Compiling and using a parallel corpus for research in translation

ANA FRANKENBERG-GARCIA
ISLA & FCCN, Portugal

There are so many variables underlying translation that examining anything longer than a few paragraphs of translated text at a time can become quite a daunting task. The advent of corpus linguistics, however, has made it possible to analyse enormous quantities of translated text in unprecedented ways. In line with these advances, parallel corpora can provide access to many aspects of translation that had previously not been possible to study in a systematic way. The first part of this paper discusses different types of decisions that have to be made when building a parallel corpus, with particular emphasis to compilation questions that are unique to parallel corpora as opposed to corpora in general. This is followed by an account of the choices made when creating COMPARA - a post-edited, bi-directional parallel corpus of English and Portuguese literary texts with 3 million words, freely available for research and education at http://www.linguateca.pt/COMPARA/. Finally, examples of how this parallel corpus can be (and has been) used in translation research are presented.

THE STRUCTURE OF PARALLEL CORPORA

A corpus is basically a large but principled collection of naturally-occurring, authentic texts stored in digital format. A parallel corpus, in turn, is a combination of at least two sub-corpora consisting of source texts in one language (L1) and their translations into another language (L2). The two are aligned such that source texts and translations can be examined concurrently by means of parallel concordances.[1]

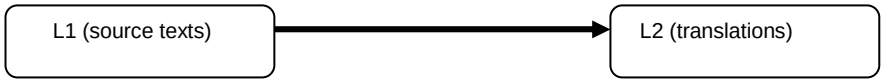

Figure 1. *Structure of a Unidirectional Parallel Corpus*

Parallel corpora can be unidirectional, bidirectional or a combination of both. The unidirectional configuration is the simplest one, with source texts in L1 and their translations into L2, as shown in figure 1. A bidirectional corpus contains source texts in two different languages (L1 and L2) aligned with their reciprocal translations into L2 and L1. This means a bidirectional structure enables researchers to analyse translations from L1 into L2 and from L2 into L1, as shown in figure 2. Parallel corpora of a mixed structure, in turn, contain a combination of unidirectional and bidirectional configurations.[2]

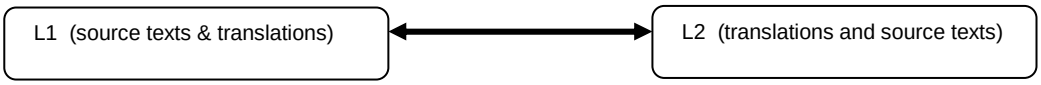

Figure 2. *Structure of a Bidirectional Parallel Corpus*

The L1-L2 alignment of unidirectional parallel corpora enables lexicographers and linguists to use them to build or improve bilingual dictionaries, computational lexicons and grammars. Professional translators and translation students can also use them to examine the different ways in which certain words or multiword segments of text have been translated.

The L1-L2 and L2-L1 configuration of bidirectional parallel corpora opens the way for a number of other analyses. To begin with, as shown in figure 3, a bidirectional structure enables researchers to develop bilingual studies from both L1 to L2 and L2 to L1 (arrows A and B). Bidirectionality can be important when, as is often the case, translation equivalents between two languages are not biunivocal: in other words, the translation of X into Y does not necessarily mean that the translation of Y will be X. As most experienced translators will know, it is not enough to reverse an L1-L2 dictionary to generate an L2-L1 lexicon. Although there may be many intersections between them, both language directions have to be considered separately in order to obtain a comprehensive picture of L1-L2 correspondences and mismatches.

If we dispense with the actual alignment, the bidirectional structure of a parallel corpus can also be utilized in studies that make use of comparable corpora of translated and non-translated texts in the same language, as depicted by arrows C and D in figure 3. The differences observed between translated and non-translated texts can help us come to a better understanding of some of the distinctive features of translation, which may include characteristics that are

considered negative, such as the phenomenon of translationese, as well as other phenomena, which are distinctive but not necessarily undesirable.[3]

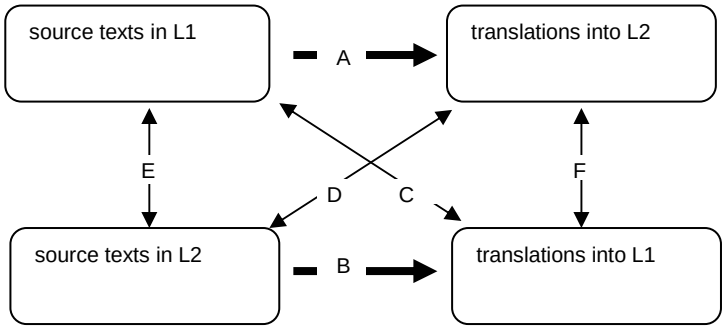

Figure 3. *Possible Directions of Analysis in a Bidirectional Parallel Corpus*

It is particularly important to note that a bidirectional structure lends itself to analyses where the findings pertaining to one specific angle of the analysis can act as a control for the findings observed along its counterpart perspective. Thus the results along the plane represented by arrow A can act as a control for the ones along the plane represented by arrow B (or vice-versa). Likewise, analyses along arrow C can be compared with analyses along arrow D. Characteristics that can be observed along complementary analyses such as these - despite intrinsic language-specific differences - may well lead to empirical evidence of translation universals[4].

Finally, the sub-corpora linked by arrow E can be used as bilingual comparable corpora, in terminological studies and contrastive linguistics, where the influence of translated language needs to be kept at bay. The sub-corpora joined by arrow F, in turn, comprise translated texts alone, and can be used in studies that examine the characteristics of translations regardless of the source texts that motivated them.

COMPILING PARALLEL CORPORA

In addition to all the factors that need to be considered when compiling a monolingual corpus - such as what genres and language varieties to include, whether to consider spoken as well as written texts, whether to use older or more recent texts, and so on - a few other decisions have to be made before setting out to build a parallel corpus. To begin with, it is necessary to decide which languages pairs are to be represented and whether their relation in the corpus is to be unidirectional or bidirectional. Then, one must also consider what kind of translations are to be included in the corpus: professional translations, learner translations, published translations, texts translated by different people, by native speakers, and so on.

The combination of the above decisions is not simple and is often oportunistic, as it will be constrained by the translated texts that are at one's disposal. Only a very small part of what people in general say or write ever gets to be translated, which seriously limits the number and types of texts available for the compilation of parallel corpora. Indeed, this is one of the main reasons why parallel corpora are usually much smaller in scale than monolingual corpora.

Another point to bear in mind is that certain language combinations are more prevalent than others. For example, while there are plenty of English-Portuguese translations in the world, not many texts get to be translated, say, from Hungarian into Portuguese. There is also an imbalance with regard to the availability of translations pertaining to different text types, genres, modes and time spans. For example, while it is fairly easy to come across translated film subtitles, translations of spontaneous speech are practically unheard of. Even when there are a fair number of translated texts pertaining to a given domain, the translations available may be mostly unidirectional. In the case of screen translation, for instance, there are many films, videos and DVDs translated from English into Portuguese, but comparatively very few translated from Portuguese into English.

Another factor that needs to be considered at an early stage of corpus compilation is whether the corpus is to be used privately, by a limited number of users, or whether it is going to be a public corpus. Although public corpora can be shared by many different people and the results obtained from them can always by verified, obtaining copyright permission to use the texts that make up a public corpus can be a time-consuming and tedious task. It must also be remembered that, for parallel corpora, double permissions are needed. It is no use obtaining copyright clearance for a source text when we are unable to secure permission to use its translation. Even if we use source texts that are in the public domain[5], it may often be the case that their translations are still protected by copyright law.

The first step to obtain permission to use a text in a corpus is to find out who, among writers, publishers, translators or even the heirs of deceased writers and translators, holds the rights to the text in question. Having done this, one of the greatest challenges of the corpus compiler is to explain to the copyright holder what a corpus is, for most people

are not familiar with corpora and fear that permission to store a copy of their texts on a server might lead to illegal downloads. It therefore is important to explain to copyright holders that the users of a corpus normally access concordances and frequency lists rather than full texts (and indeed, that it is possible to limit the amount of text retrieved from a corpus). Copyright holders can also be reassured if told that making concordances available through a corpus is a way of advertising the full text and encouraging corpus users to purchase it.

If aiming for a public corpus, it is a good idea to deal with copyright permissions before anything else. The time and effort devoted all other stages of making parallel texts searchable - digitization, mark-up, alignment, and so on - will be pointless if permission is subsequently denied.

Alignment is the one stage of corpus compilation that is unique to parallel corpora. Source texts and translations can, in theory, be aligned text by text, paragraph by paragraph, sentence by sentence, clause by clause or even word by word. The finer the alignment, the more complex it becomes. While it is fairly straightforward to align entire texts, aligning paragraphs is only trivial when paragraph structure is preserved from source text to translation. Sentence alignment further complicates the issue, because translators can (and often do) join sentences together, split a sentence into two or more smaller sentences, delete entire sentences, reorder them or even add new sentences of their own, which were not present in the source text. Clause alignment is even more complex, for it is very hard to establish clause boundaries automatically when there are no punctuation marks to set them off. Word alignment, in turn, is obviously exceedingly difficult to achieve inasmuch as languages are not translated word for word.

The level of alignment chosen will depend on the subsequent use of the corpus. Most existing parallel corpora are aligned at the level of the sentence, using one-to-many and many-to-one matches to deal with the source text and translation sentences which do not have a one-to-one correspondence. There are several automatic alignment programs capable of rendering this kind of alignment, which can then be manually revised if necessary[6].

When digitizing a text in preparation for its inclusion in a corpus, apart from header information, the only tags that are absolutely necessary in parallel corpora are alignment tags, which provide a unique identifier to a source text segment and its corresponding unit in the translation. As in any other corpus, all other mark-up and annotation is optional, and will depend on the kind of information we subsequently wish to extract from the corpus. If we want to retrieve translators' notes automatically, for example, translators' notes will have to be tagged; if we want to be able to carry out queries that involve distinguishing between nouns, verbs, adjectives, adverbs and other word classes automatically, then part-of-speech annotation will be required.

A final step involves deciding what tools will be used to manipulate the corpus. Two well-known commercial programs especially conceived for parallel corpora are Multiconcord (Wools 2000) and ParaConc (Barlow 2002). It is also possible to navigate through parallel corpora using sophisticated corpus processors such as the IMS-CWB (Christ et al 1999).

THE COMPARA PARALLEL CORPUS

This section summarizes the main decisions underlying the building of one parallel corpus in particular - the COMPARA corpus (Frankenberg-Garcia and Santos 2003). COMPARA is a bidirectional parallel corpus of English and Portuguese literary texts. Full details about the corpus, as well as free, online access to it are available at www.linguateca.pt/COMPARA/.

Structure

The bidirectional structure of COMPARA allows users to carry out studies that involve analysing:

(a) Portuguese source texts and English translations
(b) English source texts and Portuguese translations
(c) Non-translated and translated Portuguese texts
(d) Non-translated and translated English texts
(e) Non-translated Portuguese texts and non-translated English texts
(f) Translated Portuguese texts and translated English texts

Text selection

The corpus is made up of text extracts taken randomly from the beginning, middle and end of books. Extracts were preferred over full texts in order to facilitate the procurement of copyright permissions for a corpus that was to be made available online. The texts in the corpus were published between 1837 and 2002, but only 16% of them are in the public domain, with 68% of the source texts and all but one translation still being protected by copyright law.

Only published source texts and translations were admitted in the corpus. This decision was motivated by the fact that, having gone through an independent process of selection for publication and editorial revision, the texts selected should contain fewer typographical, language and translation mistakes than average. In addition to this, only direct Portuguese-English and English-Portuguese translations were considered for inclusion, in order to curb the potential effects of intermediary languages and relayed translations.

Literary texts were chosen for two reasons. The first one was to ensure bidirectionality. Although English literature in Portuguese translation is far more common than Portuguese literature in English translation, there are enough exemplars of the latter to guarantee the bidirectionality of the corpus. The second reason is that literary texts are generally known to make wider use of lexis than other genres, so a greater coverage of the general vocabulary of English and Portuguese could be obtained with a relatively small corpus. COMPARA assembles the work by original fiction writers from Angola, Brazil, Mozambique, Portugal, the United States, Britain, Ireland and South Africa, and by professional translators from Brazil, Portugal, Britain and the United States. Other varieties of Portuguese and English can be added to the corpus, but so far only the above mentioned varieties are represented.

Despite the different language varieties and wide range of dates of publication of the texts in the overall corpus, COMPARA was designed so that users can at any time restrict the corpus so as to work with a tailor-made sub-corpus consisting of texts selected according to their own specific criteria. In addition to the options available for selecting specific language varieties and publication dates automatically, users can also initiate their queries within source texts alone (excluding translations) or within translations alone (excluding source texts), or they can limit their searches to the texts of specific authors, or even select the texts they wish to analyse one by one.

The latest version of the corpus (v.10.1.5) contains around 3 million words from 72 source texts and 75 translations. The reason why there are more translations than source texts is that the corpus admits multiple translations, and three of the source texts have been aligned with two translations each.

Digitization

The texts in COMPARA were digitized such that page numbers, columns, figures, diagrams and other extra-linguistic elements that are not immediately relevant to the study of translation were removed from the corpus files. Obvious misprints detected were corrected and recorded on a separate file in case future reference to them is needed. Because of the importance of differentiating between direct speech in literary texts, direct speech markers and punctuation that might be confused with such markers were standardized to ensure that direct speech could be analysed consistently. Thus Portuguese *travessões* or m-dashes are rewritten as double hyphens (--), while hyphens and bullets receive the n-dash mark (-); double quotes are marked («) to open and (») to close and single quotes are represented by the grave accent (`) to open and the acute accent (´) to close, while apostrophes are rewritten as single, non-directional quotation marks (').

Mark-up

Only very light mark-up that was felt to be relevant to translation studies was introduced during the digitization process. This comprised authors' and translators' notes, and text segments that were highlighted (in capital letters, bold, italics, a different type of font or by indentation) in print editions. The latter differentiates between titles, foreign words, named entities, emphasis and changes of voice in the narrative.

Annotation

Although the texts in COMPARA had not been annotated when the corpus was first made available to the public in 2001, grammatical annotation was introduced in 2005 for Portuguese and in 2008 for English. The Portuguese texts in COMPARA were annotated with the PALAVRAS parser (Bick 2000) and the English ones with CLAWS (Garside & Smith 1997). The output of both parsers is currently being revised manually. Another recent feature of the corpus is semantic annotation. In 2007, the semantic field of colour was introduced using a lexically-driven approach followed by human revision.

Alignment

Unlike most parallel corpora, which do not distinguish between one-to-many and many-to-one alignment, the basic unit of alignment in COMPARA is the source-text sentence. Whenever there was not a one-to-one sentence correspondence between source and translation, the translation sentences were split or joined together with adjacent sentences to match the way sentences were originally divided in the source text. Thus an alignment unit in COMPARA 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 were left out of the translation were aligned with blank units. Sentences that were added to the translation with no corresponding text in the original were fitted into the nearest preceding alignment unit. The sentences that were reordered in translation follow the same alignment rules, with the reordering being marked separately. This means that the alignment is directional (always from source text to translation) and can be one-to-many or one-to-part or one-to-zero, but cannot be many-to-one. The alignment was carried out automatically (using the IMS Corpus Workbench EasyAlign 1.0 tool) and subsequently edited so as to conform to the directional alignment criteria described above.

The directional alignment of COMPARA facilitates the alignment of source texts with multiple translations and the comparison of not only source texts and translations, but also of different translations of the same source, with the source text acting as a common denominator to several translations. In addition to this, the 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.

Encoding and access
The corpus is encoded into the IMS Corpus Workbench format (Christ et al. 1999) and can be searched online at www.linguateca.pt/COMPARA. COMPARA is free and can be used for research and education by anyone who has an internet connection. The corpus interface was conceived to cater for the needs of both experienced users wishing to carry out sophisticated queries and novice users who have never used corpora before. Both a Portuguese and an English-language service is available - so knowledge of Portuguese is not essential for those who wish to try out the corpus, as users with little or no Portuguese can select the English interface and carry out searches in English. Concern with usability has meant the interface has undergone several improvements since it was first launched in 2001.[7] Full details about all the steps involved in the building of COMPARA are available on the corpus website, at http://www.linguateca.pt/COMPARA/construcao_compara.php.

USING COMPARA IN TRANSLATION RESEARCH

As noted in the first part of this paper, the L1-L2 alignment of parallel corpora enables researchers to build or improve bilingual dictionaries, computational lexicons and grammars, by examining the different ways in which certain words or multiword units have been translated. One of the most common uses of a parallel corpus is the analysis of the different translations of a polysemous word. Ribeiro & Dias (2005), for example, compared the human translations of the Portuguese adjective *grande* into English in COMPARA with the equivalent machine translation output rendered by Babel Fish. They were then able to identify some of the limitations of the latter and suggest ways of overcoming them. Similarly, Specia et al. (2005) and Oliveira-Netto (2005) have used COMPARA to develop word-sense disambiguation modules to be utilized in the improvement of Portuguese-English machine translation programs.

In order to find out what the more frequent English translations of a given Portuguese word in COMPARA are, it is best to begin by restricting the corpus such that only Portuguese source texts and English translations are used for the query[8]. This restriction is advisable when using a bidirectional corpus like COMPARA because, as mentioned in the beginning of this paper, the translation of a word is not necessarily biunivocal. By restricting the corpus in this way, we exclude from the analysis all the back-translations from Portuguese into English. If we then carry out a search for a polysemous word like *tempo*, for example, we obtain parallel concordances like the ones presented in figure 5. From the partial results illustrated in that figure, we can see that *tempo* was translated five times into *time* and one time each into *weather*, *moment* and *while*.

PBJS1(1220):	Não temos muito **tempo**, senhor Holmes.	We don't have much time, Mr. Holmes.
PBJS1(1263):	-- Sei que o senhor tem a melhor das intenções, mas posso lhe afirmar que não temos **tempo** para praticar nenhuma cerimônia de iniciação.	«I know that you have the best of intentions, but I can assure you we have no time to practice any kind of initiation ceremony.»
PBJS1(1372):	Enfrentando o mau **tempo**, um sem-número de entusiastas acompanhou o carro que levava a Divina ao Grande Hotel depois do espectáculo, numa estrondosa ovação, e os gritos de «Viva Sarah Bernhardt» e de trechos da *Marselhesa* ecoaram por todas as ruas até de madrugada.	Braving the bad weather, innumerable admirers, with a thundering ovation, had accompanied the carriage that took the Divine One to the Grande Hotel after the show. Cries of *« Viva Sarah Bernhardt !»* and passages of the *Marseillaise* had echoed through the streets until early morning.
PBJS1(1393):	Anna Candelária olhou-o por um **tempo**, como se avaliasse a possibilidade:	Anna Candelária looked at him for a moment, as if weighing the possibility.
PBMA1(31):	Não foi; deixou-se ficar, algum **tempo**, a olhar para os móveis.	He didn't go. He allowed himself to stay there for a while, gazing at the furniture.
PBMA1(220):	Rubião fiou do **tempo** que este projeto lhe passasse, como tantos outros; mas enganou-se.	Rubião was positive that with time this project would pass like so many others, but he was mistaken.
PBMA1(255):	Suportando menos a sede, Rubião pôde alcançar que bebesse leite; foi a única alimentação por algum **tempo**.	He was bothered more by thirst. Rubião managed to get him to drink milk. It was his only nourishment for some time.
PBMA1(290):	O santo e eu passamos uma parte do **tempo** nos deleites e na heresia, porque eu considero heresia tudo o que não é a minha doutrina de Humanitas; ambos furtamos, ele, em pequeno, umas pêras de Cartago, eu, já rapaz, um relógio do meu amigo Brás Cubas.	The saint and I have spent a portion of our time in pleasures and heresy, because I consider heresy everything that isn't my doctrine of Humanitas. We've both stolen things, he, as a boy, some pears in Carthage, I, a young man already, a watch from my friend Brás Cubas.
	Item, impunha-lhe a condição, quando morresse o cachorro, de lhe dar sepultura decente em terreno próprio, que cobriria de flores e	Item, the condition is imposed that when the dog dies it is to be given decent burial in its own plot, which will be covered with

Figure 5. *Parallel Concordances for* "tempo" *in COMPARA 10.1.4*

In order to obtain a more complete picture of the correspondences between *tempo* and these four possible translations, we can then look them up as alignment restrictions for *tempo* and request combined distributions of the Portuguese and the English search expressions. The results obtained are summarized in figure 6.

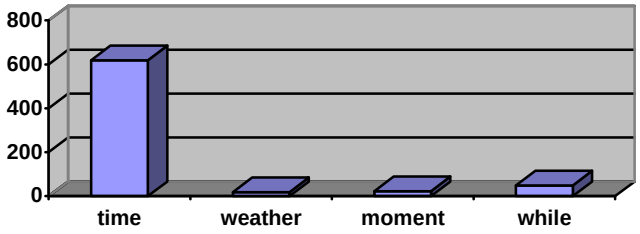

Figure 6. *Results for* "tempo" *aligned with* "time", "weather", "moment" *and* "while" *in COMPARA 10.1.4*

Note that to interpret these results as if they were translation equivalents, one must allow for a small error margin resulting from the fact that the corpus is not aligned at the level of the word. That is to say, the results in figure 6 summarize the number of occurrences of *tempo* on the Portuguese side of the concordance that match *time*, *weather*, *moment* and *while* on the English side of the parallel concordance. Although this will generally mean that one word has been translated into the other, this may not always be true. For example, in the concordance below, we find *tempo* on the Portuguese side and *weather* on the English side, but the latter is not the translation of the former: the translation match of *tempo* happens to be *dating*, while the source text word that motivated *weather* is *frios*. However, in 16 out of the 18 times that *tempo* and *weather* coincide in the alignment, *weather* is in fact the translation of *tempo*.

PPJSA2 (291) Já se percebeu que a casa é antiga, sem conforto, de um **tempo** espartano e bronco, quando sair para a rua, na altura dos **frios** maiores, ainda era o melhor remédio para quem não dispusesse senão de um corredor gélido onde aquecer o corpo em pequenos exercícios de marcha.	Easy to see that the house is old and lacking in comfort, **dating** from more spartan and primitive times, when to go outdoors with the **weather** at its coldest was still the best solution for anyone who had nothing better than a freezing corridor where he could march up and down in an effort to keep warm."

Another interesting point to be made is that although *time* seems to be a more likely translation for *tempo*, with 67.3% concordance matches, the opposite is not actually true. If we look up *time* in English source texts, it will only match *tempo* in 35.8% of the Portuguese translations, confirming that translation equivalence is not biunivocal.

This is not the place, however, for a detailed study of specific lexical equivalences from the viewpoint of bilingual lexicography. I will therefore conclude with a few examples of how a bidirectional parallel corpus like COMPARA can also be used in descriptive analyses in the field of translation studies. As already mentioned in the beginning of this paper, a bidirectional corpus can also be used to compare translated and non-translated language. Thus COMPARA can be used in studies that compare both translated and non-translated English and translated and non-translated Portuguese.

There have been quite a few corpus-based studies devoted to the former. For example, using the Translational English Corpus and the British National Corpus as two comparable corpora of translated and non-translated English, Olohan and Baker (2000) found that the use of the relative pronoun *that* after reporting verbs seemed to be a lot more frequent in translations than in texts that were not translations. Using the much smaller, translated and non-translated English components of COMPARA, Frankenberg-Garcia (2002) obtained remarkably similar results, replicating Olohan and Baker's findings.

Translated and non-translated Portuguese read differently from each other too, but there do not seem to be many empirical studies comparing the two. In one of the few studies available, Frankenberg-Garcia (2008) used COMPARA to examine the distinctive distribution of lexis in translated and non-translated Portuguese. Using a top-down, corpus-driven approach, the study identified the lexical lemmas which were most markedly over and under-represented in the translations. For example, the Portuguese adverb *enfim* was very frequent in original Portuguese but conspicuously absent from the translated texts, while its synonymous English cognate, *finalmente*, was comparatively very frequent in translated Portuguese and rather unusual in original Portuguese. In addition to confirming translators' intuitions regarding distinctive lexical distributions in translated and non-translated texts, the study disclosed a number of unexpected contrasts that would not have been discernible without recourse to corpora.

Last but not least, bidirectional corpora enable researchers to carry out analyses where the findings pertaining to the translations into one language can be confronted with the equivalent findings for the opposite translation language direction in the corpus. As pointed out in the beginning of this paper, characteristics that can be observed in both sets of findings - despite the language-specific differences underlying them - may well constitute empirical evidence of

translation universals. Frankenberg-Garcia (2005) used this cross-analysis approach to examine the use of loan words in translated and non-translated texts in COMPARA and, among other things, found that irrespective of whether translating from Portuguese into English or from English into Portuguese, translators tended to treble the number of loans originally present in source texts. Frankenberg-Garcia (2009), in turn, examined the complex relation between explicitation, text length and translation. Using a balanced bidirectional sub-corpus of comparable Portuguese and English source texts and translations from COMPARA (in order to cancel out the language-dependent bias of word counts), the translations were found to be on average significantly longer than the source texts. It was concluded that the observed increase in the number of words in the translations was more likely to be due to differences between source texts and translations than due to lexico-grammatical differences between Portuguese and English, and that this supported the phenonemon of explictation.

CONCLUDING REMARKS

This paper addressed different types of decisions that have to be made when creating a parallel corpus, highlighting aspects of corpus compilation that are unique to parallel corpora. This was followed by a brief description of the choices made in the building of one parallel corpus in particular - the COMPARA corpus. The final part of the paper provided a few examples of how the corpus has been used in translation research. It is hoped that the methodologies used and the findings observed can encourage translation researchers to build new parallel corpora or use existing ones to carry out analogous studies based on different language combinations and text types.

ACKNOWLEDGEMENT

COMPARA is part of the Linguateca project, which was jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC.

NOTES

[1] Note that some scholars have used different terminology in the past, using the term translation corpora to refer to parallel corpora. The standard term used today by most researchers in the area appears to be parallel corpora.

[2] In multilingual parallel corpora, the relation between the various language pairs upon which they are based can also be unidirectional or bidirectional. One of the most renowned corpora of this kind is the Oslo Multilingual Corpus (c.f Johansson 2007).

[3] See Baker (1993) for a discussion of the distinctive features of translated texts.

[4] See Maurenen & Kujamäki (2004) for studies that look into translation universals.

[5] Texts by writers who have died more than 70 years ago are considered to be in the public domain.

[6] A well-known example is the Translation Corpus Aligner (Hofland & Johansson 1998).

[7] Further details about a log-based, usability study of COMPARA are available in Santos & Frankenberg-Garcia (2007).

[8] This can carried out in the Advanced Search mode. For information on how to use the Advanced Search in COMPARA, see <http://www.linguateca.pt/COMPARA/docum/Tutorial.pdf>

REFERENCES

Christ, O., 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). Available online: < http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/>

Baker, M. 1993. Corpus linguistics and translation studies. Implications and applications. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair* (pp 233-250). Amsterdam: John Benjamins.

Barlow, M. 2002. ParaConc: Concordance software for multilingual parallel corpora. In proceedings of *LREC-2002: Third International Conference on Language Resources and Evaluation,* (pp.20-24) Las Palmas.

Bick, E. 2000. *The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus: Aarhus University Press.

Frankenberg-Garcia, A. 2002. Using a parallel corpus to analyse English and Portuguese translations. Paper presented at *Translation (Studies): a crossroads of disciplines*, University of Lisbon, 14-15 November.

Frankenberg-Garcia, A. 2005. A corpus-based study of loan words in original and translated texts. In P. Danielsson & M. Wagenmakers (eds.) *Proceedings of the Corpus Linguistics 2005 conference*, Birmingham, UK, 14-17 July.

Frankenberg-Garcia, A. 2008. 'Suggesting rather special facts': a corpus-based study of distinctive lexical distributions in translated texts. *Corpora*, vol. 3.2 (in press)

Frankenberg-Garcia, A. 2009. Are translations longer than source texts? A corpus-based study of explicitation. In Beeby, A., Rodríguez, P. & Sánchez-Gijón, P. (eds.) *Corpus use and learning to translate (CULT): An Introduction.* Amsterdam: John Benjamins (in press)

Frankenberg-Garcia, A. & Santos, D. 2003. Introducing COMPARA, the Portuguese-English parallel translation corpus. In F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in Translation Education*, (pp. 71-87) Manchester: St. Jerome Publishing.

Garside, R., & Smith, N. 1997. A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech & A. McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, (pp. 102-121) Longman: London.
Hofland, K. & S. Johansson 1998. The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In S. Johansson & S. Oksefjell (eds.) *Corpora and Cross-linguistic research. Theory, Method and Case Studies* (pp 87-100). Amsterdam: Rodopi.
Johansson, S. 2007. *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies.* Amsterdam: John Benjamins.
Mauranen, A. & P. Kujamäki (eds.) 2004. *Translation Universals. Do They Exist?* Amsterdam: John Benjamins.
Olohan, M. & Baker, M. 2000. Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1/2, 141-158.
Ribeiro, G. & Dias, M.C. 2005. Two Corpus-based Studies on the Translation of Adjectives in English and Brazilian Portuguese, In P. Danielsson & M. Wagenmakers (eds.) *Proceedings of the Corpus Linguistics 2005 conference*, Birmingham, UK, 14-17 July.
Santos, D. & Frankenberg-Garcia, A. 2007. The corpus, its users and their needs: a user-oriented evaluation of COMPARA, *International Journal of Corpus Linguistics* 12, 335-374.
Specia, L., Nunes, M.G.V. & Stevenson, M. 2005. Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation, *Recent Advances in Natural Language Processing* (RANLP-2005), Borovets, Bulgaria, 21-23 September.
Wools, D. 2000. From purity to pragmatism; user-driven development of a multilingual parallel concordancer. In S. Botley, A. McEnery & A. Wilson (eds.) *Multilingual Corpora in Teaching and Research.* Amsterdam: Rodopi.

ANA FRANKENBERG-GARCIA
INSTITUTO SUPERIOR DE LÍNGUAS E ADMINISTRAÇÃO
ESTRADA DA CORREIA 53
1500-210 LISBOA
PORTUGAL