Corpora for Terminology Extraction – the Differing Perspectives and Objectives of Researchers, Teachers and Language Services Providers

Belinda Maia
Faculdade de Letras
Universidade do Porto
Via Panorâmica s/n
4150-563 Porto
Portugal
bmaia@mail.telepac.pt

Abstract

Using corpora to find correct terminology is an activity that is interpreted rather differently according to the final objectives of those involved. This paper will try to show how the perspectives and objectives of researchers, teachers and language services providers do not always coincide, and how this lack of mutual appreciation and understanding can sometimes cause confusion. We shall first look at the more speculative aspects of current terminology research for the possibilities they offer in the future, even though some of this work is not directly related to translation, and consider the reasons why correct terminology is growing in importance in the lives of both domain specialists and language services providers. We shall then briefly consider both the older prescriptive notions of standardisation and the descriptive approach made feasible by technology and corpora today. Corpora in the broadest sense – from formally constructed and officially approved collections of texts to the disposable, do-it-yourself corpora anyone can now collect off the Internet for information on a specific subject – come as part of the information revolution provided by technology. They provide possibilities for any user of language and knowledge that were unthinkable a few years ago, but there are also problems and drawbacks.

1. Introduction

The compilation of terminology used to consist largely of collecting the words and phrases considered to be specific to a certain domain and bringing them together to form glossaries, with or without definitions or information on how or where the information was gathered. Since translators often had a vested interest in finding, or providing recognised equivalents in several languages, these glossaries would often become bi- or multilingual at a later stage. With the increase in availability of electronic text, the advantages of using corpora for term extraction are now generally recognised, particularly since the prescriptive view of terminology work has given way to a more descriptive approach, and the storage of definitions and other information on the terms has been made possible by relational databases.

This paper assumes that there are three classes of people with a particular interest in this terminology work. First there are the researchers in various areas of linguistics in general, as well as more specific terminology research. Many, but not all of these people, are also the teachers who try to train the professional language services providers needed today. The word 'linguist' as someone proficient in two or more languages has become ambiguous since the advent of 'linguistics' as an academic discipline, and the tasks required of someone with a good knowledge of languages are increasingly varied. I have therefore chosen the term 'language services provider' to refer to those who not only provide traditional translation and interpreting services, but also those who write and revise texts professionally, specialise in localisation, sub-titling, dubbing and making web pages, create terminological databases and translation memories, work with machine translation, and both use and take advantage of the information technology now available for a wide variety of projects and customers.

2. Terminology research

Those involved in this workshop on translation work and research will tend to see terminology research as primarily interested in supplying the needs of the translator for specialised terminology, but this is only one aspect of the overall picture. A good deal of terminology research is monolingual in nature and directed at the standardisation and categorisation of the relationship between concepts belonging to certain domains of knowledge and the terms used to describe them. This type of work is typically carried out by the domain experts, with or without the assistance of linguists, and, more often than not, in major languages like English, French and German. The subsequent translation of these standardised terms into other languages is by no means as simple or as well organised as it might be, despite official efforts to the contrary.

Standardisation of terminology has a long history, and its objectives have typically been to prevent confusion in the transmission of knowledge, with all the economic, social, legal and political consequences involved. Some areas of knowledge, like engineering, have a long-standing tradition in producing standardised terminology, but even they find it difficult to keep up with technical and scientific developments. Many other domains have little or no organised terminology resources and what exists is often 'local' in nature, in the sense that it is the property of certain organisations, companies and other entities, of varying size and importance.

The information revolution caused by the Internet, however, has led to demands for better systematisation of knowledge and improved accessibility. For this reason, the computational side of terminology research today is increasingly orientated towards facilitating information retrieval and knowledge engineering (see Budin, 1996, and Charlet et al, 2001). Traditional terminology work

tends to be painstaking and slow, and is not adapted to coping with the exploding need for retrieving knowledge. For this reason, efforts are being made by computational linguists and computer scientists to speed up the process of identifying, extracting and processing terminology (see Bourigault et al (Eds.) 2001, and Veronis (Ed). 2000).

3. Computational terminology

So much information is now processed in computer-readable form that there are obvious advantages to be drawn from this for machine (assisted) translation, translation memories and their related terminology databases. The corpora required for this type of research need to consist of texts that are not just well written, in the sense that they represent texts normally produced in a particular domain of knowledge: they need to use terms that are generally accepted in the community that works in that domain. When translations exist of these texts, they, too, need to conform to the same standards of text and terminology in the target language if one is to produce good aligned parallel corpora.

The experimental work done in computational terminology usually involves standardised texts in which both originals and translations are considered to be of high quality. Some of these texts have been provided by organisations like XEROX (see Bourigault 1994). The texts are often chosen for their linear compatibility (See Blank, 2001), which allows for easy alignment at, at least, sentence level, and the standardisation of their technical terminology. This is understandable, since it will only be possible to proceed with the analysis of a wider variety of texts when some sort of procedure has been worked out on the basis of these controlled corpora – rather as machine translation is better at translating controlled language than Shakespeare.

There is, of course, a lot of textual material that apparently conforms to the needs of this type of research. The European Commission has worked hard at making as many of its multilingual texts available as possible. In order to do this, the translation services have effectively created enormous translation memories full of texts translated by themselves, and one can presume that the terminology used is usually supported by the EURODICAUTOM database, which is itself the result of many years of effort by a large number of people. The large multinational companies that have invested heavily in translation memory software and terminology databases could also provide a vast amount of material. Organisations like the International Standards Organisation could provide invaluable material once its standards are efficiently translated in other languages. After all, not only do these standards and their translations represent ideal parallel corpora, but the very purpose of the texts themselves is to standardise the terminology used.

4. 'Real-life' terminology

There can be no doubt that a lot of the work to which we have just referred is impressive and of high quality and, therefore, a reliable source of information for the most necessary function of all these texts – the communication of knowledge. However, anyone who has worked seriously on producing terminology with the collaboration of experts will realise that the notion of 'one concept = one term' is an ideal, not a reality. International classifications that do exist have sometimes tried to escape the problems of normal language in different ways, as when natural species are classified in Latin, or chemical and mathematical concepts use formulas and symbols.

There are various reasons why the 'one concept = one term' notion is an ideal. It is easy enough for the linguist to understand the fluidity of the lexicon. After all, one of the perennial problems of general linguistics is how to deal with it in an easily classifiable way, hence all the work with projects like Wordnet (at: http://www.cogsci.princeton.edu/~wn/). On the other hand, experts in any particular domain are also aware of the fluidity of concepts and probably spend a good deal of time arguing about how to stabilise them for practical purposes - and stable terminology is only one aspect of this problem. In practice, they often resort to diagrams, images and other pictorial representations in order to circumvent or supplement the limitations of language. The general public, however, likes to believe in the stability of both language and concepts, and, for the practical purposes of communication, we all accept that there has to be some sort of 'social contract' whereby we agree to this stability in order to understand each other.

Prescriptive terminology has usually aimed at providing this stability in an organised fashion and most specialised dictionaries and glossaries are the result. The technology of databases, however, allows for a more descriptive approach, with all the implications this has for including all the information terminologists collect in the course of their work. When one is no longer limited by space on paper – a major factor in previous lexicographical work – the prospects of including all the information available and/or prescribed by international standards for terminological databases are, to say the least, tempting. These prospects may seem unnecessary to the more immediate problems of communication, but they contribute in no small way to various visions of the systematisation and documentation of knowledge.

Terminology is not the simple accumulation of words, their equivalents in other languages, definitions and a certain amount of grammatical information. Nor is it the simple matching of term to concept. One has to deal with all the usual problems of language - social, geographical, historical, political, and other aspects of style and register. At the level of standardisation, one can even become involved in authentic battles between academics or commercial companies who want to see the words they use to describe their particular theories or products prevail.

5. 'Real-life' corpora

When one is not working for the interests of computational terminology, one will probably not have access to the type of standardised corpora already described, except for the online documentation of the European Commission. Besides this, these standardised texts, no matter how well written or translated, tend to

reflect a degree of deliberate homogenisation of style and register across languages. In the more routine terminology work carried out in universities and other institutions, every terminology project will come up against a different situation, and circumstances will play an important role.

First of all, one has to find what texts are available in the domain one is studying and it is more than likely that the most important ones will not be in digital form. We have found that this is often the case when one wants to use first-class academic texts published by well-known publishers. Working with industrial or commercial institutions or companies is one way of obtaining texts, but we have not yet tried this, partly because it will require careful negotiation, and partly because we have found several academic partners interested in cooperating on a serious and more unbiased basis.

One can always scan texts, and there are, of course, plenty of texts already in digital form. It is often easy enough to obtain permission to use these texts if one explains why one needs them and what one intends to do with them, as there is plenty of interest among domain experts to see their terminology systematized. The Internet, as we all know, can provide an enormous amount of material in certain areas, but is less useful in others. For example, we have found it of limited interest for certain engineering terminology projects because both the high level expert-to-expert type of academic article and the more didactically orientated teaching text are not freely available to the general public. Too often one ends up with commercial sites trying to sell certain types of engineering equipment, and the information thus obtained is not necessarily very reliable. In the area of population geography, however, where one is dealing with a subject that cuts across the disciplines of geography, sociology and demography, one project group was able to find a sizeable amount of material in several languages, of both a parallel and comparable nature, precisely because there are plenty of official or governmental institutions who want to publish such material on-line. The other interesting aspect of this area is that the subject is relatively new and the relative instability of the terminology was observable in the texts found.

As our projects must have a Portuguese component, one of the problems we have found is that some languages are more equal than others. If the languages involved are English, French or German, there is a chance that one will be able to find reliable texts of a parallel or comparable nature, but the same will not be true of less used languages. We have found this to be true at all levels of text we look for. We have also found that the translations of websites - whatever the original language - are often of poor quality and cannot be used as parallel corpora.

6. Teaching and Project work

The type of project work we have done over the years started as a typical translation exercise in vocabulary research that owed much of its dynamics to the fact that the translation classroom contained PCs connected to the Internet. Our curriculum had been formulated by believers in the notion that 'general translation', together with six months placement at the end of the course, was sufficient for training Modern Languages students to become translators. Our experience, and that of our graduates, soon told us that this was far from enough and we developed specialised subject project work as a way of training students in LSP (see Maia, 1997 and Maia, 2000) within the limitations of the curriculum. We have now moved on to interdisciplinary postgraduate training in terminology and translation work, working with professors from the Engineering Faculty and History and Geography departments. Our early wordlists processed in Word have now developed into more sophisticated terminology work in Excel and Multiterm, and include definitions, sources, images and other data fields. We soon hope to have our own database system and make it available online.

Corpora have always been obligatory elements of our project work but, although we have collected quite a lot of specialised mini-corpora over the years, we admit that they have not always been the most successful part of the projects. There are various reasons for this. On the one hand, perhaps the biggest enemy of terminology related corpora work is the large number of existing on-line glossaries on everything under the sun that our students soon discover from each other. One can, of course, argue that these glossaries, which are often easy to copy or download, are in themselves language resources of the type we are discussing here. However, they are usually monolingual, largely in English, often rather general in scope, and infrequently backed up by any form of official recognition. When the glossaries are good, complete, and officially recognised, adding Portuguese terminology to them is usually beyond the scope of an undergraduate project. Of course, one might argue that beginners could do worse than discover how to convert them into their own languages.

The big problem here is that such work merely encourages the idea that finding the 'right word' is enough. This means they miss out on the didactic strengths of making mini corpora - the understanding of the subject itself, brought about by having to find and read texts, the appreciation of different types and styles of text gained while doing this, and the extraction of terms in context. Although students are encouraged to use software like Wordsmith to look for keywords and to study concordances of both general language words and specialised terminology, there is always a preliminary stage when the actual reading of the texts is necessary – at least from a pedagogical point of view. If they are lucky, they will also find definitions in the texts, although these are not as frequent, or as reliable, as the literature on the subject would have us believe.

There are successful types of glossary work that do not require corpora, such as some excellent ones our students have done on tools of various types – e.g. carpentry and gardening tools - in which the 'corpora' were largely catalogues with images, and students had to work hard to make the words in both languages match the pictures provided, a process that involved plenty of questioning of individuals, but little text work.

7. Conclusions

Corpora and terminology research can work well together, but they are not always equal partners. Ideally, students should be able to find good texts and extract terms, definitions and other information from them. When mini-corpora form the basis for terminology work, the process of producing the terminology project is didactically more valuable, and it is an easy step from collecting and aligning texts, and then using concordancing, to understanding the theory behind translation memories and other software and making them work in practice. As we have said, however, valuable terminology work can be done without resort to corpora. Perhaps the most important attitude to adopt towards project work is flexibility, since each domain brings its own circumstances and problems. If at the end of the experience our undergraduate students have learned how to take special languages seriously, the main objective has been achieved. Our postgraduate students already know how important they are and need to learn how to progress further, and perhaps even join the process of research into computational processes that will speed up the accumulation of valuable resources for all of us who do not want to see the world speaking only one language.

References

Austermühl, Frank, 2001 *Electronic Tools for Translators*, Manchester: St. Jerome Publishing.

Blank, I., 2001. Terminology extraction from parallel technical corpora. In D. Bourigault, C. Jacquemin and M-C. L'Homme. 237-252.

Bourigault, D., 1994. *LEXTER, un Logiciel d'Extraction de TERminologie, Application à l'acquisition des connaissances à partir de textes.* PhD thesis. Paris: École des Hautes Études en Sciences Sociales.

Bourigault, Didier, Christian Jacquemin, & Marie-Claude L'Homme, (Eds.) 2001. *Recent Advances in Computational Terminology.* Amsterdam & Philadelphia: John Benjamins Publishing Co.

Budin, G., 1996. *Wissensorganisation und Terminologie*. Tübingen: Gunter Narr.

Charlet, J., M.Zacklad G.Kassel D.Bourigault, 2001. *Ingénierie des connaissances.* Paris: Éditions Eyrolles.

Maia, B. 1997. Do-it-yourself corpora ... with a little bit of help from your friends! In Barbara Lewandowska-Tomaszczyk and Patrick James Melia (Eds.) *PALC '97 Practical Applications in Language Corpora.* Lodz: Lodz University Press. 403-410.

Maia, B. 2000, Making corpora – a learning process. In Bernardini, S. & F. Zanettin, (eds). 2000: *I corpora nella didattica della traduzione.* Bologna: CLUEB pp.47-6.

Maia, B., (forthcoming),'Comparable and parallel corpora – and their relationship to terminology work and training', paper presented at the *CULT - Corpus Use And Learning To Translate*. Bertinoro, Italy, November 3-4, 2000.

Maia, B. (forthcoming). ' Terminology – where to find it, and how to keep it', *Proceedings of III Jornadas sobre la formación del traductor e intérprete, Universidad Europea de Madrid 7 -10 March 2001.*

Veronis, Jean (Ed). 2000. *Parallel Text Processing – Alignment and Use of Translation Corpora.* Dordrecht: Kluwer Academic Publishers.

Wright, Sue Ellen e Gerhard Budin, 1997. Handbook of Terminology Management – Volume 1: Basic Aspects of Terminology Management. Amsterdam & Philadelphia: John Benjamins Publishing Co.

2001. Handbook of Terminology Management – Volume II: Applications-oriented Terminology Management. Amsterdam & Philadelphia: John Benjamins Publishing Co.