DO-IT-YOURSELF, DISPOSABLE, SPECIALISED MINI CORPORA – WHERE NEXT? REFLECTIONS ON TEACHING TRANSLATION AND TERMINOLOGY THROUGH CORPORA

Belinda Maia
Universidade do Porto

1. Introduction

Over the last few years we have seen an increasing number of teachers using the creation of 'do-it-yourself' mini corpora (Maia, 1997 and 2000) and (Zanettin, 2002), and 'disposable' corpora (Varantola, 2000) as a part of their teaching methodology. The product has usually been a small corpus of specialised texts and a glossary of terms extracted from it, but the educational process involved has been the major objective for most of us. By looking for texts, reading them, analysing their content and searching for suitable terms, students not only acquire the terms, but also knowledge about the subject and a familiarity with the style and register of texts associated with it. The process not only provides a partial solution to the problem of finding information on specialized subjects, it also dovetails nicely with current pedagogical theory, which exhorts teachers to become facilitators who encourage students to become active participants in the teaching/learning process.

When the right conditions exist, compiling small corpora and glossaries on specialized domains is, without doubt, an excellent

way of making students active participants in the learning process. However, the exercise is subject to the usual problems found by advocates of this methodology, as will be demonstrated in this paper. The other factor that must be taken into consideration is that, although one can provide general guidelines on this type of research, each project undertaken will need to adapt to circumstances.

Collecting mini-corpora is a simple form of 'information retrieval' or 'text mining', expressions that have developed from the current preoccupation with how best to extract information from all the digital resources now available. The corpora usually consist of 'raw' text, i.e. they are not annotated or tagged for morphological, syntactic or semantic purposes, and the emphasis is on encouraging students to observe how language behaves at all levels, from the lexicon to the text, rather than on providing a fully developed corpus for serious research. However, since one thing always leads to another, we shall also describe how this pedagogical technique is related to and can develop into something more far-reaching.

The methodology described presumes that both teachers and students have easy access to PCs, a scanner with a good OCR, a concordancer, a connection to the Internet and other digital resources. This is all technology that is becoming commonplace in most universities.

2. Do-it-yourself, disposable mini corpora

In the 1960s, most linguistic research was based on small corpora of examples or texts put together without the help of computers, until Chomsky pointed out that any such corpus was, by its restricted nature, skewed, and that it was safer to rely on the intuition of the intelligent native speaker. In those days, that was a justifiable objection, but the present large corpora are sufficiently representative of varieties of language to provide more reliable instruments than the intuitive method. Many people are now working

on the feasibility of small corpora for studying specific aspects of language (see Ghadessy et al, 1996).

This is particularly true of corpora for terminology extraction and the analysis of special registers and styles. Although the 'magic number' for such a corpus is one million words, some interesting work has been done on much smaller corpora. Bi or multi-lingual parallel or aligned corpora, or original texts + their translations, also provide interesting material for translation studies, as do comparable corpora, or corpora consisting of original texts in two or more languages in the same domain, register or style. As we shall see, they also provide research material for more ambitious projects than translation pedagogy.

The mini-corpora of the kind produced in the undergraduate projects described by Maia (1997) were the result of the realisation that the Internet and other sources could be used for this purpose. The idea of 'do-it-yourself' corpora suddenly became technically possible and the implications for teaching were obvious to those interested in corpora. However, since these early papers produced frissons among more traditional corpus makers, who were well aware of the dangers of infringing copyright, Varantola (2000) took care to coin the term 'disposable' corpora for collections of texts, which, once used for teaching purposes, were safely disposed of before copyright infringement became an issue.

The Internet had a decisive role in the emergence of small corpora for teaching purposes. After all, here was all this information, encoded in text that was easy enough to copy/paste to our PC and examine at our leisure. There is no doubt that – despite copyright - there is nothing much anyone who posts a text on the Internet can do to stop the individual next door or on the other side of the world doing precisely this. In fact, the function of many Internet sites is essentially to divulge information, and many have copyright notices more as protection against plagiarism than to prevent the individual from using such text for research into either its subject content or its linguistic form. For example, the European

Commission publishes an enormous amount of text online with the intention of providing information to the general public. We also know that all or most of it is stored in giant translation memories and that the Commission would probably welcome the fact that we are using these texts to train our students or do research (see Wagner 2002).

The position as regards specialized texts, of course, varies considerably. From the point of view of the translation teacher, there are plenty of texts available in certain domains, but it is difficult to generalise as to what can be found on the Internet, as the motives for posting information are so varied. For undergraduate work, a lot can be found in certain areas and up to a certain level of sophistication, but it is often a question of luck.

As the Internet has expanded, the original enthusiasm for free information for all, no holds barred, has become tempered by the real world of commercial interests and an increasing awareness of the problems of copyright and plagiarism. For example, the Encyclopaedia Britannica site – http://www.britannica.com - used to provide free access to the encyclopaedia, but now consultation is restricted. Some texts can be retrieved from CD-ROMs, including those of the Encyclopaedia Britannica, and it is possible, although perhaps questionable legally, to use these texts for research.

However, once one goes beyond the officially permitted, educationally orientated type of text, the information obtainable varies immensely, and if one is going on to study special domains in depth, the Internet's provision of suitable texts is patchy, to say the least. It is generally agreed that the best type of texts from which to extract terms + definitions are the books used to introduce any academic discipline. Quite naturally, both the authors and the publishers of any successful book of this kind will be too protective of its contents to put them on-line – although some now put a sample chapter at our disposal as an incentive to buy the book. As far as specialist writing is concerned, it is true to say that being published in print still carries more prestige, as this kind of publishing is monitored, whereas anyone can publish anything on the Internet

and a lot of rubbish is to be found. In spite of this, more and more specialists are becoming aware of the likelihood that their texts will be read more widely if they are on the Internet, rather than in (expensive) journals or the proceedings of conferences published by an individual university. The fact that many academic authors now contribute to the expense of publishing their work is proving another incentive to online publishing. In order to protect the copyright of their work, however, the texts can be read or printed from .pdf or .ps files, neither of which is easy to cut/paste into a do-it-yourself corpus.

3. Special domains and terminology

The training of translators to cope with specialised domains is by no means easy. Any discussion of the problem inevitably leads people to suggest that the best solution is to teach the domain specialist languages and the techniques of translation, rather than hand highly technical texts to a humanities trained linguist. However, although there are exceptions to the rule, specialists usually want to work in the area of specialisation itself, and look upon translation work as a secondary, if not unwelcome, aspect of their occupations.

The curricula of translator training institutions are usually designed by people from the humanities, often with little or no real appreciation of the complexities of specialised language. Although some translation teachers – particularly those who are also professional translators - have long made efforts to provide a variety of 'real-life' texts for their students, the fact is that many still do little more than skate around the problem, perhaps because their own research priorities involve literary rather than professional translation. However, the analysis of non-literary texts, often related to corpora analysis, means that studying languages for special purposes (LSP) is gaining importance and respectability in the world of academic research. This factor, together with a growing interest

in and need for good terminology, should help teaching and research interests at universities with translation courses to become more compatible. In the meantime, it is possible to develop a teaching and learning methodology for discovering how to specialise in any subject.

One of the claims often made in favour of a humanities education is that it trains people to think, and to analyse and manage knowledge. In order to justify such a claim, however, we must accept that the knowledge involved be wide-ranging and not merely restricted to the traditional, albeit multiple, interests of the humanities faculties. If we can do this, both in theory and in practice, we are on the right track for training translators who must, if they are to succeed in their profession, be able to become interested in any subject that is the topic of a text and its translation. It is this curiosity, and the knowledge of how to satisfy it, that should be encouraged. The belief – so common in literary faculties - that once one can translate literary texts, one can translate anything, is a convenient myth for those who do not want to face up to the texts produced by a wide range of professionals in the real world.

Domain specialists, on the other hand, are only too aware of the need for correct terminology. Many teaching textbooks in specialised domains include glossaries, and the large number of glossaries on everything under the sun on the Internet is further proof that the need for this type of information is often acute. However, their presence in print or on the Internet does not mean that they are reliable, and these glossaries are sometimes false friends to the translator. Despite all the research and consultation of experts that goes into preparing EURODICAUTOM - http://europa.eu.int/eurodicautom/login.jsp, there are still people who indignantly contest terms they find here. If this terminology can be criticised, how much more so the lists of terms – and their definitions – published everywhere, often by well-meaning secondary school and university teachers.

The fact is that terms, like words, need a context, and differing opinions on the correct term for X owe much to the geographical,

professional, social and even personal context in which it appears. At a formal level, we are talking about the battles that go on inside standardisation committees, as respected academics and captains of industry fight to have their favourite terms accepted. Such recognition may make one individual, school, or company (seem) more important than another. This means that, further down the social scale, the possibility of variation increases. For example, the construction engineer will probably use one term with his peers, another with the foreman at the warehouse, and yet another with the workers on the building site to designate a particular type of what the general public calls *brick*.

Standardisation committees deal with what is known as prescriptive terminology and, whatever theoretical reservations one may have, such terminology will make an important difference when the understanding of what constitutes a particular type of brick is essential to the safety of the building. However, in the field of terminology, as elsewhere in academic disciplines, prescriptive attitudes are giving way to descriptive methodology, and this has been helped by an understanding of the flexibility and capaciousness of terminology databases, as opposed to the need for lexicographical and terminological concision when creating paper resources.

The descriptive approach to terminology has led to plenty of interesting work that goes under the designation of socio-terminology, in which the social factors of terminology use are examined and, since Temmermann (2000), there has been considerable interest in the cognitive dimension of terminology and the use of metaphor. Besides this, the conceptual fields the terminologist has always dealt with, and which are implicit in the organisation of thesauri and other classification systems, are taking on a new importance as people struggle to further classify the world for reasons I shall describe below.

Corpora, or at least large quantities of electronic text, are receiving increasing attention as sources of information. It is partly for this reason that domain specialists are quicker to understand the

possibilities of specialized corpora than linguists. Although they acknowledge the truism that some of them do not necessarily write 'good' texts, they are more aware of what constitutes effective style and register in text than they are given credit for. They also understand the importance of context for terminology and, although those involved in standardisation committees are fully conscious of the need for standardised definitions, they also swiftly comprehend the need for the simpler, more didactically orientated 'definitions' that can be extracted from corpora.

4. Information retrieval – soft and hard

'Information retrieval' is a term that has developed with the appearance of the Internet. For example, in pre-Internet days, those translator trainers who recognised the need for specialised vocabulary struggled to provide it using general 'technical' dictionaries – which never seemed to specialise in what was actually needed - and the few specialised dictionaries they could find, or their institution could afford. More enlightened institutions offered their translation students introductions to law and economics. This was fine in itself, but was frustrating in that the information provided was rather like that given by a map of Europe, when what one needed was an Ordnance Survey map of a small area of the British Isles.

It was only natural, therefore, that translation trainers should very quickly discover the Internet as an invaluable source of information. As someone who encouraged my students to surf the Net with enthusiasm from the beginning, I have fought many battles with those who saw it as an innovation to be handled with deep suspicion. Now, some years later, I continue to be an enthusiast, but I have learnt to look for the rocks and dangerous currents, as well as judge waves which never break, or peter out, leaving us nowhere.

There is no doubt that often, even with just one word, we can find what we are looking for, and even find plenty of useful information associated with it, but success depends a lot on knowing how to choose one's word. It is also true that excellent project work can be done if a good conscientious individual or group of students works in close collaboration with the teacher and subject specialists. This is the type of work that makes the enthusiasts for the do-it-yourself method so optimistic. Both the process and the product of the project are very satisfying for all concerned.

However, teachers do not always have the time to accompany each and every student of a large class on their journey, and one of the dangerous currents is that which pulls the lazy or weak student into one of the glossaries referred to above. Once they have come to rest in one of these apparently safe havens, they tend to refuse to venture out in search of more information, particularly textual information, or corpora. Unless the teacher has time to control every movement they make, they will probably settle down to translating the glossary, usually in English, into their own language, using normal dictionaries and even, in the case of the weakest, actually selecting only the terms that are easily translated! During the year, the harassed teacher asks how work is going and is told that everything is fine, only to discover later – as s/he surfs the 'webliography' given – the strengths and (often severe) limitations of the work actually done. Some weaker projects produce corpora consisting of numerous, short, very repetitive texts. Online 'encyclopaedias' of natural species, like the Botany.Com Encyclopedia of Flowers and Plants at http://www.botany.com/, are fertile hunting grounds for the hard-working but not-so-bright student who believes quantity is better than quality. I have received work on sharks, amphibians, and reptiles with 'corpora' constructed out of such texts from similar web sites. I now discourage anything that involves a lot of cut-and-paste (often with nice pictures) but little genuine understanding of the subject chosen, as it is poor preparation for the real detective work required of a good translator.

Certain glossary projects can, of course, function very well without recourse to corpora. This is particularly true of studies of equipment and tools, where the 'corpora' are often catalogues found on the Internet, complete with pictures, a system that even tends to dispense with definitions. If the project then takes the student to local shops and supermarkets to check the words used in their own language to describe the object in the picture, a lot can be learnt about the limitations of the vocabulary used and about the relative reliability of different sources. In these cases the lessons learnt from the process complement or make up for the product of the project.

One can also argue that good glossaries in one language in themselves present a challenge to the student who proceeds to use traditional terminology methodology to find the correct term in another language. However, on these occasions, students rarely use corpora when they do this sort of work, and once the project focuses on the original glossary as the main source of information and the new glossary as the end product, we are drifting away from the main point of do-it-yourself corpora.

Although a good informative text should provide a selection of the technical vocabulary needed – and even some definitions – it will rarely provide the larger number of terms provided by a glossary. It is probably for this reason that students - who so often ask 'how many words do you want?'- find corpora collection a little frustrating. In order to get them to persevere, therefore, one must explain how the corpora should consist of texts that may actually teach them about the subject they are researching. If one points out to them that one is asking them to learn about the subject, rather than just collect words, they find it easier to understand the relevance of texts. If one then makes them analyse the type of text they are finding as examples of the style, register and general context in which the terms appear, and shows them how concordancing software like WORDSMITH can be used to find terms in context, they will usually develop the necessary enthusiasm. The important point text collection must make is that, in order to do a good

translation, we must know something about the subject and the type of text used to discuss it. Finding the 'right word' is not enough.

Once one moves from training undergraduates to be curious, to the more complex sphere of serious terminology training, the finding and use of specialised texts becomes more difficult. Cabré (1993; 1999) draws attention to all the traditional reference material used for finding terms – technical thesauri, dictionaries, glossaries, standards, etc., – but does not focus the use of corpora. Sager (1990), on the other hand, refers to the use of corpora for extracting terminology as common practice – although he does not give any concrete examples. Pearson (1998) investigates the type of corpora needed for extracting terms and ends up choosing one corpus consisting of texts written by experts for initiates, another of textbooks used by teachers, and a third of texts written for expert-to-expert communication (see Pearson 1998:64-66).

Every pressure is being put on linguists to use corpora for terminology extraction, and one objective is to encourage terminologists to consider terms in context, rather than in the isolation of word-lists. With databases and corpora accessible at a couple of touches of the mouse, access to the information thus provided is quick and easy, unlike the consultation of heavy, complex specialised dictionaries, and often inaccessible texts by experts.

As the parsing and tagging of corpora has become easier, computational linguists are searching for quicker ways of terminology extraction. There are three main ways of doing this:

1. One tags the texts for parts of speech and searches for the combinations of tags which find us noun phrases (the typical syntactic form of terms) - a system which over-produces possible items, or what is called 'noise';

2. One consults the domain expert for keywords with which to search the text – a system that may ignore words or word combinations that the domain expert overlooks, and leads to what researchers call 'silence';

3. One uses 'clues' like the verbs *be*, *mean*, and *consists of* and phrases like *part of, type of* etc.

Research of this kind is reported in works such as Bourigault et al (2001), Charlet et al (2001) and Rodriguez & Araujo (2002 - the LREC 2002 Proceedings). As one reads these works, one soon becomes aware that not only linguists, both computational and otherwise, are working in this area. Computer scientists, too, are talking about 'keywords', 'summarisation', 'informational retrieval', 'ontologies', 'semantic networks', and related subjects.

Impatient with the slow reaction of most linguists, and under heavy pressure from a world anxious to surf the net more efficiently, the computer scientists are working on making the above three methods more efficient, or resorting to statistical methods of text analysis to solve their problems. They talk of 'semantic tagging', and go where no linguist dares to tread, devising sets of tags that remind one of the componential analysis of the '70s, which can be used to differentiate between synonyms, but which proved unmanageable as a way of describing general language. The more conservative work on ontologies builds on criteria like the Universal Decimal System and adds the finer classifications supplied by traditional thesauri in specialised areas, which the advent of hypertext makes more flexible and easy to use. Then there is some interesting work in semantic networks that work by association of words as explained in Maia (forthcoming). Some of this work can be found if one explores sites like Wordnet or Wordsmyth for work on general language, and ONTOLINGUA and Semantic Web for specialised language (see Bibliography for site references).

5. Where next?

Although it is obvious that professional translators under pressure will rarely have time to check all their sources, this does not mean

that we should not train them towards an ideal that includes corpora and properly developed terminological databases. They will only learn how to aim for high standards if they are aware of what these standards are. Therefore, although every effort should be made to encourage translation teachers to take full advantage of every resource available, including the Internet, the emphasis should be on learning:

- about the subject
- how to recognise useful information
- how to check the reliability of information sources
- how to recognise and evaluate the different types of texts related to the domain in question

Acquiring these skills is more important than getting a quick answer to the translation problem of the moment. Training translators in terminology management will usually require more sophisticated corpora than the Internet can provide, such as introductory textbooks in special domains, ISO norms and other standards, expert-to-expert texts and other high-level technical documents, but high-flying translators will already recognise the need to use such documentation anyhow.

Training in the making of corpora also prepares translators to make translation memories using parallel texts. Increasingly, translators are finding that they are expected to prepare translation memories, research into and produce terminology databases, and work with machine translation and other technological aids. Corpora or text databases of different kinds are essential to these tasks.

Naturally, it is impossible to ask translators to keep abreast with the more ambitious types of research, but it is as well for them to have an idea of the big picture, because the world of work changes quickly, and often as a result of this research. We must learn to

follow the tide and also accept that there is a gulf between the slow, conscientious work of the traditional terminologist and today's demand for instant information that must be bridged somehow. Computational linguists and computer scientists are working to speed up the process, and terminologists, translators and general linguists need to work with them for the common good.

Bibliography

Bernardini, S. & Zanettin, F. (eds) (2000) *I corpora nella didattica della traduzione*. Bologna: CLUEB.

Botany.Com, the Encyclopedia of Flowers and Plants at: http://www.botany.com/

Bourigault, D., C. Jacquemin, & M.-C. L'Homme (eds.) (2001) *Recent Advances in Computational Terminology*. Amsterdam & Philadelphia: John Benjamins Publishing Co.

Cabré, M. T. (1993) *La Terminología; Teoría, Metodología, Aplicaciones*. Barcelona: Editorial Antártida / Empúries.

Cabré, M. T. (1999) *Terminology: theory, methods and applications*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

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

Encyclopaedia Britannica site at: http://www.britannica.com

EURODICAUTOM site at: http://europa.eu.int/eurodicautom/login.jsp

Ghadessy, M.A.H. & R.L. Roseberry (1996) *Small Corpora Studies and ELT*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Maia, B. (forthcoming) *Ontology, Ontologies, General Language and Specialised Languages*, in preparation by Centro de Linguística da Universidade do Porto.

Maia, B. (2000) "Making corpora – a learning process", in Bernardini, S. & F. Zanettin, (eds). 2000: *I corpora nella didattica della traduzione*. Bologna: CLUEB pp.47-61.

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. pp 403-410.

Maia, B., J. Haller & M. Ulrych (eds.) (2002) *Training the Language Services Provider for the New Millenium*. Porto: Universidade do Porto.

ONTOLINGUA site at: http://www-ksl-svc.stanford.edu:5915/

Pearson, J. (1998) *Terms in Context*. Amsterdam & Philadelphia: John Benjamins Pub. Co.

Rodriguez, M. G. & C.P.S. Araujo (eds.) (2002) *LREC 2002 –Third International Conference on Language Sources and Evaluation – Proceedings*. Distributed by ELRA.

Sager, J. (1990) *A Practical Course in Terminology Processing*. Amsterdam & Philadelphia: John Benjamins Pub. Co.

Semantic Web site at: http://www.semanticweb-org/

Temmermann, R. (2000) *Towards New Ways of Terminology Description - The Sociocognitive-Approach*. Amsterdam & Philadelphia: John Benjamins Pub. Co

Varantola, K. (2000) "Translators, Dictionaries and Text Corpora", in Bernardini, S. & Zanettin, F (eds) (2000) pp 117-133.

VERONIS, J. (ed) (2000) *Parallel Text Processing – Alignment and Use of Translation Corpora*. Dordrecht: Kluwer Academic Publishers.

Wagner, E. (2002) "Contacts between Universities and the EU Translation Services", in Maia, B., J. Haller and M. Ulrych (eds.) (2002) pp 397-406.

Wordnet site at: http://www.cogsci.princeton.edu/ ~ wn/

Wordsmyth site at: http://www.wordsmyth.net/home.html

Zanettin, F. (2002) "DIY corpora: the WWW and the translator", in B. Maia, J. Haller and M. Ulrych (eds.) (2002) pp 239-248.