COPLE2 – Corpus of Portuguese FL/L2

Amália Mendes, Sandra Antunes, Anabela Gonçalves
{amalia.mendes, sandra.antunes}@clul.ul.pt; a.goncalves@letras.ulisboa.pt

Center of Linguistics of the University of Lisbon
(UID/LIN/00214/2013)

1. Introduction

COPLE2 is a new corpus of Portuguese FL/L2 that encompasses written and spoken data produced by foreign learners of Portuguese at the University of Lisbon. Following the trend towards learner corpus research applied to less commonly taught languages, it is our aim to enhance the learning data of Portuguese L2.

COPLE2 corpus aims at collecting Portuguese learning data produced by students that attended Portuguese FL/L2 courses or by applicants to accreditation exams between 2010 and 2014.

2. Written corpus

The written corpus is composed by **966 free handwritten essays** produced by **424 students** in evaluation tests, in a total of **156.691 words**.

2.1. Learner metadata

(i) **Age**: between 18 and 40 years old (80% are aged between 18-30 years old).

(ii) **Mother tongue**: 14 different L1s based on the minimum requirement of 6 informants per L1 (30% have Chinese, either Mandarin or Cantonese, as L1).

(iii) **Nationality** (relevant for languages that are spoken in more than one country).

(iv) **Knowledge of other foreign languages**.

(v) **Proficiency** (the average is elementary (40%) and intermediate (31%)).

(vi) **Type of course**: annual/summer course or accreditation exam (65% of learners attended summer courses).

(vii) **Number of years of Portuguese study**.

L1	Inf.	Masc.	Fem.	Age	Texts	Words	Words/Text
Chinese	129	33%	67%	22	232	57.377	178
English	65	34%	66%	24	142	21.610	152
Spanish	52	42%	58%	29	139	21.200	153
German	39	38%	68%	27	76	12.548	165
Russian	25	8%	92%	25	70	9.697	139
French	23	26%	74%	29	43	7.808	181
Japanese	23	26%	74%	23	50	6.809	136
Italian	20	30%	70%	25	34	5.875	172
Dutch	11	18%	82%	23	15	1.993	133
Tetum	9	56%	44%	31	22	3.163	144
Arabic	8	25%	75%	30	13	2.206	170
Polish	8	25%	75%	26	22	2.810	128
Korean	6	17%	83%	24	9	1.530	170
Romanian	6	0%	100%	26	8	2.057	257
Total	**424**	**32%**	**68%**	**26**	**966**	**156.691**	**163**

2.2. Task metadata

(i) **Genre**: opinion (36%), recount (19%), personal letter (13%), formal letter (10%), informative (10%), dialogue (6%), message (6%).

(ii) **Topic** (the essays cover a wide range of topics, depending on genre).

(iii) **Type of test** (diagnostic, intermediate or final test; accreditation exam).

(iv) **Time limit for writing or untimed**.

(v) **Use of reference tools**: dictionaries, grammars, notes, etc.

3. Spoken corpus

The compilation of the spoken corpus is still in progress and will include **28 recordings** collected in accreditation exams performed by **54 applicants** with the same profile (at the moment, 12 recordings were transcribed).

- Conversations between 2 or 3 learners of different proficiency levels moderated by the examiner, on topics such as: presentation of the students; simulation of communicative situations; discussion of particular subjects, presenting arguments and opinions.

L1	Inf.	Masc.	Fem.	Age	Words	Proficiency
Romanian	7	4	3	31	6.554	Beginning
Moldavian	5	3	2	32	2.688	Beginning
Russian	3	3	0	34	2.173	Beginning
Spanish	3	1	2	27	3.910	Beg./Adv./Prof.
Ukrainian	2	2	0	40	1.646	Beginning
Chinese	2	0	2	31	2.364	Beg./Prof.
English	1	1	0	--	554	Beginning
Greek	1	0	1	24	1.107	Advanced
Total	**24**	**14**	**10**	**31**	**20.996**	**Beginning (83%)**

4. Transcription

Regarding the **written corpus**, all the essays were digitalized and transcribed:

- The transcriptions are encoded in XML, following the TEI guidelines, and anonymized.
- Each file is composed by a header (with the metadata), and the transcription, which includes: (i) all the changes made by the student (deletions, additions, etc.); (ii) the correction and comments made by the teacher.

```
<p>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.
```

- Each transcription has an equivalent plain text with the final version of the text.

Normalmento, Eu acordo às oito horias de manhã, e tomo o duche e o pequeno-almoço. (…)

Concerning the **spoken corpus**, the recordings were transcribed following the CHILDES guidelines, and they were text-to-sound aligned using the EXMARaLDA editor.

5. Linguistic annotation and interface platform

The XML files are imported to the TEITOK – The Tokenized TEI Environment platform, which enables:

- the visualization of different versions of the texts: XML, transcription (faithful to the handwritten document), final version of the student, text-to-sound alignment;
- the linguistic annotation: lemmatization; PoS information, error codification;
- corpus search (for word, lemma, PoS, metadata, etc.), using the CQP query system.

6. Conclusion

Due to the large variety of L1s, this corpus will constitute a good resource for teachers and researchers, since it will provide empirical data to conduct studies based on Contrastive Interlanguage Analysis, and consequently: (i) identify general errors in the learning of Portuguese L2; (ii) develop textbooks and other material targeting specific groups of students; (iii) implement teacher training materials; (iv) illustrate the writing-speech interaction.

Contactos
Av. Professor Gama Pinto, 2, 1649-003 Lisboa, Portugal
Tel.: +351 21 790 47 00 | Fax: +351 21 796 56 22
http://www.clul.ul.pt

CLUL | Centro de Linguística da Universidade de Lisboa

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA
U LISBOA | UNIVERSIDADE DE LISBOA
FLUL LETRAS LISBOA
FUNDAÇÃO CALOUSTE GULBENKIAN
ICLP Instituto de Cultura e Língua Portuguesa
caple CENTRO DE AVALIAÇÃO DE PORTUGUÊS LÍNGUA ESTRANGEIRA