COMPARA, language learning and translation training

Ana Frankenberg-Garcia
Translation Department, ISLA, Lisbon

This paper is an introduction to COMPARA and to how it can be used in language learning and translation training. COMPARA is a machine-readable and searchable collection of Portuguese-English and English-Portuguese source texts and translations. The present corpus is made up of published fiction. However, COMPARA is open-ended, and other genres will be added to the corpus at a later stage. COMPARA is freely available on the Web and has been made for people who have never used corpora before as well as for experienced corpus users. COMPARA's criteria for text alignment allow users to investigate translational discourse changes such as when and where translators have chosen to join, separate, delete, add and reorder sentences. Other innovative features are that users can inspect translators' notes, and that the corpus admits more than one translation per source text. COMPARA is encoded according to the IMS Corpus Workbench system, developed at the University of Stuttgart, and is distributed on the WWW via the DISPARA interface, developed in collaboration with the Computational Processing of Portuguese project. In addition to countless theoretically-oriented contrastive studies of language, COMPARA also lends itself to quite a significant number of practical applications. It can be used in the development of bilingual lexicography and terminology, and for refining machine-translation programs. The final part of this paper will focus on some of the more immediate uses of COMPARA. A couple of practical examples of how it can be used in second language learning, teaching and translator training will be presented.

I. Introduction

This paper is an introduction to COMPARA, the Portuguese-English parallel corpus. The word corpus is being used here to refer to a collection of texts held in a machine-readable form so that they can be automatically processed by a text-retrieval program. Notable examples of monolingual corpora include the Bank of English and the British National Corpus, both of which are extremely useful to help us understand the English language as it is used today. For Portuguese, one of the most impressive corpora that exist is CETEMPúblico, which contains around 180 million words of machine-searchable contemporary European Portuguese. In addition to monolingual corpora, there are also corpora that contain more than one language, like the English-Norwegian Parallel Corpus (Johansson et al. 1999). Modelling itself on the core structure of the latter, COMPARA is a machine-readable and searchable collection of source texts originally written in Portuguese and in English that have been aligned with their respective English and Portuguese translations.

Two special features of COMPARA are that it is fully searchable via the Internet and that it has been made for people who are not necessarily corpus-literate as well as for experienced corpus users. Potential users include Portuguese learners of English, English learners of

Portuguese, students and teachers of translation, professional translators, bilingual dictionary makers, developers of machine translation software and whoever else might be interested in translation language and in the similarities and differences between Portuguese and English.

The advantages of using a corpus to compare and contrast English and Portuguese are that corpus-based analyses can be more objective, more systematic and a lot more extensive than analyses based on conventional introspective linguistics. In order to use a corpus well, however, it is important to know what the corpus is made of and how it is structured.

II. Selecting Texts

When selecting texts for COMPARA, all varieties of Portuguese and English were considered, and no priority was given to any particular variety. In terms of date of publication, both contemporary and non-contemporary texts were accepted. In addition to this, the possibility of having a source text aligned with more than one translation was not ruled out. Having established this, it was decided to begin the corpus by assembling an initial collection of published fiction, although other genres are to be included in the corpus at a later stage.

The decision to leave COMPARA open-ended was taken partly so that it could grow in whichever direction proved to become important to its users, and partly because this meant the texts incorporated in the corpus could be put to use as soon as they were processed. The second of these two reasons is not trivial: it meant that it was possible for the corpus to become operational within a reasonable amount of time.

III. Copyright permissions

At the time this paper was written, COMPARA had permission to include extracts of 60 different Portuguese-English text-pairs by authors and translators from Angola, Brazil, Mozambique, Portugal, South Africa, the United Kingdom and the United States. These texts represent the combined product of the work of 33 authors and 31 translators[1].

Because COMPARA allows for the inclusion of more than one translation of the same source, some interesting text-pair combinations have emerged. For example, permission has been obtained to include extracts from a couple of novels by David Lodge paired up with both their Portuguese and Brazilian translations, which can be useful for the study of similarities and differences between Brazilian and European Portuguese. Another interesting example is that

of a Brazilian nineteenth century Romantic classic, Iracema, which has been paired up with a contemporary English translation published by Oxford University Press less than a year ago and a contemporaneous translation which dates back to 1886 - this could be interesting for a diachronic study of translation.

IV. Preparing texts

The procedure for preparing texts for COMPARA is as follows:

1. The texts in the corpus that are not available in electronic form are scanned and submitted to an optical character recognition (OCR) program.
2. The OCR is revised (if the text was scanned) and all non-translational material such as page numbers, pictures and diagrams is removed.
3. Marks for titles, foreign words and expressions, emphasis and translators' notes are introduced so that these elements can later on be retrieved automatically.
4. Source text and translation are aligned in a way that enables the text-retrieval software to interpret which parts of the source text and the translation match.
5. The texts are automatically encoded so that they can operate within the IMS Corpus Workbench system.

V. Alignment Problems

Aligning source texts and translations is not a simple task, for translators do not always translate texts in a predictable and linear manner. Source-text sentences are sometimes divided into two or more sentences in the translation. Translators may also join source-text sentences together, rendering them as a single translation sentence, or they may leave things out and insert elements that were not present in the source text. In addition to this, translators sometimes reorder elements so that the order in which they appear in the translation differs from that in which they appear in the source text. The way these problems have been dealt with in COMPARA is described below.

VI. Aligning texts in COMPARA

The basic unit of alignment in COMPARA is the source-text sentence. Whenever there is not a one-to-one sentence correspondence between source and translation, it is the translation that is split or joined up to conform to the way sentences were originally divided in the source text. Thus an alignment unit is always one orthographic sentence in the source text and the corresponding text in the translation, whether it is one, more than one, or even only part of a

sentence. Source-text sentences that have been left out of the translation are aligned with blank units. Sentences that have been added to the translation with no corresponding text in the original are fitted into the nearest preceding alignment unit. Figure 1 below summarizes these alignment criteria.

Figure 1: COMPARA criteria for text alignment

SOURCE		TRANSLATION
S	→	S
S	→	S,S
S	→	½ S
S	→	ø
S	→	S(+S)

Apart from the above, if there are any sentences that have been reordered in the translation, they are aligned with the sentences that prompted them in the source texts.

One of the advantages of aligning the corpus in this way is that, as the source texts in COMPARA are always divided in the same way, it is possible to align a source text with multiple translations and compare not only source text and translation, but also different translations of the same source, in which case the source text can act as a common denominator to several translations. In addition to this, this alignment procedure enables one to search automatically for translational discourse changes such as where and when translators have decided to join, split, delete, add or reorder sentences. It is important to note, however, that it is not possible to automatically retrieve the addition or deletion or reordering of units smaller than the sentence such as individual words, phrases and clauses.

VI. COMPARA in May 2001

Preparing texts for the corpus is a time-consuming task. The COMPARA corpus project began in mid-October 1999, and ten pairs of texts had been fully processed at the time this paper was written. The part of the corpus that is available for research in May 2001 is summarized in table 1 below.

Table 1: Composition of COMPARA in May 2001

COMPARA May 2001	**Portuguese language**	**English Language**	**Total**
Source Texts	7	2	9
Translations	2	8	10
Words	91,142	99,911	191,053

The above figures mean that in May 2001 COMPARA is still a relatively small corpus. Although small corpora are not recommended for lexicographic studies, syntactic analyses do not require very large corpora (Biber, Conrad and Reppen, 1998). COMPARA is already a reasonably good-sized corpus for comparing certain aspects of Portuguese and English syntax, but still has limitations with regard to contrastive studies of Portuguese and English lexis. In addition to this, because for now COMPARA contains only fiction texts, words and expressions that do not belong to the language of fiction cannot be expected to be found in the corpus.

VII. Using COMPARA

COMPARA can be accessed free of charge at http://www.portugues.mct.pt/COMPARA/. Its Web interface, DISPARA, has been developed in collaboration with the Computational Processing of Portuguese project, and serves as a bridge between the IMS Corpus Workbench software and the specific requirements of COMPARA. Two search options are available in DISPARA. The Simple Search was made for people who have never used corpora before. It allows users to search the entire corpus either in the Portuguese-English or in the English-Portuguese direction. The instructions on how to conduct a Simple Search are extremely simple. Users only have to write a word or expression in English or Portuguese and press the search button. No special training is required.

The Complex Search was made for those who find the Simple Search too restrictive and want to conduct more sophisticated queries. We have endeavoured to make the Complex Search as user-friendly as possible, so that newcomers to corpus studies should feel confident enough to exploit its potentialities. Users are guided through the following four search steps:

1. Users are asked to choose their search direction. As in the Simple Search, they can search from Portuguese to English or from English to Portuguese. However, in the Complex Search,

instead of searching the whole corpus, users can also tell the system that they only want to search from source-texts to translations, or only from translations to source texts. It is an option to consider if the directionality of translation is relevant to a particular query.

2. Users are asked if they want to narrow down the corpus, and, if so, they are asked to choose which texts within the corpus you want to use. This is a very important step because, as COMPARA is an open-ended corpus, it is here that users will be able to control which texts they are going to use if their queries require a balanced corpus or a specific subset or other of the corpus. COMPARA can be automatically narrowed down so as to search only within specific varieties of Portuguese and English. It is also possible to narrow down the corpus by date of publication. Users who are not interested in non-contemporary language, for example, can automatically remove source texts and translations published before a particular date. The third narrowing-down option available allows users to select any manual combination of texts. Users can determine exactly which texts they want to use for their search queries, and create their own, tailor-made sub-corpora of COMPARA. They are thus able to conduct searches within texts by only one particular author, translator, group of authors, and so on. Eventually, when other genres are added to the corpus, there will also be an option that allows users to select texts automatically by genre.

3. Users can select how they want their results to be presented. The options available include concordances, distribution of forms, distribution of sources (how a search expression is distributed in the texts within the corpus) and a quantitative wrap up (the distribution of the search expression in the two languages, for searches that involve alignment constraints - see below).

4. Users are asked to enter their search queries. The IMS Corpus Workbench syntax[2] can be used here to refine searches so as to include in a single query access to different spellings of a word (for example, *analyse* and *analyze*), different morphological variants of a word (for example, *walk, walked, walks*, etc.), a word and a collocate with any number of elements in between (for example *make* and *decision*), and so on.

In addition to a search word or expression, in the Complex Search it is also possible to enter an alignment constraint. For example, users searching for the Portuguese translation of *even*,

which is often rendered as *até*, can retrieve just the cases in which *even* is translated into *até* or just the cases in which *even* is translated into something other than *até*.

Some searchable features that are very specific to COMPARA are already directly available through the DISPARA interface. DISPARA allows users to inspect the translators' notes and alignment properties associated with each search string. In addition to this, users can leave the space for entering a search string blank and search directly for translators' notes, emphasis, foreign words and expressions, and titles. And because of the way the texts in COMPARA have been aligned and encoded, it is also possible to inspect when and where translators have decided to join, separate, delete and add sentences to the translation. The possibility of looking at reordered sentences was not yet operational at the time this paper was written.

VIII. Search results

The users of COMPARA are welcome to use the results of their search queries for research and education.

The concordances are displayed in two vertical columns, with the Portuguese or English search item appearing in bold on the left-hand side, and the corresponding text in English or Portuguese on the right-hand side. The context within which the results appear is one full source-text sentence and the corresponding text in the translation.

Next to each parallel concordance displayed, there is a link to the full reference of the pair of texts from where the parallel concordance was retrieved. When looking up a reference, users also get information on copyright, on language variety, and on the number of words and alignment units for the extracts in question.

IX. Using COMPARA in language learning and translation training

The first thing to do in order to use COMPARA in language learning and translation training. is to identify ways in which comparing and contrasting English and Portuguese might be useful to students. As argued in (Frankenberg-Garcia 2000a), it doesn't make sense to use a parallel corpus to make students aware of language differences that do not affect their learning. Sometimes, however, Portuguese learners of English and English learners of Portuguese find it difficult to establish clearcut boundaries between the two languages.

Frankenberg-Garcia and Pina (1997) have identified a number of problems of crosslinguistic influence that are common among Portuguese learners of English. In Frankenberg-Garcia (2000b), reference is made to a a couple of language transfer problems that are typical of English learners of Portuguese.

Having identified aspects of Portuguese and English which tend to get mixed up, COMPARA can be very useful in helping to unmix them. Appendix 1 contains a cloze exercise adapted from COMPARA search results in which students are required to give the Portuguese translation of *even*, a word susceptible of creating confusion among native speakers of Portuguese given its different meanings and translations.

To prepare the exercise, a search for *even* was carried out in COMPARA and the results were saved as an html file. The file containing those results was opened directly from within Word (which automatically converts the file into a Word document). It was then very simple to edit the results. The *table* menu was used to delete the row where the corpus reference links appear (which are not necessary for the exercise), and Word's *replace* function was used to replace the Portuguese translations of *even* (*até, mesmo, sequer* and *ainda*) by blank spaces. The exercise helps students understand the different meanings and translations of *even*.

Another very common problem of Portuguese-English crosslinguistic influence involves the use of negative prefixes. Appendix 2 contains the first page of a worksheet prepared out of COMPARA in order to help students understand different uses of negative prefixes in English and Portuguese. Based on the principle of data-driven learning (Johns, 1991), students are asked to look at the concordances extracted from COMPARA and underline the Portuguese words that correspond to English words beginning with the negative prefix *un*. The exercise helps students realize that negative prefixes seem to be used much more sparingly in Portuguese, and that translators use different strategies to deal with them. Out of the 57 occurences of English words beginning with the negative prefix *un* contained in the exercise, less than half were translated into a word containing a Portuguese negative prefix (i.e., *i, im, in* or *des*). About a third were translated into a word preceded by a negative particle (i.e., *não, sem, difícil de* or *pouco*). In seven cases the translation was an antonym of the root of the English word. For example, a word like *untrue* was translated into an antonym of *true*, i.e., *false.* And in three cases the English word was simply not translated.

The above are just two examples of the many of ways COMPARA can be used in second language learning, teaching and translator training. As the corpus is free and accessible to all those who have an Internet connection, learners can also be trained to look things up for themselves in COMPARA.

X. Conclusion

COMPARA first became available less than a year ago and still has a long way to go. It is hoped that feedback from users will contribute towards the development of COMPARA and that this might take place alongside a growing interest in the use of corpora for research and education.

Notes

1 For a full regularly updated list of copyright permissions, see:
http://www.portugues.mct.pt/COMPARA/CorpusContents.html

2 The IMS Corpus Workbench Syntax is explained in:
http://www.ims.uni-stuttgart.de/CorpusWorkbench/CPQSyntax.html
http://www.ims.uni-stuttgart.de/CorpusWorkbench/CPQExamples.html

References

Bank of English. Available at http://titania.cobuild.collins.co.uk/boe_info.html

Biber, Douglas, S. Conrad & R. Reppen (1998) *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

British National Corpus. Available at http://info.ox.ac.uk/bnc/what/index.html

CETEMPúblico. Available at http://cgi.portugues.mct.pt/cetempublico/

COMPARA, the Portuguese-English Parallel Corpus. Available at
http://www.portugues.mct.pt/COMPARA/

Computational Processing of Portuguese Project. Available at http://www.portugues.mct.pt

Frankenberg-Garcia, Ana (2000a) 'Using a Translation Corpus to Teach English to Native Speakers of Portuguese' in *Op.Cit. - A Journal of Anglo-American Studies* Vol. 3, 65-78.

Frankenberg-Garcia, Ana (2000b) 'Using a translation corpus to sort out Portuguese-English crosslinguistic influence' Seminar presented at the Oxford University Language Centre. February 2000.

Frankenberg-Garcia, Ana and Pina, M.F. (1997) ‘Portuguese-English Crosslinguistic Influence’ in *Actas do XVIII Encontro da Associação Portuguesa de Estudos Anglo-Americanos*, Instituto Politécnico da Guarda Vol. 1, 69-78.

IMS Corpus Workbench. Available at http://www.ims.uni-stutgart.de/CorpusWorkbench/

Johansson, Stig (1998) ‘On the role of corpora in cross-linguistic research’. In Stig Johansson & S. Oksefjell (eds.) *Corpora and crosslinguistic research: theory, method and case studies*, Amsterdam: Rodopi, 3-24.

Johns, Tim (1991) ‘Should you be persuaded: two examples of data-driven learning’ in *ELR Journal* Vol.4, 1-16.

APPENDIX 1

INSTITUTO SUPERIOR DE LÍNGUAS E ADMINISTRAÇÃO - CURSO DE TRADUÇÃO

Prof. Doutora Ana Frankenberg-Garcia

Handout based on *COMPARA http://www.portugues.mct.pt/COMPARA/ [15-May-2001]*

Fill in the gaps with an appropriate Portuguese translation for *even*:

I had the ideas; I **even** made notes.	Tinha as ideias; _____ coligi notas.
They had planned the trip in detail, had their hair specially curled for the occasion, and had **even** stolen flowers for the girls.	Tinham planeado a visita em pormenor, tinham ondulado especialmente o cabelo para a ocasião, e _____ tinham roubado flores para as raparigas.
Memories came out of hiding, but not emotions; not **even** the memories of emotions.	Surgiam as recordações, mas não as emoções; nem _____ recordações de emoções.
The other rooms contained medical instruments of the eighteenth and nineteenth centuries: heavy metal relics coming to sharp points, and enema pumps of a calibre which surprised **even** me.	As outras salas tinham instrumentos médicos dos séculos XVIII e XIX: pesadas relíquias de metal que terminavam em pontas agudas e irrigadores de um calibre que _____ a mim me surpreendia.
She keeps the adored relic beside her, and **even** takes to saying her prayers while kneeling before him.	Guarda junto de si a relíquia adorada e começa _____ a dizer as suas orações ajoelhada na sua frente.
A cheeky bird, inducing affection, **even** reverence.	Um pássaro atrevido, que suscitava afecto, respeito _____.
All that remains of Flaubert's residence is a small one-storey pavilion a few hundred yards down the road: a summer house to which the writer would retire when needing **even** more solitude than usual.	Tudo o que ficou da residência de Flaubert é um pequeno pavilhão a poucas centenas de metros da estrada: uma casa de Verão para onde o autor se retirava quando precisava de _____ maior solidão do que a habitual.
Then I realised the fallacy in this: Flaubert, after all, hadn't been given a choice of parrots; and **even** this second one, which looked the calmer company, might well get on your nerves after a couple of weeks.	Depois descobri a ironia disto: Flaubert, apesar de tudo, não tinha podido escolher o papagaio; e _____ o segundo, que parecia uma companhia mais calma, podia muito bem tornar-se irritante depois de umas semanas.
In early manhood he is extremely attractive to women and his speed of sexual recuperation is, by his own account, very impressive; but **even** in later life his courtly manner, intelligence and fame ensure that he is not unattended.	Nos primeiros anos da juventude é extremamente atraente para as mulheres e a sua velocidade de recuperação sexual é, segundo ele próprio diz, impressionante; mas _____ mais tarde os seus modos corteses, a sua inteligência e fama asseguram-lhe companhia.

APPENDIX 2

INSTITUTO SUPERIOR DE LÍNGUAS E ADMINISTRAÇÃO - CURSO DE TRADUÇÃO

Prof. Doutora Ana Frankenberg-Garcia

Handout based on *COMPARA http://www.portugues.mct.pt/COMPARA/ [10-May-2001]*, using extracts from:

Julian Barnes

1984 *Flaubert's parrot* Corpus text based on London: Picador, 1985, pp 11-65.
1988 *O papagaio de Flaubert* translated by Ana Maria Amador. Corpus text based on Lisboa: Queztal Editores, 1988, pp 11-74.

Read the extracts below and underline the Portuguese word or words that correspond to English words beginning with the negative prefix *un*.

The thrower remained a stylish, temporary statue: knees not quite **unbent**, and the right hand ecstatically spread.	O jogador ficou como uma estilizada estátua temporária: os joelhos um pouco dobrados e a mão direita erguida e estática.
Let me start with the statue: the one above, the permanent, **unstylish** one, the one crying cupreous tears, the floppy-tied, square waistcoated, baggy-trousered, straggle-moustached, wary, aloof bequeathed image of the man.	Vou começar pela estátua: a de cima, a permanente, a sem estilo, a que chora lágrimas de cobre, a que lega à posteridade a imagem circunspecta de um homem com um laço desajeitado, colete quadrado, calças largas como sacos, bigode em desalinho.
If so, then how tantalising are the **unfinished** books.	Se assim é, então que excitantes são os livros inacabados.
The **unwritten** books?	Os livros que não se escreveram?
Dot, dash, dash, dash went: the concrete caissons, with the **unhurried** water between them.	Ponto, traço, traço, traço, faziam as caixas de cimento, separadas umas das outras pela água calma.
I was close to where friends had died - the sudden friends those years produced - and yet I felt **unmoved**.	Ali perto amigos meus tinham morrido - os amigos inesperados que esses anos nos dão - mas não me sentia comovido.
But here, in this **unexceptional** green parrot, preserved in a routine yet mysterious fashion, was something which made me feel I had almost known the writer.	Mas aqui, neste vulgar papagaio verde, perservado de uma maneira vulgar e no entanto misteriosa, havia algo que me fazia sentir que quase tinha conhecido o escritor.
It's about a poor, **uneducated** servant-woman called Félicité, who serves the same mistress for half a century, unresentfully sacrificing her own life to those of others.	É acerca de uma pobre criada ignorante chamada Félicité, que serve a mesma patroa durante meio século, sacrificando sem ressentimentos a sua vida à dos outros.
It's about a poor, uneducated servant-woman called Félicité, who serves the same mistress for half a century, **unresentfully** sacrificing her own life to those of others.	É acerca de uma pobre criada ignorante chamada Félicité, que serve a mesma patroa durante meio século, sacrificando sem ressentimentos a sua vida à dos outros.