Introducing COMPARA, the Portuguese-English parallel[1] corpus

Ana Frankenberg-Garcia (ISLA, Lisbon) & Diana Santos (SINTEF, Oslo)

This paper is an introduction to COMPARA. COMPARA is a machine-searchable, open-ended collection of Portuguese-English and English-Portuguese source texts and translations. It was made for people who have never used corpora before as well as for experienced corpus users. COMPARA's encoding and alignment criteria allow users to inspect translators' notes and to investigate when and where translators have chosen to join, separate, delete, add and reorder sentences. Another innovative feature is that the corpus admits more than one translation per source text. COMPARA is encoded according to the IMS Corpus Workbench system and is freely accessible on the WWW via the DISPARA interface.

Basic characteristics of COMPARA

This paper is an introduction to COMPARA, the Portuguese-English parallel corpus. Modelling itself on the part of the English-Norwegian Parallel Corpus devoted to parallel texts (Johansson et al. 1999), COMPARA is a machine-readable and searchable collection of texts originally written in Portuguese and in English that have been aligned with their respective English and Portuguese translations.

The basic characteristics of COMPARA are that it is:

1. open-ended
2. for people who are not necessarily corpus-literate as well as for experienced corpus users
3. searchable via the Internet

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. The trade-off, of course, is that at the time this paper was written (just over one year after the project began), COMPARA did not lend itself to analyses requiring large and representative language samples.

COMPARA was made for anyone interested in the study of Portuguese and English contrasts. 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 in and in the similarities and differences between Portuguese and English.

One of our main concerns was to make sure that COMPARA could be used not only by experienced corpus users, but also by people who have never used a corpus before.

Access to COMPARA is provided free of charge at:

http://www.portugues.mct.pt/COMPARA/Welcome.html

The above site is maintained by the Computational Processing of Portuguese project, which is also responsible for distributing other Portuguese language resources apart from COMPARA[2].

Every page in the COMPARA website is available in both Portuguese and English, so that people with very little Portuguese or very little English can still read them.

Text Selection

When selecting texts for the corpus, 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.

Copyright permissions

For the initial, fiction part of the corpus, most efforts have so far been directed at obtaining copyright clearance in the Portuguese to English direction, because of the fewer Portuguese to English translations available to choose from. However, since COMPARA is to remain open-ended, obtaining a balanced corpus is not crucial. The responsibility of achieving balance is in fact being deliberately transferred to the users of COMPARA, who are expected to pre-select the texts they want to use for their search queries if and when the question of balance is important.

Although not all copyright permissions applied to were granted, the overall response from authors, translators and publishers was quite encouraging, especially when considering the fact that they actually gave permission for their texts to be freely searchable on the Internet.

At the time this paper was written, COMPARA had permission to include extracts - usually 30% of a complete work[3] - 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 translators[4].

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 of a couple of novels by David Lodge (Lodge 1975; 1995a) paired up with both their Portuguese and Brazilian translations (respectively, Lodge 1995b, 1995c, 1997, 1998), 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 (Alencar 1865), which has been paired up with a contemporary English translation published by Oxford University Press only a few months ago (Alencar 2000) and a contemporaneous translation which dates back to 1886 (Alencar 1886) - this could be interesting for a diachronic study of translation.

Corpus composition in December 2000

The COMPARA corpus project began in mid-October 1999, and very few texts had been fully processed at the time this paper was written. The part of the corpus that was available for research in December 2000 is summarized in table 1.

Table 1:Composition of COMPARA in December 2000

COMPARA December 2000	**Portuguese language**	**English Language**	**Total**
Source Texts	5	1	6
Translations	1	6	7
Words	62,039	72,402	134,441

Encoding Aims and Options

The overriding aim of all text encoding options adopted in COMPARA was to provide accurate examples of how sentences have been translated from Portuguese into English and from English into Portuguese, and, within those sentences, to provide "co-textualized" examples of how words and phrases have been translated.

Because COMPARA is a sentence-driven corpus, text divisions that lie above the level of the sentence - such as chapter and paragraph divisions - were not encoded. In fact, since COMPARA did not acquire permission for the actual physical redistribution of texts, no attempt was made to preserve the texts in a format that would allow exact future replication. Page layout, typeface, pictures, diagrams and all other material that is not immediately relevant to the study of language contrast and translation were simply removed without being replaced by omit tags.

Also, since the texts in COMPARA can only be used in COMPARA, one did not feel obliged to follow the Text Encoding Initiative (TEI) guidelines (Sperberg-McQueen & Burnard, 1994) or any other standard for text encoding to the letter, although an attempt was made to follow the TEI’s spirit and syntax whenever dealing with phenomena coped with by the TEI.

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 has been 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(s), 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. Sentences that have been reordered in the translation are aligned with the sentences that prompted them in the source texts. Thus if the order of sentences A, B and C in the source text has been changed to A, C and B in the translation, A is aligned with A, B is brought back so that it can be aligned with B, and C is aligned with C. Table 2 summarizes these alignment criteria.

Table 2: Alignment criteria

	SOURCE TEXT	TRANSLATION
Sentence preserved	1	1
Sentence split	1	2
Sentence joined	1	🗁
Sentence deleted	1	0
Sentence added	1	1 + [1]
Sentence reordered	A, B, C	A, C, B ↶

For all the above cases, special alignment markup is inserted so that corpus users can search for translational discourse changes such as when and where translators have chosen to join, split, delete, add or reorder sentences in the translation.

On the one hand, it is important to note that the alignment markup in COMPARA does not capture the addition or deletion or reordering of units smaller than the sentence such as individual words, phrases and clauses. On the other hand, the strength of this alignment procedure is that it enables one to align a source text with more than one translation, and to compare not only source and target text, but also more than one translation of the same source (where the source would, in this case, act as a common denominator to the two translations).

Text Preparation: from print to Web

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), all non-translational material is removed, and marks for titles, foreign words and expressions, emphasis and translators' notes are introduced. The notes themselves are inserted at the point where their identifiers appear.
3. Source text and translation are aligned manually, paragraph by paragraph.
4. The texts are submitted to a program of automatic tokenization and sentence separation developed by the AC/DC project (Santos et al. 2000) and to a program of automatic sentence alignment - the IMS Corpus Workbench Easyalign.
5. The alignment results are revised semi-automatically so as to conform to COMPARA's alignment criteria. Alignment markup for sentence joining and reordering is introduced manually at this point.
6. The remaining alignment markup (for sentence deletion, addition and splitting) is inserted automatically.
7. Alignment markup revision (sentence addition markup has to be discriminated from sentence splitting markup manually).
8. IMS-Corpus Workbench automatic encoding[5].

Searching COMPARA: the DISPARA interface

COMPARA can be searched via the DISPARA interface, which has been developed as a bridge between the IMS Corpus Workbench system used and the specific requirements of COMPARA. Although DISPARA was conceived to cater for the specific needs of COMPARA, it can also be very easily adapted to other parallel corpora that are encoded according to the IMS Corpus Workbench system[6].

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. A user-testing session is to take place shortly so as to provide us with feedback on what can be improved in the Complex Search option. As it stands, users are guided through four relatively simple search steps (see appendix 2).

Step 1

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 tell DISPARA 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 syntax 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, and so on[7]. Although for now users still need to learn to use the IMS Corpus Workbench syntax to have access to those details, DISPARA has plans to develop a more user-friendly interface for the types of queries that prove to be more popular among users.

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 *yes*, which is usually rendered as *sim*, can retrieve just the cases in which *yes* is translated into *sim* or just the cases in which *yes* is translated into something other than *sim*.

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 adddition 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 maximum number of concordances per query shown is 500, because many of the texts in COMPARA are still in copyright. If the corpus user chooses to narrow down the corpus in any way (e.g. by language variety, by date of publication, or by selecting a specific text pair), then the maximum number of concordances per query is further limited to 200. Whenever the results exceed these numbers, a random selection of respectively 500 and 200 concordances are presented instead. However, even when it is not possible to *show* all the concordances, the total number of solutions found is always given. Thus the user will get a message saying that, for copyright reasons, only 500 (or 200) random concordances out of the x>500 (or x>200) found can be presented.

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. Instead of a key-word-in-context (KWIC) concordance with a fixed number of characters to the left and to the right, the user gets a KWIC concordance where the context is one full source-text sentence and the corresponding text in the translation (see appendix 3). There are plans to allow the user to expand the amount of co-text given within the limits of fair-use, but this feature was not yet operational at the time this paper was written.

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

The COMPARA project began only a year ago and is still very much in its infancy. There is a long list of fiction texts for which copyright clearance has been acquired, but which are still waiting to be processed. There are plans to expand the corpus so as to include genres other than fiction, but this phase has not yet begun. The DISPARA interface to COMPARA has reached a reasonable stage of development, but feedback from users is essential to further its objectives of becoming truly user-friendly and accessible to all. This is just the beginning, and we hope some of the innovative choices made during the process of creating COMPARA and DISPARA will contribute towards future developments in the conception and distribution of parallel corpora.

Notes

[1] Parallel 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. See also the introduction of Véronis (2000).

[2] The Computational Processing of Portuguese project is concerned with the creation, evaluation, cataloguing and public distribution of Portuguese language computational resources. It is financed by the Portuguese Ministry of Science and Technology. For further information, see Santos (2000) and http://www.portugues.mct.pt.

[3] Unlike the texts in the English-Norwegian Parallel Corpus, the extracts in COMPARA are not required to be of about the same length nor taken only from the beginning of novels. The ENPC's attempt to achieve homogeneity in this respect was found to be a disadvantage by Santos and Oskefjell (1999) when attempting to validate corpus-based contrastive work.

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

[5] The IMS Corpus Workbench (Christ, 1994; Christ et al. 1999) has been singled out in previous occasions as the best corpus system available given the general context of the Computational Processing of Portuguese project, which is responsible for creating the DISPARA Web interface to COMPARA. Motivations for its use can be found in Santos (1998), and Santos & Ranchhod (1999).

[6] DISPARA is a general system for DIStributing PARAlell corpora on the Web.

[7] For a detailed description of the options available, see the IMS-CPQ User's Manual at http://www.ims.uni-stutgart.de/CorpusWorkbench/

References

Alencar, José de (1865) *Iracema,* http://www.vbookstore.com.br/nacional/josedealencar/iracema.shtml [06/12/1999]. Digital text prepared by the Biblioteca Virtual do Estudante Brasileiro, based on 24th ed.,São Paulo: Ática, 1991

----- (1886) *Iracema, the honey lips: a legend of Brazil,* London: Bickers.

----- (2000) *Iracema*, New York: Oxford 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]

Lodge, David (1975) *Changing Places*, London: Secker & Warburg.

----- (1995a) *Therapy*, London: Secker & Warburg.

----- (1995b) *A Troca* [Changing Places], Porto: Asa.

----- (1995c) *Terapia* [Therapy], Lisboa: Gradiva.

----- (1997) *Terapia* [Therapy], São Paulo: Scipione.

----- (1998) *Invertendo os Papéis* [Changing Places], São Paulo: Scipione.

Santos, Diana (1998) "Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts", in A.Rubio, N.Gallardo, R.Castro and A.Tejada (eds) *Proceedings of The First International Conference on Language Resources and Evaluation,* Vol. 1, pp.475-481.

------ & S. Oksefjell (1999) "Using a Parallel Corpus to Validate Independent Claims", *Languages in Contrast*, Vol. 2(1):117-132.

----- & E. Ranchhod (1999) "Ambientes de processamento de corpora em português: Comparação entre dois sistemas", in *Actas do IV Encontro sobre o Processamento Computacional da Língua Portuguesa* (Escrita e Falada), PROPOR [“Portuguese corpora processing: comparing two systems”, in *Proceedings of the IV Encounter on the computational processing of written and spoken Portuguese*] (Évora, 20-21 September 1999), pp. 257-268.

----- & E. Bick (2000) "Providing Internet access to Portuguese corpora: the AC/DC project", in Gavriladou M., G. Carayannis, S. Markantonatou, S.Piperidis & G. Stainhaouer (eds) *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000*, pp.205-210.

----- (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.

Sperberg-McQueen, C. & Burnard, L. (eds) (1994) “Guidelines for Electronic Text Encoding and Interchange” *TEI P3*. Association for Computers and Humanities/ Association for Computational Linguistics/ Association for Literary and Linguistic Computing. Chicago and Oxford.

Véronis, Jean (ed) (2000) *Parallel Text Processing*, Dordrecht: Kluwer Academic Publishers.