Using a Parallel Corpus in Translation Practice and Research

Ana Frankenberg-Garcia
Instituto Superior de Línguas e Administração, Lisboa, Portugal
ana.frankenberg@sapo.pt

Abstract

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. Using the technology of corpus linguistics, however, it is possible to analyse enormous quantities of translated text in unprecedented ways. A parallel language corpus, i.e., a computerized collection of texts in one language aligned with their translations into another language, can provide automatic access to countless features of translated texts that up to now have not been possible to study in a systematic way. COMPARA, a translation tool developed by Linguateca [1], is the largest public, edited online parallel corpus of English and Portuguese in the world. In its current version 7.04, it provides access to almost three million words of original and translated fiction published in Portuguese and English. The aim of this presentation is to offer a brief description of the corpus and to demonstrate how it can be used in translation practice and research.

Key words: parallel corpora, Portuguese-English, translation.

A brief introduction to the COMPARA corpus

COMPARA is an extensible bidirectional parallel corpus of English and Portuguese. At present, version 7.04 contains 69 extracts of published fiction (dating from 1837 to 2000) by 33 different authors from Portugal, Brazil, Mozambique, Angola, the United Kingdom, the United States and South Africa. These texts are aligned with 72 published translations dating from 1886 to 2002, totalling almost three million words in July 2006. Only published texts, and only English translated directly from Portuguese and Portuguese translated directly from English are admitted in the corpus. Although the present corpus contains only fiction, plans are currently underway to add a non-fiction section to COMPARA [2].

---

[1] The corpus is free and is available at http://www.linguateca.pt/COMPARA/. The project has received funding from the Fundação para a Ciência e Tecnologia and from the Ministério da Ciência e Ensino Superior/POSI (POSI/PLP/43931/2001).

[2] For an updated account of the corpus contents, see http://www.linguateca.pt/COMPARA/Contents.html.

Modelling itself on the bidirectional structure of the English-Norwegian Parallel Corpus [1], COMPARA allows users to analyse a lot more than just (a) the translation of Portuguese into English and (b) the translation of English into Portuguese. As shown in figure 1, it is also possible to use COMPARA to compare (c) original Portuguese with translated Portuguese, (d) original English with translated English, (e) untranslated Portuguese and English, and (f) translated and untranslated language in general.

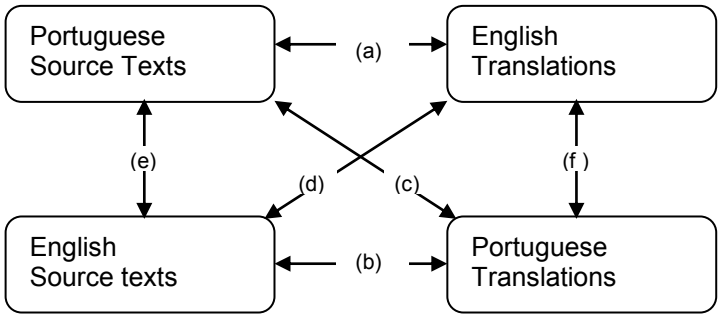


Figure 1 – Bidirectional structure of COMPARA.

COMPARA is encoded according to the IMS-CWB system [2], and can be consulted online via the DISPARA interface [3]. Access to the corpus is free and requires no registration. The service is available in English and Portuguese, and offers two different search facilities: the simple and the complex search.

The simple search enables users to retrieve parallel concordances from the entire corpus, in both the English to Portuguese and the Portuguese to English direction. As the name says, it is simple and easy to use (see figure 2). It suffices to type in a word, a prefix, a suffix, a string of words and so on in Portuguese or English and press the search button to retrieve parallel concordances containing the search term in question. Figure 3 illustrates the results for a simple search for the word *"paper"*. The search term appears on the left-hand side of the screen within a broader context of text [3] and its equivalents in Portuguese appear on the right-hand side. By clicking on the text codes on the right-hand column, users can check the source-text and translation references for each concordance line. Copyright has been cleared so that the results can be used for non-commercial, educational and research purposes.

The complex search allows users to carry out simple searches and a lot more. To begin with, the complex search allows for queries involving alignment constraints. By entering the query *"paper"* plus the alignment constraint *"papel"*, for example, users can retrieve parallel concordances containing English concordances with the word *paper* aligned with Portuguese concordances without the word *papel*. This function is useful for finding out what equivalents of *paper* other than *papel* there might be in the corpus. The complex search also allows users to retrieve translators' notes, foreign words, the titles of books, songs, plays, films (and so on) cited in the corpus texts, words that have been set off for emphasis, named entities (i.e., proper names that have been highlighted), and sentences that have been added to, deleted from, joined, split and reordered in translation. When using the complex search, users are not required to search the whole corpus, and can restrict their queries to different types of sub-corpora. For example, they can carry out searches within a specific variety of Portuguese or English, they can use texts published before or after a certain year, they can determine that searches be conducted only from source texts to translations or only from translations back to source texts, and they can query only one particular text, or the texts by

[3] The context is always one full source-text sentence and whatever matches it in the translation.

just one particular author or translator. Finally, the complex search allows users to select different output options. In addition to parallel concordances, they can retrieve the distribution of forms (for example, how frequent different words are in a particular text), the distribution of sources (in what texts different search terms appear), the distribution in original and translated texts (for example, how frequent a word is in translated and untranslated language), the distribution according to language variety (for example, how frequent a word is in European and Brazilian Portuguese) and the combined distribution in two languages (for example, in a search for *paper* and *papel* at the same time, how many times the two appear on the same concordance line and how many times *paper* does not match *papel* and *papel* does not match *paper*).

In addition to this, because the Portuguese part of the corpus has been recently annotated [4], users can also carry out queries that involve part-of-speech tags (for example, the word *casa* only as a noun, or only as a verb, or what prepositions follow a particular verb, and so on). There are plans to introduce part-of-speech annotation to the English part of the corpus in the near future.

The target users of COMPARA include not only language engineers, corpus and computational linguists, terminologists, lexicographers and machine-translation experts, but also language learners, language teachers, university lecturers, students and professional translators with little or no prior experience of using corpora. A tutorial is available at http://www.linguateca.pt/COMPARA/Tutorial.doc to help novice users become acquainted with the corpus, and the complex search facility is linked to a help file so that even non-experts are encouraged to give it a try. According to our latest statistics, the corpus has been receiving on average 6000 queries per month from Brazil, Portugal, the United Kingdom and many other countries.

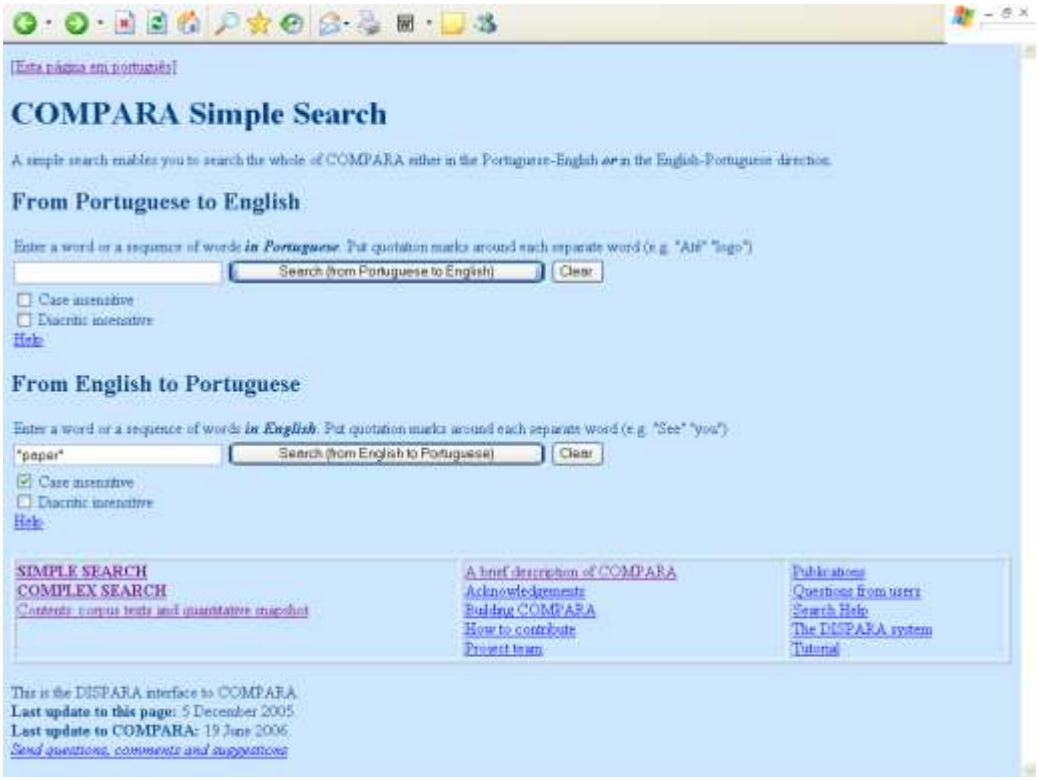
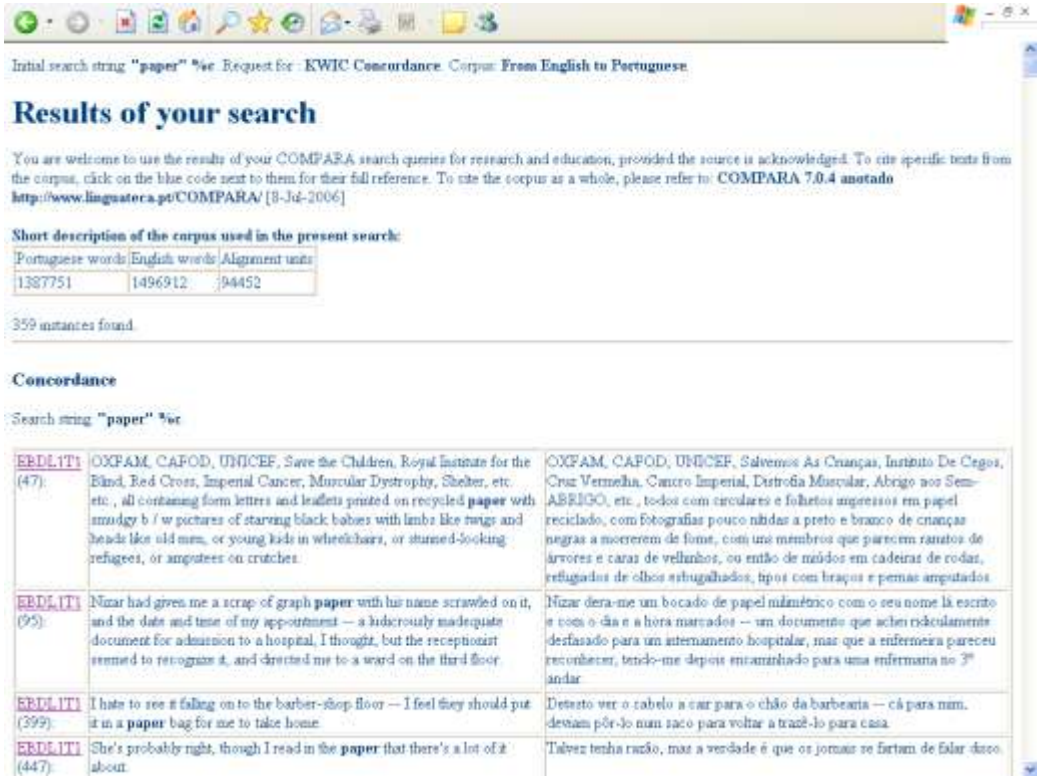


Figure 2 – The simple-search interface of the COMPARA corpus.



Initial search string: **"paper" %vr**. Request for : **KWIC Concordance**. Corpus: **From English to Portuguese**.

Results of your search

You are welcome to use the results of your COMPARA search queries for research and education, provided the source is acknowledged. To cite specific texts from the corpus, click on the blue code next to them for their full reference. To cite the corpus as a whole, please refer to: **COMPARA 7.0.4 anotado http://www.linguateca.pt/COMPARA/** [8-Jul-2006]

Short description of the corpus used in the present search:

Portuguese words	English words	Alignment units
1387751	1496912	94452

359 instances found.

Concordance

Search string: **"paper" %vr**

ERDL1T1 (47):	OXFAM, CAFOD, UNICEF, Save the Children, Royal Institute for the Blind, Red Cross, Imperial Cancer, Muscular Dystrophy, Shelter, etc. etc., all containing form letters and leaflets printed on recycled **paper** with smudgy b / w pictures of starving black babies with limbs like twigs and heads like old men, or young kids in wheelchairs, or stunned-looking refugees, or amputees on crutches.	OXFAM, CAFOD, UNICEF, Salvemos As Crianças, Instituto De Cegos, Cruz Vermelha, Cancro Imperial, Distrofia Muscular, Abrigo aos Sem-ABRIGO, etc., todos com circulares e folhetos impressos em papel reciclado, com fotografias pouco nítidas a preto e branco de crianças negras a morrerem de fome, com uns membros que parecem ramos de árvores e caras de velhinhos, ou então de mútilos em cadeiras de rodas, refugiados de olhos esbugalhados, tipos com braços e pernas amputados.
ERDL1T1 (95):	Nizar had given me a scrap of graph **paper** with his name scrawled on it, and the date and time of my appointment -- a ludicrously inadequate document for admission to a hospital, I thought, but the receptionist seemed to recognize it, and directed me to a ward on the third floor.	Nizar dera-me um bocado de papel milimétrico com o seu nome lá escrito e com o dia e a hora marcados -- um documento que achei ridiculamente desfasado para um internamento hospitalar, mas que a enfermeira pareceu reconhecer, tendo-me depois encaminhado para uma enfermaria no 3º andar.
ERDL1T1 (399):	I hate to see it falling on to the barber-shop floor -- I feel they should put it in a **paper** bag for me to take home.	Detesto ver o cabelo a cair para o chão da barbearia -- cá para mim, deviam pôr-lo num saco para voltar a trazê-lo para casa.
ERDL1T1 (447):	She's probably right, though I read in the **paper** that there's a lot of it about.	Talvez tenha razão, mas a verdade é que os jornais se fartam de falar disso.

Figure 3 – Parallel concordances for *paper* from COMPARA 7.0.4.

Observing source texts and translations

The first type of study that can be carried out with the help of COMPARA involves analysing how words, parts of words and multi-word expressions have been translated from Portuguese into English and from English into Portuguese: i.e., arrows (a) and (b) in figure 1. These studies can be used to improve bilingual dictionaries and machine-translation programs. For example, Frankenberg-Garcia [5] looks at several different Portuguese translations of the verb *nod* in COMPARA and compares the results with the verbal entries for *nod* in six different English-Portuguese bilingual dictionaries. Ribeiro & Dias [6] use the corpus to examine the English translation of *grande,* a vague and highly polysemous adjective that is often problematic in machine translation. Specia [7] and Specia et al. [8] use the corpus as a tool for word-sense disambiguation.

In translation practice and language teaching, analyses types (a) and (b) can be used as contextualized bilingual dictionaries. It suffices to enter a search expression in one language and click enter to see how it has been rendered in the other language. Figure 4 illustrates this with examples of English equivalents for the Portuguese word *festa.* Frankenberg-Garcia [9] contains examples of how COMPARA can be used in translator training and Frankenberg-Garcia [10] and [11] discuss how parallel concordances can be used in language pedagogy.

Contrasting Portuguese and English

A bidirectional parallel corpus like COMPARA can also be used as two comparable corpora of original English and Portuguese fiction texts: i.e., arrow (e) in figure 1. In what ways do the two differ? In one study, Frankenberg-Garcia [12] observed that contemporary native-English writers used over eleven times more loan words than their native-Portuguese

counterparts. Also, while the former made use of loan words from thirteen identifiable languages, the latter only used loans from English, Latin, French and German.

– Está a dar uma **festa**? – perguntou.	«Having a **party**?» he said.
… faz tempo que não vejo Karl Kroop nas **festas da faculdade**.	… you don't often see Karl Kroop at **faculty social gatherings**.
... semelhante ao som de um antigo sistema de altifalantes numa **festa de aldeia** ...	... like the sound of an old-fashioned tannoy system at an English **village fete** …
... os ciclos anuais eram marcados por **festas de família** ...	... annual cycles were punctuated by **family occasions** …
Na **festa** do Corpo de Deus, dizia ela, toda a aldeia da família se associava ...	At the **feast** of Corpus Christi, she said, her family's whole town combined …
... os cultos, as **festas**, as religiões que floresciam na sua mocidade.	… the cults, **festivals** and religions that had flowered in his youth.
Toda a vida ela sonhara a **festa**.	She had dreamed of a **reception** all her life.
... *Tu Bisvat* é o nome de uma **festa** judaica ...	... *Tu Bisvat* being the name of a Jewish **holiday** …
Ela transou com o garoto Ritchie na **festa** de Ano Novo.	She had it off with young Ritchie at the New Year's Eve **do**.
... não se queria meter em **festa** alheia ...	… not wishing to interfere in other people's **celebrations** …
... não tem graça nenhuma, mas, para o resto dos filiados, é uma **festa** ...	… is no kind of fun; but for the rest of the members it's a **ball** …
... será tudo **festa** e foguetes!	… it's all **feasting** and fun!´
... numa pequena **festa** dada pela tia ...	at a little **entertainment** given by her aunt
«É sempre Páscoa, a **festa da Passagem**» – replico.	«It is always **Passover**,» I reply.
como se tivessem consciência de estarem a participar em **festa** alheia.	as though conscious of being present at someone else's event.
Tentou estender a mão para me **fazer uma festa**.	She tried to **stroke** me with her hand
Por pouco não lhe **fiz uma festa** nas orelhas	I nearly **tickled** him under the ears.
Fiz-lhe uma **festa** na cabeça ...	I **patted** him on the head …

Figure 4 – Selected parallel concordances from Compara for *festa*.

Comparing translated and untranslated language

Bidirectional parallel corpora can also be used to compare translated with untranslated language, i.e., arrows (c) and (d) in figure 1. Intuitively, many people are aware that translated texts do not read like texts that have been originally written in a particular language. But in what ways do they differ? A very simple study that can be carried out using a corpus like COMPARA is to examine the relative frequency of certain words in source texts and translations. Figure 5 summarizes the distribution of *diferente(s), simplesmente* and the verb *rezar* in the Portuguese source texts and translations of COMPARA 7.04. As can be seen, *diferente(s)* is two times more frequent in translated Portuguese, while *simplesmente* occurs three times as often. The lemma *rezar* on the other hand, is more frequent in texts originally written in Portuguese.

	Original Portuguese	Translated Portuguese
diferentes(s)	15.4	30.7
simplesmente	5.1	15.6
lemma "rezar"	12.4	5.6

Figure 5 – Relative frequency (per 100 thousand words) of words in original and translated Portuguese.

There are many other studies that can be carried out to compare translated and untranslated language. Frankenberg-Garcia [12], for example, found that while translated Portuguese contained more loan words than original Portuguese, untranslated English had more loans than translated English. Using the British National Corpus and the Translational English Corpus, Olohan and Baker [13] found that the syntactically optional relative pronoun

that following the reporting verb *tell* occurs more frequently in translated than in untranslated English. Frankenberg-Garcia [5] replicated these findings using the translated and untranslated English component of COMPARA 2.2.

Examining the characteristics of translated texts

A fourth type of study that can be carried out using a bidirectional parallel corpus like COMPARA involves examining the characteristics of translated texts, i.e., arrow (f) in figure 1. Multidirectionality is important here, for what might be a characteristic of the translation of language X into language Y may not apply in the translation of language Y into language X, or of language X into language Z. To put it differently, in order to find out whether translated texts share certain features irrespective of the translation languages involved, it is important that more than one translation language pair be analysed. Otherwise, the results may be skewed. In a study using a purposefully balanced sub-corpus of COMPARA where an attempt was made to cancel out the language-dependent bias of word counts, Frankenberg-Garcia [14] found that translations tended to be longer than source texts in both the English-Portuguese and the Portuguese-English directions. Several other studies concerning the search for translation universals are discussed in Maurenen & Kujamäki [15].

Concluding remarks

People have different opinions about translation and remarks about what translations are or what they should be are often controversial and full of allegations based on anecdotal evidence. Without proper empirical investigation, it is not possible for theory to advance. The present paper attempted to show how a bidirectional parallel corpus and corpus techniques can be used to analyse translation data from different perspectives and for different purposes, in ways than would not have been possible before the existence of computerized language corpora. It is hoped that this brief introduction to COMPARA and to some of the ways in which it has been used in translation practice and research will encourage many other uses of the corpus.

References

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

[2] Christ, O., B. Schulze, A. Hofmann & E. Koenig, *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).

[3] Santos, D. «DISPARA, a system for distributing parallel corpora on the Web», in N. Mamede & E. Ranchhod (eds.), *Advances in Natural Language Processing: Third International Conference, Proceedings*, Berlin/Heidelberg: Springer-Verlag, 2002, pp. 209-218.

[4] Santos, D. & S. Inácio. «Annotating COMPARA, a grammar-aware parallel corpus», in N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik & D. Tapias (eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)* Genova, 2006, pp. 1216-1221.

[5] Frankenberg-Garcia, A. «Using a parallel corpus to examine English and Portuguese translations», paper presented at *Translation (Studies): a crossroads of disciplines,* Faculdade de Letras da Universidade de Lisboa, November 2002.

[6] Ribeiro, G. & M. C. Dias, «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 2005, ISSN: 1747-9398.

[7] Specia, L. «A Hybrid Model for Word Sense Disambiguation in English-Portuguese Machine Translation», *Proceedings of the 8th Research Colloquium of the UK Special-interest Group in Computational Linguistics (CLUK-05)* (Manchester, 11 January), pp.71-78.
[8] Specia, L, M.G.V. Nunes & M. Stevenson, «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 2005.
[9] Frankenberg-Garcia, A «COMPARA, language learning and translation training», in B. Maia, J. Haller & M. Ulyrch (eds.), *Training the Language Service Provider for the New Millennium*. Porto: FLUP, 2002, pp.187-198.
[10] Frankenberg-Garcia, A. «Lost in parallel concordances», in G. Aston, S. Bernardini & D. Stewart (eds.), *Corpora and language learners*. Amsterdam/Philadelphia: John Benjamins, 2004, pp.213-229.
[11] Frankenberg-Garcia, A. «Pedagogical uses of Monolingual and Parallel Concordances», *English Language Teaching Journal* 59.3, July 2005, pp.189-198.
[12] Frankenberg-Garcia, A. «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 2005, ISSN: 1747-9398.
[13] Olohan, M. & Baker, M. «Reporting *that* in translated English: Evidence for subconscious processes of explicitation?». *Across Languages and Cultures* 1(2), 2000: pp.141-158.
[14] Frankenberg-Garcia, A. «Are translations longer than source texts? A corpus-based study of explicitation», *Third International CULT (Corpus Use and Learning to Translate) Conference,* Barcelona, 22-24 January 2004 and *CIC-ISLA Working Papers*, Lisbon: ISLA, 1-2004, pp. 2-9.
[15] Mauranen, A. & Kujamäki, P. (eds.) *Translation Universals. Do they exist?* Amsterdam: John Benjamins, 2004.

BIOGRAPHICAL NOTE

Ana Frankenberg-Garcia holds a PhD in Applied Linguistics from Edinburgh University and is an auxiliary professor at ISLA, in Lisbon, where she teaches English, translation and the use of corpora in applied translation. She is joint project leader of the COMPARA parallel corpus of English and Portuguese, a public, online tool developed at Linguateca with funding from the Portuguese Foundation for Science and Technology. Her current research interests focus on the use of corpora for language learning and translation studies, parallel corpora, corpus usability and user behaviour, learner autonomy, crosslinguistic influence and second language writing