Terminology and Translation – bringing research and professional training together through technology

BELINDA MAIA
University of Porto, Porto, Portugal
bmaia@mail.telepac.pt

RÉSUMÉ
La terminologie devient plus descriptive que prescriptive tant au niveau théorique que pratique et, parallèlement, la traduction des langues de spécialité devient plus professionnelle. Ces processus se sont accélérés grâce aux possibilités offertes par les technologies de l'information et de la communication. Cette communication présentera notre suite d'outils en ligne permettant la compilation et l'analyse de corpus de spécialité, l'extraction terminologique et la gestion de bases de données terminologiques : le « Corpógrafo ». Celui-ci a été développé par des ingénieurs informaticiens, en étroite collaboration avec les enseignants et les étudiants d'un *Master* en Terminologie et Traduction. Nous insisterons ici sur ses implications au niveau de l'enseignement et de la recherche.

ABSTRACT
As terminology has become more descriptive than prescriptive in theory and practice, translation of special domain language has become more professional, and these processes have been accelerated by the possibilities offered by information technology. This paper will describe our online suite of tools for special domain corpora compilation and analysis, term extraction and term database management, Corpógrafo, which has been developed by computer engineers in close cooperation between the teachers and students of a Master's degree in Terminology and Translation. Special emphasis will be given to its implications for teaching and research.

MOTS-CLÉS/KEYWORDS
information technology, Corpógrafo, translation tools, pedagogy, terminology

Introduction

The underlying philosophy of this paper is that, although we believe translators must be educated to use and understand their own language and culture and those of their other languages, and to appreciate intercultural differences, they must also know how to use the relevant information technology and be educated to recognize that most well-paid translation is of a technical or scientific nature, that terminology is not 'jargon' but the expression of knowledge, and that most translation clients dream of a domain expert with training in translation.

However, few domain experts want to translate and translator education is traditionally for 'linguists'. Anyone involved in translator education recognizes the perennial problem of how to provide meaningful 'specialization' in the curriculum, rather than introductions to certain key areas. As academics, we are well aware that translation theory has tended to be for literary analysis, with an emphasis on intercultural and linguistic relativism. The advance of technology, in the form of machine assisted translation and machine translation, on the other hand, increasingly forces the language of both the original and the translation to be as standardized and as literal as possible, thus creating an even greater gulf between the theory and practice of translation.

This paper will describe how we have used technology creatively in research and in the classroom to accustom students to the applications of technology and to understand its underlying advantages and disadvantages. Our online suite of tools for special domain corpora compilation and analysis, term extraction and term database management, Corpógrafo, offers real possibilities for serious research, but it can also be used at a pedagogical level to encourage undergraduate students to reflect seriously on levels and types of text and terminology.

Attitudes to terminology and specialized translation

For many years terminology was by nature prescriptive, the objective being to clarify and organize terminological usage with the objective of encouraging unambiguous communication between people working in special domains. However, the nature of committees like the International Standards Organization, which exist for this purpose, lead to the terminology often being unrealistically restricted to that of a group of experts, and the price of the resulting standards often means that the results are of difficult access.

The unprecedented advances on all technical and scientific fronts in the 20th century led to a need for a more descriptive approach that took local factors into account. These factors exist, whether we are talking about different countries speaking the same language, competing schools of academic thought or commercial developments, or simply the variety of words used to discuss any specialized process at different social and cultural levels. As with other areas of knowledge, the emphasis is now on being descriptive, rather than prescriptive, and technology has provided the possibilities for this process. As organizations like the European Commission opened up their databases for public use over the internet, the increased accessibility to terminology has led to a greater awareness of its importance.

Terminology users are usually either experts or vested interests in certain special domains, or people who need to deal with technical communication and translation in these same domains. However, terminology has now become important to those involved in areas like Information retrieval and Knowledge engineering. Standardized terminology has its uses, but having the 'right' word is essential to finding relevant information, and the focus is now on perfecting search engines like Google, and structuring knowledge so it can be found fast.

From 'Do-it-yourself' corpora to information retrieval?

To make a claim that we have moved from simple classroom collection of texts in digital form for studying with the help of the tools of corpus linguistics (see Maia, 2003, 2002) to being able to venture into the world of information retrieval (see Oliveira et al: forthcoming) would be an exaggeration at this stage. However, the association of the Master's in Terminology and Translation at the University of Porto to LINGUATECA (see http://www.linguateca.pt) has proved to be an interesting experiment in combining the objectives of Natural Language Processing (NLP) with those of terminology production and specialized translation.

The objectives of the LINGUATECA project, with poles at Oslo, Lisbon, Braga and Porto, are to create and evaluate resources and tools for the computational processing of Portuguese. For some time, LINGUATECA has provided general linguists and translators with resources like the CETEMPÚBLICO corpus of 180 million words in Portuguese, and COMPARA – a parallel corpus of approximately 1 million words each of Portuguese and English literary texts, not to mention a variety of tools for NLP. The Porto group, PoloCLUP, has also produced a meta-engine for machine translation (MT), METRA, an environment for evaluating MT, TrAva, and a corpus of sentences + their MT equivalents, CORTA. However, the main focus of our efforts at PoloCLUP has been Corpógrafo.

Corpógrafo results from LINGUATECA's interests, the collaboration of a computer engineer with interests in artificial intelligence and information retrieval, the terminology and translation master's degree, and the interest of the domain experts we work with. Naturally, such a range of disciplines means that there are often arguments on priorities. The focus of Corpógrafo, therefore, has been to see the Big Picture, create an overall framework, get feedback from the users, and develop according to real research needs, rather than simply developing tools for the sake of it. This approach has left us with a flexible environment where we can add details and improve techniques as needed.

This means that the future of Corpógrafo may well result in new tools that will be of use to a variety of people. For this paper, however, we shall concentrate on its applications to the study of terminology and translation.

Working with Corpógrafo

It is important to remember that Corpógrafo is a suite of integrated tools for **individual** or **group** research. Restrictions to university computers in the past had made research and project work very difficult, so Corpógrafo was designed to allow access to these tools online from anywhere. Each username/password gives the user a space of 10 MB on our server but, as with other translation software, it consists of the tools and an empty space which the user has to fill with data.

One can acquire a username and password almost immediately after filling in the necessary form online. At present the instructions are in Portuguese, but there are tutorials and a glossary of expressions in English in .pdf format that can be printed out and used to guide the user through the instructions provided.

The File Manager

The first area one will need to work in is that of the 'Gestor de Ficheiros' or File Manager. In this area each individual or group can upload texts to their space on the server and convert most text formats to .txt, including .pdf and .ps text files that have not been actively protected. This is an important factor, since so much important terminology can be found in scientific and technical work published online using this type of file.

When the file has been uploaded, users are expected to fill in the metadata on the file. Apart from registering the name of the file, the title of the document and the language, they are expected to register the authors, the source, and date of publication, as well as define the special domain, sub-domain and text genre. If this is done properly, relevant information on authors and sources will automatically be inserted in the terminology database when a term is chosen from the text later on.

The next phase is to 'clean' the texts of unnecessary formatting residues and text. This can be done within the programme or using a word-processor like Word. Since we have worked a lot with engineering texts, there is also a function for eliminating non-text formulas and mathematical items. The next phase is to use the 'frasear' function which segments the text at sentence level, using the full stop as the natural divider of sentences.

Once one has uploaded a number of files, one can proceed to produce text selections or small corpora from these files. Again, one is expected to register the metadata on the corpus but, although each corpus should be restricted to one language, the user is free to group or re-group the files at will. For example, one may want to work on all texts of one special domain at one point but, on another occasion, the objective of research may be the study of a particular genre. In this case one can re-select files from several special domains into a new corpus without disturbing the previous selections. It is also possible to eliminate unwanted files and corpora as needed. The new version also allows for one to import and export a corpus from or to another environment.

Corpora analysis area

In this area there are concordancing tools that are typical of those used for general lexicography and linguistic analysis. They allow for Sentence and KWIC concordancing, and the observation of collocations. There is also an n-gram tool which allows for wordlists of one word as well as occurrences of combinations of more than one word to be found and sorted in alphabetical order or in order of frequency of occurrence The tools in this area and can be developed further if and when the need arises.

The ‘Centro de Conhecimento’ or Knowledge Centre

Since our research has been primarily devoted to terminology work, the tools for term-candidate extraction are kept separately from the corpora analysis tools, and they function within the overall framework of the conceptual databases. The expression ‘conceptual database’ was chosen over ‘terminology database’ because the overall objective is to create an environment that goes beyond the mere storing of terms and their multi-lingual equivalents and allows for conceptual and semantic organization as well.

The database for each domain must be created and then linked to the corpus or corpora to be used with it. Term candidates can be extracted using a term extraction tool that will be described in detail by Sarmento (2005: forthcoming). The terminologist can observe the term candidate list, check the validity of the terms by looking at related concordances and referring to the authors or sources of the original text, and transfer the term to the database, together with the metadata.

The terminology database itself provides further traditional terminology database fields such as information on the term’s status and morphology, as well as definitions, multilingual equivalents, and semantic relations. There are also tools for the semi-automatic extraction of definitions and semantic relations from the corpus using underlying patterns of words and phrases that tend to occur in suitable contexts. A simple concordance can also be made to assist in the finding of possible clues as to how to manually write a definition or define a semantic relationship. The new version of Corpógrafo will allow for the diagrammatic presentation of these relations, as well as ontologies, hypertext links, and links to multimedia representations of aspects of the terms considered of interest.

Progress made

The progress from the original methodology of encouraging students to collect mini-corpora in .txt form, analyze texts with Wordsmith and extract terminology, more or less manually, has been considerable. Formerly, student project work, often requiring considerable research and hard work, was done on a variety of software and, given the ensuing impossibility of coordinating the work, left forgotten in cupboards once its purpose of obtaining good grades had been achieved. Now, not only does Corpógrafo allow for much more sophisticated and rigorous research to be done relatively easily, it permits all the work done with it to be kept, added to and re-formulated. Its new import and export features will make it ideal for interdisciplinary and even inter-university cooperation, in a way similar to that advocated by Schmitz (2002:365-70) with the project Webterm.

Teaching Applications for translators

As several people have argued for some time (see Zanettin et al. 2003) training in general corpora use is invaluable when teaching translators. It trains students to go beyond the dictionary, observe words in context and select the ‘right’ word from the point of view of collocation, probably the single most difficult part of translating out of one’s mother tongue (something all professionals agree should not be done, but which is demanded by the market in many countries).

Special domain corpus building is also useful pedagogically, as it forces students to look for and evaluate different types of texts. In an educational environment which so often favours literary texts, it is important that they should learn to find and evaluate text according to genre, register and degree of information content.

Since this is also an exercise in how to find out about the very specialized texts that appear in professional situations, it is important that they should be guided to choose very specific subjects, for example, ‘Parkinson’s disease’, but not ‘neural diseases’ and certainly not ‘medicine’. They should be encouraged to start with encyclopedia type articles and pedagogical texts, as in this way they will actually manage to understand something about the subject, while they collect the basic terminology. Besides, these texts are often rich in definitions and explanations which will be useful for the database. As they progress, they will find that they search the Internet with increasingly specialized terms and gradually work towards finding peer-to-peer scientific texts with the most advanced terminology.

Although a project of the kind may seem restricted in terms of the domain chosen, the process teaches trainee linguists a methodology for dealing with any specialized domain or text type they encounter in the future.

There are, of course, other lessons to be learnt from this process. It does not work with all domains and in all languages. For example, although one can usually find texts in English, there is often a problem with languages like Portuguese. Even in English, however, a lot will depend on the local culture of the domain being explored. Engineering and Medicine, for example, seem to favour the publication of a wide variety of text types online, from the generally instructional to state-of-the-art academic papers. In other areas, especially in the humanities, there still seems to be considerable resistance to publishing online. However, although this means that corpus building can be frustrating in these areas, the difficulties in judging the quality of the texts involved can be educational.

Corpógrafo and Research

Corpógrafo offers a wide variety of possibilities for research. We have concentrated on making it work for our interest in terminology, and it certainly provides an interrelated set of tools that we hope will be useful to a wide variety of users doing normal terminology research.

However, there is also plenty of research to be done on the tools themselves, and this can involve both computational and more traditional linguists. Our term extraction tool has been created by observing how terms function in context and is the result of cooperation between linguists and an engineer. A similar method has been followed to help find definition and semantic relation patterns semi-automatically, but there is room for plenty more research here too. Researchers could reflect on and analyze the data found with a view to perfecting the patterns used for searching for further data. For example, they could investigate the nature of definitions and semantic relations as they appear in context, and they could use text analysis techniques to come up with ways of evaluating texts and finding further relevant texts. Although this type of research is not particularly new, Corpógrafo can help with the task and give it a better focus.

Although the need to support research into more general linguistics or lexicography has not been our priority, the fact that an individual can now create a personalized corpus, analyze it, and store results in a database within the same environment offers a variety of possibilities for research. To take a small example, let us imagine that someone wishes to investigate a particular syntactic or lexical feature using Corpógrafo. One would normally turn first to a large monolingual corpus like the BNC or CETEMPúblico to find a representative number of examples, process them manually, or by using a tool like Wordsmith, and then create one's own database for keeping the results of one's observations. Corpógrafo offers the researcher the possibility of forming a 'corpus' of the concordanced examples from the bigger corpus, editing and analyzing this corpus and sending the results to the existing database, all within the same environment. Although the present conceptual database contains fields that may be irrelevant to a study of this type, there is no reason why the researcher should not use it for this effect. Should the occasion arise in the future, it is perfectly possible for the database to be adapted and developed for research of this kind.

Lessons learnt

Corpógrafo has taught those involved the difficulties and advantages of interdisciplinary cooperation, on the one hand between language orientated disciplines and areas like engineering, geography and medicine and, on the other hand, between apparently such close areas as natural language processing (NLP) and terminology and translation.

Domain experts do not find it difficult to understand the implications of feeding Corpógrafo texts and extracting terms, definitions and other information. In fact, they become enthusiastic about the idea, and we hope to experiment in the near future with training domain experts to make their own databases, with or without the help of a trained terminologist. Since no one is yet prepared to pay a trained terminologist properly in our country, this exercise might demonstrate the need to do so, rather as the gradual realization of the difficulties of good translation is beginning to create a market that will pay good translators properly.

Lack of cooperation between computer-orientated and traditional linguists is patently obvious at any conference that tries to bring the two together. In particular, terminologists and translators traditionally tend to view computational methods with disdain - when they see the errors made, suspicion - as they begin to understand the methodologies used, and worry - when they understand the sheer ambition of the NLP engineers. However, since translation is ever more in demand, and correct terminology is a 'must' for information retrieval and related tasks, the fact is that NLP experts are expected to use their knowledge of computers to discover means of circumventing this human effort. Traditional linguists should not ignore what is being done, but be grateful for how much can be done semi-automatically and build on it. Corpógrafo is an attempt to make everyone realize this and it can be accessed freely at http://www.linguateca.pt/corpografo.

REFERENCES

MAIA, B. (2003): 'Training Translators in Terminology and Information Retrieval using Comparable and Parallel Corpora', in ZANETTIN, F., BERNARDINI, S. and STEWART D. (eds.): *Corpora in Translator Education*, Manchester, St. Jerome Pub, p. 43-54.

MAIA, B., HALLER J. and ULYRCH, M. (2002): *Training the Language Services Provider for the New Millenium - Proceedings of Encontros III de Tradução da AsTra-FLUP* 25-26 Maio de 2001, Porto, FLUP.

MAIA, B. (2002): 'Nothing is inherently boring – reflections on training translators in terminology', in MAIA, B., HALLER J. and ULYRCH, M. (2002): *Training the Language Services Provider for the New Millenium - Proceedings of Encontros III de Tradução da AsTra-FLUP* 25-26 Maio de 2001, Porto, FLUP, p. 355-364.

OLIVEIRA, D., SARMENTO, L., MAIA, B. and SANTOS, D. (forthcoming): 'Corpus Analysis for Indexing: when corpus-based terminology makes a difference', in *the Proceedings of Corpus Linguistics 2005*, Birmingham.

SARMENTO, L. (2005, forthcoming):'A simple and robust algorithm for extracting terminology'.

SARMENTO, L, MAIA, B. and SANTOS, D. (2004): 'CORPÓGRAFO – a web-based environment for corpora research', Poster presented at the LREC 2004 conference.

SCHMITZ, K-D. (2002): 'Webterm - Terminology from Academic Theses for the Web Community', in MAIA, B., HALLER J. and ULYRCH, M. (2002): *Training the Language Services Provider for the New Millenium - Proceedings of Encontros III de Tradução da AsTra-FLUP* 25-26 Maio de 2001, Porto, FLUP, p. 365-370.

ZANETTIN, F., BERNARDINI, S. and STEWART, D. (2003): *Corpora in Translator Education* Manchester: St. Jerome Pub.