The pedagogical and linguistic research implications of the GC to on-line parallel and comparable corpora

Belinda Maia
Faculdade de Letras de Universidade do Porto
bmaia@mail.telepac.pt

1. Monolingual corpora

Before going on to discuss parallel and comparable corpora we should just remind ourselves of what kinds of monolingual corpora exist and how they are used pedagogically and for research. The following is a short list of such corpora:

- Very large general language corpora - e.g. Cobuild, BNC, Mannheimer, PELCRA;
- Newspaper corpora - e.g. CETEMPúblico, Reuter's
- Small general corpora - e.g. LOB, Brown;
- Special genre corpora - e.g. literary, scientific, technical;
- Special mode corpora - e.g. written, transcribed spoken text, recorded speech, Internet 'chat', e-mails;
- Learners' corpora - e.g. ICLE - International Corpus of Learner's English - for studying second language learners' language output;
- Translational corpora - e.g. a project involving translations into English at UMIST, for studying how translated language behaves in contrast with original English.

Monolingual corpora were first thought of as an excellent basis for lexicographical research but, as time went on, research projects extended to various kinds of analysis, from the lexicon, to syntactic patterns to various aspects of text structure. Nowadays the study of the general lexicon covers the analysis of single and multi-word units in context, collocations, lexical groups, lexical 'bundles' (Biber, 2003), and 'priming' (Hoey, 2003). There is also a lot of interest in studying specialized lexicons, observing terminology in context, and the (semi-) automatic extraction of terminology. The pedagogical uses of monolingual corpora tend to be reserved for more sophisticated language study and, despite efforts to promote the use of corpora as an alternative or supplement to the dictionary as a language resource, it is fair to say that use of corpora by the general reading public is still in the future (Maia & Sarmento), 2003).

2. Multilingual corpora

Similar observations on the design, construction and use of monolingual corpora also apply to multilingual corpora. However, here we have to make some further distinctions:

- Parallel corpora = original + translation - e.g. COMPARA and most EC online documentation. Emphasis on variety of texts
- Translation memories = parallel corpora constructed for professional translation purposes usually using commercial software. Emphasis on similarity of texts
- Comparable corpora = originals in different languages that demonstrate certain facets of comparability at the level of genre, mode or domain

The research uses of multilingual corpora obviously focus on most kinds of contrastive linguistics and translation studies, and several people have explored the possibility of using them for pedagogical purposes. However, the different types of multilingual corpora are constructed and used for varying purposes.

3. Parallel corpora

Parallel corpora have been found to be very useful for studying the translation process and product at the level of the lexicon, syntax and sentence structure. Researchers, teachers and students can consult them in order to more fully understand examples of 'good' practice in translation, as well as to analyse examples of 'bad' translation.

The main disadvantage of parallel corpora is that that they pose questions of quality. They can only be constructed satisfactorily if there is considerable linear similarity between the original and the translation, and it is therefore difficult to make parallel corpora of functionally orientated translations and originals. When the texts can be aligned fairly easily in a linear fashion, the chances are that the language of the translation is skewed towards the source language.

However, despite the theorizing of academics on the subject and the professional practice among translators who find themselves obliged to produce 'translations' which could be more satisfactorily described as 'summaries', 'gists', 'paraphrases' or 'adaptations' of the originals, the fact is that technology is pulling translators and others towards increasing linearity between the original and the translation. Translation memories are constructed with a view to economising on time and money spent on texts and translations or a repetitive nature. They are more important for professional translators than for serious research, and their use for linguistic research and translation is therefore somewhat limited. However, the eCoLoRe project (http://ecolore.leeds.ac.uk/), for example, hopes to use them for research into areas like localization, and certain types of machine translation evaluation could find them useful.

4. Comparable corpora

Comparable corpora can be used in much the same way as monolingual corpora, with the added advantage of allowing people to study aspects of genre, mode or domain in different

languages using natural original texts. All corpora construction involves the choice and classification of the text types involved, but comparable corpora pose the further problem of comparability (Maia, 2003). There are several difficulties involved in finding 'balanced' comparable corpora of genuinely monolingual texts, and in building tools to search comparable corpora. Several of the problems involved are familiar to those involved in research into Information Retrieval.

Large monolingual corpora built according to comparable criteria (e.g. the BNC and PELCRA corpora) can be used for general language research and one can also construct comparable newspaper corpora for similar purposes. However, the emphasis of those involved in constructing comparable corpora is on the need for very specialised (narrow) corpora.

From a pedagogical point of view, several people involved in teaching translation and / or second languages have been encouraging students to design and use their own 'mini', 'do-it-yourself', or 'disposable' corpora as a means of studying various aspects of language, but especially for terminology extraction and genre analysis. (Maia, 1997; Varantola, 2000; Zanettin, 2002; and others). As the possibility of expanding the horizons of this sort of work has developed, it has led researchers in the areas of linguistics, terminology and translation to the point where their interests coincide with and complement those of researchers working on information retrieval and language engineering.

5. The research objectives and pedagogical applications of the GC

In order to encourage co-operation, Linguateca, and in particular the PoloCLUP, has been working on providing the means that will allow researchers in the areas of linguistics, translation, information retrieval and related areas to test their hypotheses on a wide variety of corpora. The objective is to create the necessary corpora and tools that will permit on-line corpus analysis, terminology extraction and co-operative construction of terminologies and ontologies. In order to avoid the conflicts of interest involving copyright restrictions, the working environment will be organized in a way that will permit people to work on material varying from generally accessible corpora to the texts restricted to the use of the individual researcher.

Tools for machine translation evaluation have also been devised, and we are working on an experimental pilot project that we hope will lead to further research in this area. We also intend to create online instruction modules, and a guide to relevant bibliography and online resources.

This may seem an ambitious project, but some of it is already on-line, and a good deal more is already prepared, or under development. We have received help, encouragement and suggestions from various sources and we hope that this event will provide us with the opportunity for discussion and an exchange of ideas.

Bibliography

[1] Bernardini, S, Zanettin, F (eds) 2000 I corpora nella didattica della traduzion. Bologna, CLUEB.

[2] Biber, D (forthcoming) 'What does frequency have to do with grammar teaching?' Plenary lecture at the PALC 2003 (Practical Applications of Language Corpora) conference at the University of Lodz, Poland, 4-6 April, 2003.

[3] Bowker, L, Pearson, J 2002 Working with Specialized Language- a practical guide to using corpora. London/ New York, Routledge.

[4] Hoey, M (forthcoming) 'What can the corpus tell us about linguistic creativity?' Plenary lecture at the CL 3003 - Corpus Linguistics - conference at the University of Lancaster, 27-31 March, 2003.

[5] Laviosa, S 1997 How Comparable Can 'Comparable Corpora' Be? In Target 9(2), pp 289-319.

[6] Lewandowska-Tomaszczyk, B, Melia, P.J (eds.) PALC '97 Practical Applications in Language Corpora, Lodz, Lodz University Press.

[7] Maia, B 2003 'What are Comparable Corpora?' In the proceedings of the pre-conference workshop on Multilingual Corpora: Linguistic requirements and technical perspectives, at the CL 3003 - Corpus Linguistics - conference at the University of Lancaster, 27-31 March, 2003.

[8] Maia, B 2000 'Making corpora: a learning process'. In Bernardini, S. & F. Zanettin, (eds.) pp 47-6.

[9] Maia, B 1997 'Do-it-yourself corpora ... with a little bit of help from your friends'. In Lewandowska-Tomaszczyk, B, Melia, P.J (eds.) pp. 403-410.

[10] Maia, B, Haller, J, Ulrych, M (eds.) 2002 Training the Language Services Provider for the New Millennium. Porto, Universidade do Porto.

[11] Maia, B & L. Sarmento (forthcoming) 'Corpora and the 'general public' '. Paper presented at the PALC 2003 (Practical Applications of Language Corpora) conference at the University of Lodz, Poland, 4-6 April, 2003.

[12] Pearson, J. 1998 Terms in Context. Amsterdam/Philadelphia, John Benjamins Publishing Company.

[13] Varantola, Krista. 2000 'Translators, Dictionaries and Text Corpora.' In Bernardini, S. & Zanettin, F (Eds). 2000, pp 117-133.

[14] Zanettin, F 2002 'DIY corpora: the WWW and the translator'. In Maia, B, Haller, J, Ulrych, M (eds.), pp 239-248.