COPLE2 – Corpus of Portuguese FL/L2[1]

Amália Mendes, Sandra Antunes, Anabela Gonçalves
Centro de Linguística da Universidade de Lisboa
{amalia.mendes, sandra.antunes}@clul.ul.pt; a.goncalves@letras.ulisboa.pt

1 Introduction

In this poster, we present COPLE2, a new corpus of Portuguese FL/L2, which encompasses written and spoken data produced by foreign learners of Portuguese at the University of Lisbon. Over the past few years we are seeing a substantial growth in the area of learner corpus research applied to other languages besides English. For Portuguese, a less commonly taught language, few corpora have been compiled[2] and it is now our aim to enhance the learning data of Portuguese L2. COPLE2 will constitute a good resource for teachers and researchers, since it will provide empirical data to: (i) identify general errors in the learning of Portuguese L2 (Granger, 1996); (ii) develop textbooks and other teaching material targeting specific groups of students; (iii) implement teacher training material by taking into account the analysis of the corrections of the teachers.

In this abstract we will briefly describe the work in progress regarding the constitution and linguistic annotation of this corpus.

2 Corpus constitution

COPLE2[3] is a new corpus compiled at the University of Lisbon that aims at collecting Portuguese learning data produced by students that attended summer/annual courses or performed accreditation exams.

Regarding the written register, the corpus includes:

- 966 free essays produced from 2010 to 2012, and a total of 156.691 words;
- 424 students, aged between 18 and 40 years (80% aged 18-30);
- 14 different mother tongues: Chinese, English, Spanish, German, Russian, French, Japanese, Italian, Dutch, Tetum, Arabic, Polish, Korean and Romanian;
- different levels of proficiency (beginner, elementary, intermediate, advanced, proficient), the most frequent being elementary (40%).
- different genres (dialogue, formal and personal letters, informative, message/e-mail, opinion, recount, book review), the most frequent being opinion on several topics (36%).

We present in Table 1 some information regarding the written subcorpus (the one encoded so far).

L1	Inf.	Male	Female	Average Age	Tests	Texts	Total Words	Average Words/Text
Chinese	129	33%	67%	21.9	277	323	57.385	178
English	65	34%	66%	24.5	118	142	21.610	152
Spanish	52	42%	58%	28.3	102	139	21.200	153
German	39	38%	68%	26.5	69	76	12.548	165
Russian	25	8%	92%	25.3	52	70	9.697	139
French	23	26%	74%	29.1	40	43	7.808	181
Japanese	23	26%	74%	23.0	45	50	6.809	136
Italian	20	30%	70%	25.2	28	34	5.875	172
Dutch	11	18%	82%	23.8	14	15	1.993	133
Tetum	9	56%	44%	31.0	19	22	3.163	144
Arabic	8	25%	75%	30.2	13	13	2.206	170
Polish	8	25%	75%	26.2	16	22	2.810	128
Korean	6	17%	83%	23.8	9	9	1.530	170
Romanian	6	0%	100%	26.0	8	8	2.057	257
TOTAL	**424**	**32%**	**68%**	**26.0**	**810**	**966**	**156.691**	**163**

Table1: COPLE2 constitution

[1] The corpus compilation is funded by Fundação para a Ciência e Tecnologia (UID/LIN/00214/2013), Fundação Calouste Gulbenkian (Proc. nr. 134655) and ADFLUL.

[2] Data Collection for Learning Portuguese as a Second Language (http://www.clul.ul.pt/en/resources/314-corpora-of-ple), PEAPLE2 (http://www.uc.pt/fluc/rcpl2/dados/) and CAL2 (http://cal2.clunl.edu.pt/).

[3] http://www.clul.ul.pt/en/research-teams/547

Concerning spoken data, the corpus will include 28 recordings (interviews collected in accreditation exams performed from 2014 to 2015 by students with the same profile), transcribed and aligned text-to-sound using the EXMARaLDA software (Schmidt, 2012).

For each text, we provide detailed metadata about the profile of the candidate and the task description (Granger et al., 2009).

3 Transcription and annotation

All the essays were handwritten and were digitalized and transcribed. The transcriptions are encoded in XML, following the Text Encoding Initiative guidelines (Burnard and Bauman, 2013) and anonymized to remove personal information (Hinrichs, 2006).

Each file is composed by a header (which comprises all the metadata) and by the transcription, which includes:

(i) all the changes made by the student during the writing process (deletions, additions, transposition segments, etc);

(ii) the correction and comments made by the teacher.

An example of a XML transcription of an essay produced by a Chinese speaker, and the final version intended by the student are given below, in (1) and (2) respectively.

(1) ** **Normalmento, Eu acordo às oito horias de manhã, **<del hand="zh010">**t**</del>** e tomo o duche e o pequeno-almoço. Eu saio de casa e apanho o metro para universidade, eu chego o escritório de XX **<del hand="corrector">**á</del> **<add hand="corrector">**às**</add>** nove de manhã. **<hi hand="corrector" rend="underlined">**Eu escrevo **<add hand="zh010">**os**</add></hi>** livros de engenheiro, ou tenho curso.** **

(2) Normalmento, Eu acordo às oito horias de manhã, e tomo o duche e o pequeno-almoço. Eu saio de casa e apanho o metro para universidade, eu chego o escritório de XX á nove de manhã. Eu escrevo os livros de engenheiro, ou tenho curso.

An interface platform (TEITOK) was designed to enable the visualization of different versions of the text (cf. Fig. 1): XML; transcription (faithful to the handwritten document); final version of the student; correction of the professor. We plan to normalize the orthography and to annotate the data with part-of-speech information and lemmatization. The next step will be to label the data following a typological scheme for error annotation (Tono, 2003; Nicholls, 2003; Dagneaux et al., 2005). We also plan to make available all these different levels through the CQP query system.

Fig. 1: TEITOK platform

References

Burnard, L. and Bauman, S. (eds.) 2013. *TEI P5: - Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium Charlottesville. Virginia.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J. and Thewissen, J. (eds.) 2005. *Error Tagging Manual. Version 1.2*. Centre for English Corpus Linguistics. Université Catholique de Louvain. Belgium.

Granger, S. 1996. "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora". In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in Contrast. Text-based cross-linguistic studies*. Lund Studies in English 88. Lund: Lund University Press. Pp. 37-51.

Granger, S., Dagneaux, E., Meunier, F. and Paquot, M. (eds.) 2009. *International Corpus of Learner English. Version 2*. Presses Universitaires de Louvain. Belgium.

Hinrichs, L. 2006. *Codeswitching on theWeb. English and Jamaican Creole in e-mail communication*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Nicholls, D. 2003. "The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT". In Archer, D., Rayson, P., Wilson, A. and McEnery T. (eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University (UK). University Centre for Computer Corpus Research on Language. 28-31 March. Pp. 572-581.

Schmidt, T. 2012. "EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language". *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. Istanbul. 21-27 May. Pp. 236–40.

Tono, Y. 2003. "Learner corpora: Design, development and applications". In Archer, D., Rayson, P., Wilson, A. and McEnery T. (eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University (UK). University Centre for Computer Corpus Research on Language. 28-31 March. Pp. 800–809.