun phrases is then presented to the user for validation, sorted by frequency. For identifying possible noun-phrases, the algorithm does not rely on POS tagging: it simply looks for maximum length word sequences that are preceded by articles, numeral and prepositions and that do not contain a list of previously compiled stop words (function words, certain very frequent verbs, punctuation, etc). Despite its simplicity, this algorithm has been effective in allowing users to compile long lists of terms quite quickly, and then store them in users' databases.

An additional advantage of this algorithm is that it is easy to port to other languages, just by changing the lists of the obligatory preceding words and stop words, making Corpógrafo attractive for users interested in developing multilingual terminological databases.

3.2 Extracting Definitions

For each term stored in the database, Corpógrafo allows users to search for possible definitions in the specialized corpora they collect. This is achieved by searching archetypical definition patterns at sentence level. For example, considering that the user is searching for the definition of term "TERM", Corpógrafo tries to find sentences which match patterns like "TERM (is|are) * that...", "We understand by TERM a * that...", "...* are (named|known as) TERM.", and many others. Our experience shows that these patterns do occur in technical texts often enough to make this an effective method for obtaining candidate definitions, which have then to be validated before storage in the database. Our users have reported that, compared with manual inspection, this method is capable of speeding up the task of finding definitions by several orders of magnitude. Currently, Corpógrafo uses a pattern base of nearly 135 of these patterns for Portuguese, about 120 for English, and a few dozens for other languages such as Spanish, Italian and French, and we plan to make it possible to have users contribute to (their own, text specific) definition patterns.

3.3 Finding Relations

Corpógrafo is also intended to help the user finding semantic relations between terms, using a process similar to the one described for definitions, but extending patterns to obtain evidence of possible relations such as hyperonymy, hyponymy, holonymy, meronymy, class-of, instance-of, and several functional relations. To find which terms (already stored in the database) are related with term A in a given relation R, Corpógrafo will search the entire corpus for sentences where A and any other known term co-occur in a relation-specific context. Corpógrafo's approach is similar to Hearst's (1992), we have just extended it to deal with relations other than hyperonymy. For example, to find terms that are meronyms of A (i.e part of A), Corpógrafo will scan the selected corpus for patterns like: "[A] is composed by [OTHER_TERM]" or "[A] has several [OTHER_TERM]".

We have compiled a small base of lexical patterns for Portuguese, English, Spanish, Italian and French, to help the user find several types of relations. Our experience,

however, tells us that these patterns do not occur as frequently as the definition patterns. Most semantic relations seem to be expressed in less explicit ways, or by more complex patterns. Still, and by using less restrictive patterns when needed, we hope that Corpógrafo is able to retrieve some interesting candidates for user validation, saving a significant amount of time also in the task of harvesting relations, although we have not yet performed a formal evaluation: validation of their performance has been informally done only by user feedback.
We are currently devising means and resources for evaluating the 3 extraction algorithms (terminology, definitions and semantic relations).

4. Who is using Corpógrafo?

The main difference between Corpógrafo and all other specialized workbenches we know of is that Corpógrafo is being used by a large user mass instead of only by one or two partners at most. Also, Corpógrafo's design strategy has always been user-centered and it has evolved from actual user needs and needs from real users. Although Corpógrafo is also distributed as an open source project, our main target audience are users without any computational background, and it requires only Web access.
First of all, Corpógrafo has been extensively used in terminology and translation training at the University of Porto, as well as elsewhere in Portugal in the universities of Aveiro, Minho and Lisbon. There is also an active community in Brazil. Additionally, Corpógrafo was also installed in Universitad Pompeu Fabra (Barcelona) in the summer of 2005, and users from this university have Corpógrafo running on their own servers. This cooperation led to extending some of the linguistic capabilities of Corpógrafo to Catalan.
Table 1 and Figure 2 give an overview of Corpógrafo's user mass (650 users in February 2006), registered in Linguateca's server (an additional hundred users are registered in Corpógrafo's server at Universitad Pompeu Fabra).

Users	Organization (country)
154	Univ. do Porto (PT)
49	Univ. de S. Paulo (BR)
39	Univ. do Minho (PT)
32	Univ. de Aveiro (PT)
31	Pontifícia Univ. Cat. Rio Grande do Sul (BR)
14	Univ. Federal do Rio Grande do Sul (BR)
12	Univ. Nova de Lisboa (PT)
11	Univ. do Vale do Rio dos Sinos (BR)
11	Univ. de Lisboa (PT)
11	Univ. Salamanca (ES)
10	Univ. do Algarve (PT)
9	Linguateca (PT/NO)
267	Independent users, groups with 5 or less users

Table 1: Current users of Corpógrafo.

We believe that now the time is ripe to make Corpógrafo also known to users outside the academic world. In fact, Corpógrafo may be used by large organizations that wish to develop their own internal terminology and ontological resources based on their document collections or archives. Professional translators or translation companies could also be interested in using Corpógrafo for developing their terminological databases for translation purposes from bona fide corpora. Translating technical terminology is one of the hardest parts of the job, and having readily available bilingual terminology resources is of great use in professional translation. Again, Corpógrafo can be a great help for compiling and organizing private terminological resources.

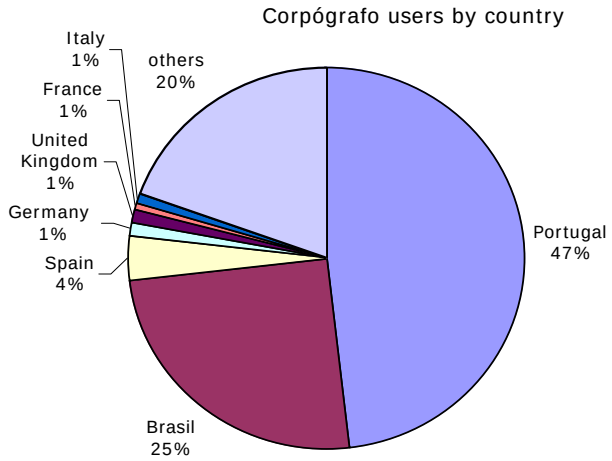

Figure 2: Corpógrafo users by country (total: 650 users)

5. Exporting Results in XML

An important property of modern information processing systems is the capability to exchange data with other systems. Up to version 2 of Corpógrafo, there was no standard way of exporting all the corpora and terminology resources produced by users to other applications. Corpógrafo V3 allows users to export all data compiled by users in standard XML files, for data-exchange and for backup purposes. Besides being able to export the corpora, users may export all data in databases into easy-to-parse XML files (Figure 3), for which the corresponding Document Type Definition file has also been made available. Users wishing to share their terminology databases may now do it in a more standard way. As more of these XML files become freely available on the web, we hope to encourage the development of semantic-aware applications, such as domain-specific (meta-)search engines, topic map interfaces or document classification tools. This will of course impose new requirements on Corpógrafo itself that will contribute to its future development.

```
<TU_ENTRY IID="38">

  <TU>axon</TU>
  <GEN_INFO lang="EN" iso_type="entrada terminológica"
iso_adm="estandardizado" iso_reg="neutro" iso_freq="usado com
frequência" iso_orig="empréstimo interdisciplinar"/>
 <MORF_INFO gender="U" number="U" animacy="U" pos="undef"/>
  ...
 <DEF_INFO CORPUS="Neurons" FILE="undef">
    <DEFINITION>The axon functions as a sort of conductor of electrical
signals .</DEFINITION>
    <COMMENT></COMMENT>
 </DEF_INFO>

 <DEF_INFO CORPUS="Neurons" FILE="undef">
    <DEFINITION>The axon is the main conducting unit of the neuron ,
capable of conveying electrical signals along distances that range from as
```

```
short as 0.1 mm to as long as 2 m .</DEFINITION>
    <COMMENT></COMMENT>
 </DEF_INFO>

 <TU_REL OTHER_IID="128" STRING="terminal button"
REL="HOLO/MERO" ROLE="HOLO" OTHER_ROLE="MERO"/>

<TU_EQUIV OTHER_IID="2" STRING="axônio" TYPE="sinónimo"/>
<TU_EQUIV OTHER_IID="210" STRING="axónio" TYPE="sinónimo"/>
<TU_EQUIV OTHER_IID="337" STRING="assone" TYPE="sinónimo"/>
<TU_EQUIV OTHER_IID="468" STRING="axon" TYPE="sinónimo"/>
<TU_EQUIV OTHER_IID="496" STRING="axón" TYPE="sinónimo"/>
<TU_EQUIV OTHER_IID="639" STRING="axone" TYPE="sinónimo"/>

</TU_ENTRY>
```

Figure 3: Snippet of the XML file containing one terminological database

5.1 Neurodemo

Neurodemo is a demonstration project whose aim is to illustrate Corpógrafo's capabilities. It started from a small project in a terminology class which resulted in two small technical domain corpora (one in English, one in Portuguese) in the area of neurology and an elementary terminology database. Currently, Neurodemo includes a terminological database with 1,885 terms in six languages (Portuguese (both European and Brazilian), English, French, Italian, German and Spanish), retrieved from corpora using Corpógrafo. The current corpora were bootstrapped by feeding keywords from the initial corpora in Google, thereby getting more texts on the subject (as well as corresponding terms in other languages). The Neurodemo corpora consist of mostly educational texts, since they tend to have explicit definitions, which makes them more suitable to test the semi-automatic extraction of definitions. Additional information about Neurodemo and its terminological database in XML is available from http://poloclup.linguateca.pt/Neurodemo.htm.

6. The future of Corpógrafo

Corpógrafo's capabilities related with technical genres and terminology extraction, as well as with the extraction of definitions and semantic relations, have reached a reasonable level of development. Still, there are obviously many possibilities for improvement, some of which we intend to pursue in the near future.

For example, for the two last mentioned functions, Corpógrafo would greatly benefit from a more comprehensive bank of lexical patterns. In order to decide how to proceed, we are currently investing some effort in validating and extending the current pattern bank. Another interesting improvement is to add bilingual terminology matching capabilities to Corpógrafo. In the current version, users have to manually match bilingual terminological entries stored in their databases, but simple heuristics, based on internal properties of terms and typical collocations, should allow Corpógrafo to suggest bilingual matches. By maintaining the semi-automatic approach, hope to simplify the work of the users considerably

Finally, we are working on the inclusion of externally compiled terminologies in Corpógrafo to allow quicker development and testing of thesauri or ontologies in technical texts.

7. Conclusion

By starting with very simple semi-automatic methods and focusing on user needs, we have developed a mature environment for terminology extraction and specialized corpora building whose source code has already been released under the GPL license. This workbench has currently a sizeable number of users, mainly students and researchers in linguistics and translation, who have also become more technical-oriented and more demanding about their tools. Now we are aiming our efforts to turn Corpógrafo into a knowledge engineering environment, suited for fast development of ontological resources from technical text, and helping developers to create semantically aware applications.

8. Acknowledgments

This work was partially supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI.

9. References

Cimiano Philipp and Johanna Völker (2005). Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. In Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems - NLDB'05 (pp. 227-238). Alicante, 15-17 June 2005.

Gulla, Jon Atle, Terje Brasethvik and Harald Kaada (2004). A Flexible Workbench for Document Analysis and Text Mining. In Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems - NLDB'04 (pp. 336-347). Manchester, 23-24 June 2004 .

Hearst Marti A. (1992). Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics (pp. 539 – 545). Nantes, France, 23-28 August 1992.

Maia, Belinda and Luís Sarmento (2003). CG - An integrated Environment for Corpus Linguistics. Poster at CL2003: CORPUS LINGUISTICS 2003 - Lancaster University (UK).

Maia, Belinda (1997). Do-it-yourself corpora ... with a little bit of help from your friends. In Lewandowska-Tomaszczyk, B. and P.J. Melia, (eds.) PALC'97: practical applications in language corpora (pp. 403-410). Lodz. Lodz University Press, 10-14 April 1997.

Sager, Juan (1990). A Practical Course in Terminology Processing. Amsterdam: John Benjamins Pub. Co. ISBN 90 272 20778 (pb).

Sarmento, Luís, Belinda Maia and Diana Santos (2004). The Corpógrafo - a Web-based environment for corpora research. In Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation (pp. 449-452). Lisboa, Portugal, 26-28 May 2004.

Sarmento, Luís (2005). A Simple and Robust Algorithm for Extracting Terminology. In the Proceedings of the META Simposium - For a Proactive Translatology. Université de Montréal, Québec, Canadá, 7-9 April 2005.