PALC '97

Practical applications in Language Corpora

Do-it-yourself corpora …..
with a little bit of help from your friends!

Belinda Maia, Ph.D.,
Universidade do Porto,
Faculdade de Letras,
Via Panorâmica,
4150 Porto,
Portugal.

1.0 Introduction

I fully realise that to talk about 'making corpora' is to place myself in what some would consider the early history of corpora making. Nowadays, those who have access to many-million word corpora are rightly interested not in the making of corpora but in working out what to do with them, and inventing software to help them in the task. I wish I were in the same position. As it is, I can only report on a teaching technique which I am trying to develop related to the making of mini-corpora.

1.1 Problems related with making and using corpora

My position is that of a corpora enthusiast in a university in which people are only just becoming interested in corpora. Corpora making tends to be regarded as a project requiring heavy hardware, software and funding, and in the past it has been only for those who undertake such projects and can afford it. Disinterest in corpora by those not involved is a reaction sometimes produced by the fact that corpora are often jealously guarded by their makers. This reaction has probably contributed to the reason why some describe themselves as theoretical linguists and look upon corpus linguists as mere 'data gatherers', as Halliday (1993:24) points out. However, as he says, they can no longer ignore the fact that 'data gathering and theorizing are no longer separate activities'.

I have followed the fortunes of corpora making for some years now, and understand something of the work involved. I particularly feel for those who find that all the blood, sweat and tears they put into building up a corpus manually, or with prehistoric forms of hardware and software, are now rendered obsolete by modern IT. I know how they feel. To use a now popular phrase - "Been there - done that!"[1] Modern technology, however, is making the making of small specialised corpora much easier.

The present position of corpora in English is not something I need to dwell on here. Suffice it to say that there are corpora of all types and sizes to be had in various forms and for varying sums of money. However, access to corpora in Portuguese is not easy. There is a corpus in Lisbon, and there are corpora in Brazil, but access to them is very restricted.

I would argue that electronic corpora will become essential to many kinds of research in the future and that their applications to the teaching of language and translation will be invaluable. An ideal situation would be for researchers, teachers and students in every (university) language course to have access to a large corpus in each language studied. However, although efforts are being made in this direction, the practice is hardly

[1] Back in the dark ages of the 80s, I prepared my doctorate using a corpus of English literary texts, largely supplied by Oxford University Computer Service, and a corpus of Portuguese literary texts, scanned by myself and a colleague. Each corpus contained nearly a million words.

widespread. For one thing, 'computing in the humanities' is still an area which provokes a strong feeling of suspicion, if not fear, among older academic staff, who often control programmes and expenditure.

1.2 Why should one want to build one's one corpora?

With so much corpora material now available, one may well ask why one should bother to build one's own. There are various reasons, however, why one is sometimes forced to do so. First of all, some of these corpora require fairly powerful hardware and software, which are not always available in humanities research centres, let alone at the level of CALL centres in undergraduate language and translation courses. Secondly, a large corpus of texts in a particular language may be non-existent or inaccessible, as in the case of Portuguese at my university. Besides this, one may require a particular type of text - as in LSP -, or one may want to use parallel and comparable texts in two or more languages.

However, there is another more subtle reason. Although one can, of course, with perseverance, persuade the powers-that-be to invest in ready-made corpora - a good persuasive tactic is to prove the desirability of such a purchase by showing what can be done with a home-made corpus. The process also helps all involved to get a clearer view of what is needed, and the main point of my paper is to show how making your own corpora can even be an educational process in itself.

2. 0 The process of corpora-making

If one is to embark on corpora-making one must ask oneself a series of questions. The first step is to decide what one wants from a corpus. Simply dreaming of producing the 'ideal' corpus of unlimited size, will get those with limited resources nowhere. The second step is to discover how far it is technically or physically possible to compile the type of corpus one requires. To demonstrate what I mean, let me take my own experience as a case-study.

2.1 Perceiving - and creating - needs

Many projects of this nature start with the perceived needs of a particular individual. I have long believed that linguistic research without sizeable corpora produces only speculative results. In my type of personal research into the cognitive 'shape' of linguistic items based on quantitative analysis, the larger and more varied the corpora, the better. However, more specific research involving the contrastive analysis of aspects of English and Portuguese has led me to build up small corpora of a comparable nature - texts of similar genres and subject matter - as well as parallel corpora.

Although our careers are dependent to a large extent on our personal research, one should never forget our responsibilities to our pupils. Apart from the fact that they are useful guinea pigs on whom to try out one's hunches, their needs often generate meaningful research. My pupils are studying to be translators and I am in contact with them in the 3rd and 4th years of their course, during their period of training and, in some cases, even to Master's degree level. I have therefore come to consider them among my 'friends'.

Experience has also taught me that the wider the application one proposes for the resources one wishes to acquire, the more the powers-that-be listen to you. It is therefore sensible to find reasons why electronic texts might be useful to other departments. Apart from linguistic analysis, one can point to the need for electronic versions of literary texts for literary analysis. People also prick up their ears when one talks of the desirability of using academic and scientific texts as a basis for specialised terminology building. This is of particular relevance with a language like Portuguese where experts often have to invent translations of terms originally created in English or other languages.

2.2 Defining objectives

It is probably over-ambitious - if not unnecessary - to aim at producing large general corpora, but it is a good idea to keep the idea in the back of one's mind. It makes one think about the more general principles behind corpora-making, and helps to make one more objective and constructive[2]. It is obvious that more general corpora are an invaluable resource for studying lexical and syntactic problems in context, but I have found that parallel and comparable corpora are a must for translation work. Parallel texts are an invaluable way of seeing how translators actually deal with specific problems at every level from the lexical to the cultural. Comparable texts, which are all originals in the mother tongue, are particularly useful for examining differences in conventions in text and sentence structure between languages and the cultures they mirror[3].

Making corpora for translation purposes also means that one can concentrate on written texts, which are easily available. Parallel literary texts have their uses for general language study as well as specific problems of literary translation. However, for those involved in earning their living through translation, comparable and parallel texts from journalism and informative texts are an invaluable source for lexical usage in a variety of fields where vocabulary evolves at a rate only the most modern dictionaries can cope with.

2.3 Activating processes

[2] Besides, building up electronic text resources for linguistic research and language teaching usually requires one to select a wide variety of useful modern texts. Any material that results may also serve as a basis for bartering with other corpora makers.

[3] See Maia (1995), (1996) and (1997).

Whatever the objectives of one's corpora, someone has to make them. The most time-honoured way of doing this is by typing and scanning texts. This is either done in exchange for payment, through slave labour, or for the love of the cause. Most of us have used these methods in the past and have also learnt to pool our resources with others.

Nowadays, however, life is easier because there are a growing number of sources of readily available electronic text. The Internet provides an almost unlimited source of texts, there are a growing number of CD-ROM's, and more and more people are being obliged by circumstances to produce their original texts in electronic form. The process of acquiring electronic texts has therefore accelerated to the extent that it has made corpora-making something that can be used in the learning process.

3.0 'Making corpora' - a learning process

As assignments, my pupils have always prepared small but specialised glossaries using not only the dictionaries available, but also texts on the subject on which they were working. Although the end-product of the glossary projects is important, the process of making them is probably more useful. Choosing a subject, finding texts, discussing the subject with an expert and analysing the specific problems of the area chosen are all part of the learning process.

3.1 Creating the incentive

This year, circumstances have offered us a marvellous opportunity to use more sophisticated Information Technology. The computers in the Translation Room are connected to the Internet and the library has also acquired CD-ROM's like the Encarta Encyclopaedia, and the Times and the Guardian on CD-ROM. An Open and Distance Learning project has also proved a valuable asset.

The experiment described here has been carried out largely with students in the 4th year who had already done one glossary project in the more traditional way. At the beginning of the year only about half the class even knew how to turn on a computer, and none of them had any experience of the Internet. However, it took very little time to generate interest and within a few weeks most of them were hooked, and there was no difficulty in getting them to work outside class time.

The translation classes in the first semester are based on texts of the type which students have been asked to translate into English during their training period[4]. In the second semester we have always moved on to specialist texts of a technical, scientific or

[4] Whatever directives translators' associations and others draw up, the fact remains that the employing public still expect translators to translate into the foreign languages they know. The same employing public still believes that 'anyone' can translate into their mother tongue.

academic nature. This year the pupils were asked to choose a specific topic at the beginning of the year as a basis for assignments. These topics included ecology, information technology, sport, art, international conflicts and electoral systems.

The topic chosen serves as a focus for several things. For one thing, it encourages them to bring texts of their own choosing to type while learning word processing. It also serves as a goal as they find their way around the CD-ROM's and the Internet. As they go, they collect texts they find interesting for their projects, and these serve as a basis for their glossary-making. As enthusiasm built up, I decided to give them the job of finding texts on their topics, preferably from a variety of genres, for their colleagues to translate

3.2 The results so far

When I started this experiment, I could only guess at the possible success of this experiment. I am glad to be able to report that so far it has exceeded my expectations.

3.2.1 Collecting texts

The process of collecting material has created an enthusiasm I hardly dared hope for. The possibility of storing texts on a diskette instead of on paper, as in the past, means that far more material can be collected at a minimal cost. Gone is the expensive photocopying of material for the teacher to see you have done the work. Instead, quantity of text can be calculated quickly by the word processor, and the teacher can concentrate on assessing the quality.

By the end of the year, when the process stops, each group will have produced a mini electronic corpus on the subject chosen. You may think that this is just a ploy to get one's pupils to make corpora, but the important point is that they are the first to recognise the usefulness of both the process and the product. The process of looking for texts electronically has led them in various directions which would have been difficult - not to say expensive - to explore using conventional paper resources.

The groups have encountered different problems in searching. The IT group has found an embarrassment of riches, and selection of relevant texts has been more important than finding them. The ecology group has found itself surfing into all kinds of sites from the serious and political to the downright wacky. The individual doing international conflict has been able to use his years of experience as a newscaster to considerable advantage, and is exploring not just newspapers but TV and radio sites. Electoral systems have been explored in various ways by a group including a Portuguese pupil involved in party politics, and ERASMUS students from other countries who have had to try and work out how to explain their systems in other languages. One of the girls doing sport said there was endless material on the Net, but complained that most of it was not specialised enough! The most difficult area has proved to be art, given the width of the subject and

the fact that people who write in depth on the subject seem to avoid computers, or at least the Internet.

3.2.2 Glossary making

Glossary making in areas such as those chosen has proved far easier using up-to-date texts than out-of-date dictionaries[5]. For example, if one consults the Times CD-ROM for last year, one will see that the phrase *common market* may occur 108 times, but its usage is either generic, or used to refer to what is now history. Similarly, *EEC*, used 99 times, is usually preceded by *former* and refers to past situations. *EC* is used 275 times and *European Community* 183. Compare these numbers to those of *EU* which is used 1714 times and *European Union* with 1662 entries and it rapidly becomes clear how terminology evolves. A quick look at the English Learner's dictionaries will show that the latter two expressions have not yet made it to the dictionary.

The text chosen to exemplify international conflict produced a similarly interesting result that could only be reached using up-to-date sources. The average dictionary will not be too clear on how to make combinations of *Bosnian, Croat, Serb* and *Muslim* into the politically acceptable versions at present being used - consulting newspaper text will.

Although the process of glossary-making has revealed plenty of examples like this, the results will only be truly apparent at the end of the year when they hand in the product of their research. One important facet of the process, however, has been that it has become abundantly clear to them that such glossaries are open-ended and need to be processed in some form which will allow for future expansion.

One by-product of the Net searching has been that several ready-made glossaries have been found. They are usually monolingual, but they serve as a useful basis for both the process and final product of the projects. We are collecting them for future use.

3.2.3 Translation

The first concrete results have been the texts which the pupils have chosen for translation, and it is clear that considerable care has been taken with the choice. The texts have to include relevant lexical items and show variety of style, and each pupil has to present them and the translation problems they pose to the class. Delegating responsibility of this kind to the pupils should not be undertaken by teachers worried about loss of face - the chances are that, since they have worked hard on the topic using

[5] Although the modern English learner's dictionaries include a quantity of up-to-date vocabulary, it tends to be general rather than specialised. If one adds to this the fact that the modernisation of English /Portuguese bilingual dictionaries, as well as Portuguese monolingual dictionaries, is being held up by the failure of the Portuguese and the Brazilians to agree on spelling conventions, it is easy to understand why the translator has to find alternative ways of 'finding the answer'.

dozens of texts, they are bound to be better-informed than you are! If loss of face does not worry you, however, the resulting gain in confidence for the pupil is well worth it, and the class becomes better aware of the relevance of their research.

3.2.4 Other developments

The Open and Distance learning project in which the class is involved has also prompted us to record our experiences in a way which is emerging as a valuable learning incentive - making your own Web page. We are at present constructing our own Translator's page[6]. It is not an original idea - there are already several on the Net - but there is always a new perspective to be found, and constructing opinions about the sites one has visited - and translating them - gives a further objective to our work.

4.0 Other Friends and ideas

While my pupils have been busy making mini corpora, I have been following up other leads - with other 'friends'. Colleagues have bartered texts of their own and there have been offers of help. An increasing interest has been shown in the applications of corpora, not all of it confined to linguists. One of the more unusual requests was from a colleague in the Geography department who was interested in investigating how Biblical texts reflect their writers' attitude to meteorological events.

Translation corpora can also be of interest to every department in the university. Specialised terminology often originates in academe, and if one sets out to create corpora from texts produced by our university colleagues, one is preparing the ground for up-to-the-minute terminology banks. Over the last year or two, colleagues in other departments have enlisted translation students in their training period in projects involving the processing of glossaries and translation of texts for disciplines such as Geography, Psychology and History of Art. This trend has taken a more serious turn with the need for terminology for World Heritage projects such as that involving Porto.

A combined interest in terminology and the analysis of scientific texts led to the realisation that the work of our colleagues is a source of useful material. Permission has been given to use the electronic versions of the articles published in the Arts Faculty journals, and this experiment could be extended to other Faculties in the future.

My circle of 'friends' is gradually growing and now includes the Translation Services of the EU who have been most helpful in providing ideas and material. In my abstract, I state that 'the other purpose of the paper is to make further 'friends' and to promote the interchange of ideas and material in this area'. I hope that this paper will help me to achieve this.

---

[6] Address: http://www.letras.up.pt/translat/i_translat.html

REFERENCES

HALLIDAY, M.A.K. (1993) 'Quantitative studies and probabilities in grammar'. In HOEY, Michael (Ed) (1993) *Data, Description, Discourse - Papers on the English Language in honour of John McH Sinclair.* London: HarperCollins Publishers.

MAIA, Belinda. 1996. ' Parallel text and the study of sentence formation' in THELEN, Marcel and LEWANDOWSKA-TOMASZCZYK, Barbara. *Translation and Meaning - Part 4.* Maastricht : Universitaire Pers Maastricht.

MAIA, Belinda. 1997. 'The Sentence as a unit of translation' in II Jornadas de Tradução - *Tradução, Cultura, Sociedade.* Porto: ISAI.

MAIA, Belinda.(forthcoming) 'Sentence structure and thematization in comparable and parallel texts' presented at the 'Transferre Necessere Est' Conference in Budapest, Hungary, 1996.

MAIA, Belinda.(forthcoming) 'Word Order and the First Person Singular in Portuguese and English' in Laviosa-Braithwaite, Sara (Ed.) *The corpus-based approach: a new paradigm in translation studies.*