

COMPARA, the Portuguese-English parallel¹ corpus

Ana Frankenberg-Garcia

Curso de Tradução, ISLA, Lisboa

This paper outlines the construction of COMPARA and demonstrates how it can be used. COMPARA is a machine-readable and searchable corpus of Portuguese-English and English-Portuguese translations. Unlike a machine-translation program, COMPARA enables its users to conduct searches within translations made by real human-beings.

COMPARA was conceived not only for experienced corpus-users, but also for people who are not necessarily corpus-literate. Language researchers can use COMPARA to compare and contrast countless different features of English and Portuguese. A translator or student of translation might find COMPARA useful to discover how different words and expressions have been translated in the past. A translation teacher might want to use COMPARA in the classroom, to tackle specific difficulties of Portuguese-English translation. Portuguese learners of English and English learners of Portuguese can use COMPARA to see how similar meanings have been expressed in the two languages. In addition to this, COMPARA may contribute towards the development of Portuguese-English machine translation programs and bilingual lexicography,

Work on building the corpus began in October 1999, and the first version of COMPARA was announced in January 2001. The corpus is can be accessed at:

www.portugues.mct.pt/COMPARA/

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 it can be automatically processed by a specific, text-retrieval program. Notable examples of monolingual corpora include the Bank of English and the British National Corpus², 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 CETEMPúblico³, 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 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

¹Parallel is being used here to refer to a bilingual collection of source texts and their translations. In the contrastive linguistics tradition, this would have been referred to as a translation corpus. Johansson (1998) predicted that the problem of conflicting terminology would eventually be resolved as the field developed and usage became more settled. This does not seem to have happened yet.

²**Bank of English** http://titania.cobuild.collins.co.uk/boe_info.html

British National Corpus <http://info.ox.ac.uk/bnc/what/index.html>

³**CETEMPúblico** <http://cgi.portugues.mct.pt/cetempublico/>

makers, developers of machine translation software and whoever else might be interested in translation language in and in the similarities and differences between Portuguese and English.

The advantage of using a corpus to compare and contrast English and Portuguese are that these 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 a corpus is made of and how it is structured.

Text Selection

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.

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 different authors and 31 translators⁴.

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.

Preparing texts for the corpus

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.

⁴ For a full regularly updated list of copyright permissions, see <http://www.portugues.mct.pt/COMPARA/CorpusContents.html>.

2. The OCR is revised (if the text was scanned), 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 system⁵.

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. Sometimes, translators join source-text sentences together, rendering them as a single translation sentence. In addition to this, translators may leave things out and insert elements which were not present in the source text, and sometimes they may reorder elements so that the order in which they appear in the translation differs from that in which they appear in the source text.

Criteria for Text Alignment

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. The figure below summarizes these alignment criteria.

COMPARA criteria for text alignment

SOURCE	→	TRANSLATION
S	→	S
S	→	S,S
S	→	½ S
S	→	∅
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 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,

⁵ <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/> - the text-retrieval software underlying the corpus. See also (Christ et al. 1999)

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.

Corpus composition 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 was available for research in May 2001 is summarized in the table below.

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 at the time this paper was written COMPARA was 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). In May 2001 COMPARA was already a reasonably good-sized corpus for comparing certain aspects of Portuguese and English syntax, but was still very limited for contrastive studies of Portuguese and English lexis. In addition to this, because COMPARA contained only fiction texts at the time, words and expressions that do not belong to the language of fiction could not be expected to be found in the corpus.

Using COMPARA

COMPARA can be accessed at www.portugues.mct.pt/COMPARA/. Its Web interface, DISPARA, has been developed in collaboration with the Computational Processing of Portuguese project⁶, 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 a corpus 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 (see appendix 1).

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 people who have never used a corpus before should feel confident enough to exploit its potentialities. Users are guided through four relatively simple search steps (see appendix 2).

Step 1

⁶ **Computational Processing of Portuguese Project** <http://www.portugues.mct.pt/>

In step one, 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. The latter is an important option to consider if the directionality of translation is relevant to a particular query.

Step 2

In step two, 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 possible to select any combination of Portuguese and English language varieties. For example, users can tell DISPARA that they want to search only Brazilian Portuguese and British English, or all varieties of Portuguese but only American English, etc.

Next, it is 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-corpus of COMPARA. They are thus able to conduct searches within texts by only one particular author, or group of authors, or translator, 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.

Step 3

The third step of the Complex Search enables users to select different displays of the results. Users can inspect 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).

Step 4

In the fourth and final step of the Complex Search, users are asked to enter their search queries. The IMS Corpus Workbench syntax⁷ 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.),

⁷ **IMS Corpus Workbench Syntax**

<http://www.ims.uni-stuttgart.de/CorpusWorkbench/CPQSyntax.html>

<http://www.ims.uni-stuttgart.de/CorpusWorkbench/CPQExamples.html>

a word and a collocate with any number of elements in between (for example *make* and *decision*), and so on⁷.

Apart from entering a given search word or expression, in the Complex Search users can also 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. Whatever the query, DISPARA allows users to inspect translators' notes and alignment properties. In addition to this, users can 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 inspecting reordered sentences was not yet operational at the time this paper was written.

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 (see appendix 3).

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.

It is possible to scroll up and down the results screen to see all the concordances displayed, and it is possible to save the results in html, text or even to cut and paste them into a word-processing program.

Conclusion

Although English language corpora are many and have been around for some time, Portuguese language corpora are still very much in their infancy. COMPARA is the first publicly accessible Portuguese-English parallel corpus. It became available less than a year ago and still has a long way to go. It is hoped that the development of COMPARA will take place alongside an increased use of corpora that include Portuguese for research and education.

References

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

Christ, Oliver, B. Schulze, A. Hofmann & E. Koenig (1999) "The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual", Institute for Natural Language Processing, University of Stuttgart, March 8, 1999 (CQP V2.2).

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

Johansson, Stig, J. Ebeling & S. Oksefjell (1999) English-Norwegian Parallel Corpus: Manual
<http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html> [Access Date 7/7/2000]

Santos, Diana (2000) "O projecto Processamento Computacional do Português: Balanço e perspectivas", in M. Graça Nunes (ed) *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada* (PROPOR 2000) [The Computational Processing of Portuguese project: balance and perspectives, in *Proceedings of the V Encounter on the computational processing of written and spoken Portuguese*], pp.105-113.