CorTrad: a multiversion translation corpus for the Portuguese-English pair

Stella E. O. Tagnin[1], Elisa D. Teixeira[2], Diana Santos[3]

Abstract

In this paper we present a new parallel corpus for the Portuguese-English pair, CorTrad, which is a joint project of University of São Paulo and Linguateca. CorTrad offers several new functionalities for translation research, such as the comparison of different versions of the same translation, from first draft to published version, and the possibility of querying specific structural parts for each text genre. For the moment, CorTrad has a 315-page book on cooking and guest entertaining, a collection of Australian and Canadian short stories, and twenty issues (2001-2003) of a Brazilian scientific news periodical.

Keywords: parallel corpora, technical translation, translator training, corpus querying, Portuguese<->English.

1. Presentation

It is uncontroversial that non-fiction parallel corpora are scarce and badly needed for teaching non-literary translation: the translation of different genres requires different skills and has different traditions that have to be followed. A parallel corpus is an abundant source of inspiration and example for future (and current) translators.

Likewise, being able to investigate improvements made to former drafts of the translated text provides an extraordinary insight into how translators and editors work, and which difficulties the language pair offers, whether their native language is the source or target language. A corpus that offers these possibilities is invaluable not only for translator training, but for contrastive linguistics as well.

The CoMET Project[4] had compiled a parallel corpus of technical texts, but was lacking a distribution infrastructure that could make it maximally useful to the community as well as to the project members themselves. On the other hand, Linguateca had developed DISPARA, a system for making available parallel corpora on the Web, considering this kind

[1] University of Sao Paulo. Projeto CoMET was financed by CNPq, grants 403120/2003-9 and 400988/2006-2.
[2] Projeto CoMET - University of Sao Paulo.
[3] SINTEF ICT, Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC, and by UMIC and FCCN.
[4] http://www.fflch.usp.br/dlm/comet

of corpus a resource of utmost importance for the computational processing of the Portuguese language. So, cooperation between both projects, which started in May 2008, has very quickly produced the CorTrad[5] service, which we will describe in some detail in what follows.

2. The CoMET project

CoMET – Multilingual Corpus for Teaching and Translation – is a multilingual corpus developed by Stella Tagnin and her group at the Department of Modern Languages, School of Humanities, University of Sao Paulo, Brazil. It has three different axes, one of which is CorTrad (translation corpora), which is the subject of the present paper.

The other two are: i) CorTec[6] - a comparable Portuguese-English technical corpus containing 15 subcorpora at present, most of them with approximately 200,000 words in each language, some with more than a million, including specialized areas such as Cooking, Ecotourism, Cultural Tourism, Business Agreements, Computer Science, Hypertension, Renal Failure, Astronomy, Electromagnetic Flowmeters, Nutritional Supplements, Coffee harvesting, processing and trading, Soccer, and Linguistics; ii) CoMAprend[7] - a learner corpus of Brazilian students learning English, French, German, Italian, and Spanish (see Tagnin 2003 and 2008 for more information).

3. The DISPARA system

DISPARA (Santos 2002) was born together with COMPARA (Frankenberg-Garcia & Santos 2003) as the technological infrastructure that allowed a user anywhere in the world with an Internet connection to query a parallel corpus. As argued in Santos (1998), to make a corpus searchable on the Web requires at least three different building blocks: the texts themselves (together with their annotation and markup), the corpus processing system (to index, align, and query efficiently) and the interface (what the user sees of the whole process, and how s/he communicates with the corpus).

DISPARA is thus a full-fledged system for making a corpus searchable on the Web, with a customizable interface, and with the IMS-CWB (Christ *et al*. 1999) as the underlying corpus processing system.

4. What CorTrad contains

[5] http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html
[6] http://www.fflch.usp.br/dlm/comet/consulta_cortec.html
[7] http://www.fflch.usp.br/dlm/comet/comaprend.html

Currently, CorTrad is freely available for search on the Web and has three different subcorpora, coming from different genres and different sources, and with specific features. DISPARA was therefore employed so that for each of these corpora the system has a slightly different behaviour, as shown in the research examples in section 7. We expect to be able, over the years, to keep on adding more texts and genres to CorTrad, to accommodate translation data in the two languages, in both directions - in that it is an open corpus.

4.1. CorTrad técnico-científico - culinária (culinary arts)

The translation of recipes, which has been experiencing growing worldwide demand in the last decades, is difficult and requires specialized knowledge of both source and target cultures (Colina 1997; Teixeira 2004, 2008; Tagnin and Teixeira 2004; see also Adab 2000 on the importance of taking cultural habits seriously). But, much as the dictionaries available for the English-Portuguese language pair may contribute to the understanding of the source text, the majority of them do not provide any information on how the terms are actually used in real texts. Thus, an English-Portuguese parallel corpus is an invaluable resource for both professional and novice translators, as well as editors.

Currently we have the texts of a complete 315-page cookbook written in Portuguese and translated into English by two Brazilian translators specialized in the area. The translated text was then revised by a native speaker of American English, also specialized in the area, and finally read by the author and both translators for some final changes. So the corpus available at CorTrad presents 4 versions of the text - original, first draft, draft revised by native speaker and published version (the latter is still being processed).

The idea is to incorporate more texts in the future, both in the Portuguese-English and in the English-Portuguese direction.

4.2. CorTrad literário (fiction)

This currently consists of a selection of Australian and Canadian short stories translated by students of the Translation Diploma Course at the University of São Paulo. The Australian stories come in four versions: original, students' first draft, revised version incorporating[8] the teacher's comments and corrections, and published translation. The Canadian stories, however, only have the original and the published translation. With these features, this corpus

[8] Not in an explicit way as in Popescu-Belis *et al.* (2002), but they can be retrieved or hinted at by direct comparison, or by searching through the CorTrad interface.

can be considered a “translation learner corpus”, highlighting not only linguistic problems, but also variant-specific cultural issues students had to deal with.

4.3. CorTrad jornalístico - divulgação científica (scientific news)

This subcorpus currently consists of texts from the journal of a leading research financing institution in Brazil, *Revista Pesquisa FAPESP*, in the Portuguese-English direction. They cover a wide variety of topics of scientific research. An interesting issue is that, while the English translation should be taken with a grain of salt, since it is performed by Brazilian translators apparently not specialized in the areas covered, non-native English in the scientific domain is pervasive and obviously deserves to be studied and documented.

In fact, translation scholars have argued (see Malmkjaer 1996) that translators who are source language natives are the best to preserve subtle shades of meaning in the original, although not so well-equipped to produce a natural text in the target language. How important this is for short news in science and technology domains, whose terminology often comes originally from English, is something that can only be assessed empirically, and CorTrad should provide an excellent testbed for this. In any case, it is widely known that published translations by professional translators still display errors and problems, and this has in fact been suggested as positive data for contrastive linguistics (Santos 2000; 2004), in addition to the uncontroversial claim that previous translations provide good examples for new ones (Isabelle 2002).

5. Corpus processing and contents: behind the scenes

Just like the other corpora made available by Linguateca, CorTrad's texts were: i) processed by Linguateca's tokenizer and sentence separator, ii) automatically marked up with relevant structural divisions, iii) parsed by PALAVRAS (Bick 2000) if in Portuguese, or tagged by CLAWS (Rayson and Garside 1998) if in English, iv) automatically aligned with the CWB-align tool, and v) encoded in CWB.

Frequency lists for each corpus were also automatically created during the corpus building process and are made available on the CorTrad site.

6. Comparison with other parallel corpora

CorTrad's closest counterpart is obviously COMPARA, since it was implemented in the same system – DISPARA – and because it is a corpus whose design and availability has influenced the Portuguese translation community. It is therefore easy to compare CorTrad with

COMPARA, in that it has most of its features (although using a different interface look and feel) and in that it tries to extend COMPARA on the few things its design was felt to need improvement, namely more than one translation for the same original, and a more flexible structuring of texts.

Of course, the main difference is that COMPARA contains exclusively published literary text, while CorTrad aims for breadth in genre. We expect that cooking, short stories and science news will just be starters. As to the contents, comparison has to take into account that COMPARA took nine years to build (and revise), while development of CorTrad has been going on for only one year. Syntactic annotation is therefore not yet humanly revised in either language.

If we compare it with the Oslo Multilingual Corpus[9] (OMC) or the English-Norwegian Parallel Corpus (Oksefjell 1999), although these were also designed for non-fiction, they did not include Portuguese non-fiction. In a way, it is also noteworthy to remark that, so far, all technical translations included in CorTrad are from Portuguese into English, whereas it is much more common to find such texts translated in the other direction.

Other parallel corpora including Portuguese (but not specifically geared towards Portuguese) are: EuroParl (Koehn 2005), the DGT Multilingual Translation Memory of the Acquis Communautaire: DGT-TM and the JRC-Acquis Multilingual Parallel Corpus[10], containing texts on European political and legal issues. None of them include the Brazilian variant of Portuguese though.

In the NATools project, also, a parallel corpus browser was developed, the NAT-QI[11]: NATools Corpora Query Interface, which serves word-aligned parallel corpora created with the NATools package (Simões and Almeida 2007). CorTrad will not provide word-alignment information in the near future.

OPUS[12] is similar to CorTrad in that it provides distribution per movie (per individual text, per domain, per author, etc. in the case of CorTrad) in the multilingual corpus of subtitles (even in a graphical way, which is something that CorTrad does not do). OPUS uses the same underlying corpus system as DISPARA.

We believe that CorTrad, although not the only one or the first, is certainly different enough and has different enough material to be seen by corpus users, especially those

[9] http://www.hf.uio.no/ilos/OMC/English/index_e.html
[10] These corpora / resources share most of the texts, but have different properties and additional information, as explained in http://langtech.jrc.it/DGT-TM.html.
[11] available from http://linguateca.di.uminho.pt/nat
[12] available from http://urd.let.rug.nl/tiedeman/OPUS/

interested in contrastive studies and in translation studies, as not just "yet another parallel corpus".

7. Examples of research with CorTrad

In this section we provide some initial examples on how to use CorTrad, in a pedagogical mood: just in terms of what can be done, and not for the sake of the results. We are assuming that readers are conversant with the usual issues that can be asked to an aligned corpus, and we will not cover them here, except for the less usual ones.

As a general parallel corpus, one can ask CorTrad for cross-distribution of a particular pattern, for example *big-grande* (original-first translation), or *grande-grande* (first-second translation) or even *big-grande-grande-grande* (in the four stages: from original to first draft, second draft and published translation).

One can also select a different point of view[13], and focus, e.g., on the final translation to see how it differs in regarding to the previous versions. It is therefore also possible to look for *furioso* (4) – not *furioso* (3). That is, cases where the adjective *furioso* was included in the final version ("4") but not in the last draft ("3").

Since the cooking parallel corpus is divided into recipes (and many other kinds of sections), one can ask for the distribution of a particular word (or phrase) per recipe. Or find out which ingredients are associated with, for example, the adjective "red": *"red"* [pos="N.*"]

On the other hand, it is also possible to restrict search to just the ingredients list, in the cooking material, or to news of a particular subject matter (or sets of subject matters) in the scientific news corpus. Conversely, one can study the distribution of lexical items, or syntactic constructions, per subject matter, for example, looking for *energ.** and its distribution per theme.

8. Final comments

We believe this resource is useful and interesting for everyone interested in the relationship between English and Portuguese and in the translation in both directions. Although CorTrad is still in its preliminary stages, in that neither has all the material been annotated or revised nor is the final display interface available, it already provides some interesting material for research.

[13] One does this in CorTrad by having always to select one of the versions as the main corpus (Principal).

References

ADAB, Beverly (2000), Cross-cultural assumptions in the translation of advertising - how realistic are these? *Across Languages and Culture* **1** (2), 193-207

BICK, Eckhard (2000), *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus: Aarhus University Press.

CHRIST, Oliver, Bruno M. SCHULZE, Anja HOFMANN and Esther 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 from: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/PDF/cqpman.pdf

COLINA, Sonia (1997), Contrastive Rhetoric and Text-Typological Conventions in Translation Teaching, *TARGET* **9**, Amsterdam: John Benjamin, 335-353.

FABRICIUS-HANSEN, Cathrine (1998), Information density and translation, with special reference to German-Norwegian-English Stig Johansson and Signe Oksefjell (Eds.), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*, Rodopi, 197-234.

FRANKENBERG-GARCIA, Ana and Diana SANTOS (2003), Introducing COMPARA, the Portuguese-English parallel translation corpus, in ZANETTIN, Federico, Silvia BERNARDINI and Dominic STEWART (Eds.) *Corpora in Translation Education*, Manchester: St. Jerome Publishing, 71-87.

ISABELLE, Pierre (1992), Bi-Textual Aids for Translators, *Screening Words: User Interfaces for Text, Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research* (Waterloo, October 18-20, 1992), 76-89.

KOEHN, Philipp (2005), Europarl: A Parallel Corpus for Statistical Machine Translation, *MT Summit.*

MALMKJAER, Kirsten (1996), Who walked in the emperor's garden: The translation of pronouns in Hans Christian Andersen's introductory passages, in ANDERMAN, Gunilla and C. BANÉR (Eds.), *Proceedings of the Tenth Biennial Conference of the British Association of Scandinavian Studies*, The University of Surrey, Department of Linguistics and International Studies.

OKSEFJELL, Signe (1999), A Description of the English-Norwegian Parallel Corpus: Compilation and Further Developments, *International Journal of Corpus Linguistics* **4.2**, 197-216.

POPESCU-BELIS, Andrei, Margaret KING and Houcine BENTANAR (2002), Towards a corpus of corrected human translations, in KING, Margaret (Ed.), *Machine Translation Evaluation – Human Evaluators Meet Automated Metrics, Workshop Proceedings, LREC'2002*, 17-21.

RAYSON, Paul and Roger GARSIDE (1998), The CLAWS Web Tagger, *ICAME Journal* **22**, The HIT-centre - Norwegian Computing Centre for the Humanities, Bergen, 121-123.

SANTOS, Diana (1998), Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts, in RUBIO, Antonio, Natividad GALLARDO, Rosa CASTRO

and Antonio TEJADA (Eds.), *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998), Vol. 1, ELRA, 475-481.

SANTOS, Diana (2000), The translation network: A model for the fine-grained description of translations, in VÉRONIS, Jean (Ed.), *Parallel Text Processing*, Dordrecht: Kluwer Academic Publishers, 169-186.

SANTOS, Diana (2002), DISPARA, a system for distributing parallel corpora on the Web, in RANCHHOD, Elisabete and Nuno J. MAMEDE (Eds.), *Advances in Natural Language Processing (Third International Conference, PorTAL 2002, Faro, Portugal, June 2002, Proceedings)*. Berlin: Springer, 209-218.

SANTOS, Diana (2004), *Translation based corpus studies: Contrasting English and Portuguese tense and aspect systems*, Amsterdam: Rodopi.

SIMÕES, Alberto and José João ALMEIDA (2007), Parallel Corpora based Translation Resources Extraction, *Procesamiento del Lenguaje Natural* **39**, 265-272.

TAGNIN, Stella E. O. (2003), COMET - A Multilingual Corpus for Teaching and Translation, *PALC 2001: Practical Applications in Language Corpora, 2003, Lodz, Poland*, Frankfurt am Main: Peter Lang - Europäischer Verlag der Wissenschaften, 535-540.

TAGNIN, Stella E. O. (2008), Disponibilização de *corpora online*: os avanços do Projeto COMET, in TAGNIN, Stella E. O. and Oto Araújo VALE (Eds.), *Avanços da Lingüística de* Corpus *no Brasil*, São Paulo: Humanitas, 95-115.

TAGNIN, Stella E. O. and Elisa D. TEIXEIRA (2004), British vs. American English, Brazilian vs. European Portuguese - how close or how far apart? – a corpus-based study, *Proceedings of the 4th Conference on Practical Applications in Language Corpora - PALC´03.* Lodz: Lodz University Press.

TEIXEIRA, Elisa D. (2004), *Receita qualquer um traduz. Será? A culinária como área técnica de tradução*, São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. [MA dissertation].

TEIXEIRA, Elisa D. (2008), *A lingüística de corpus a serviço do tradutor: proposta de um dicionário de culinária voltado para a produção textual*, São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. [PhD dissertation]. Available from: http://www.teses.usp.br/teses/disponiveis/8/8147/tde-16022009-141747/