The COPLE2 Corpus: a Learner Corpus for Portuguese

Amália Mendes, Sandra Antunes, Maarten Janssen, Anabela Gonçalves

Universidade de Lisboa – FLUL/CLUL
Alameda da Universidade, Lisboa, Portugal
{amalia.mendes; sandra.antunes}@clul.ul.pt, maarten.janssen@campus.ul.pt, A.Goncalves@letras.ulisboa.pt

Abstract

We present the COPLE2 corpus, a learner corpus of Portuguese that includes written and spoken texts produced by learners of Portuguese as a second or foreign language. The corpus includes at the moment a total of 182,474 tokens and 978 texts, classified according to the CEFR scales. The original handwritten productions are transcribed in TEI compliant XML format and keep record of all the original information, such as reformulations, insertions and corrections made by the teacher, while the recordings are transcribed and aligned with EXMARaLDA. The TEITOK environment enables different views of the same document (XML, student version, corrected version), a CQP-based search interface, the POS, lemmatization and normalization of the tokens, and will soon be used for error annotation in stand-off format. The corpus has already been a source of data for phonological, lexical and syntactic interlanguage studies and will be used for a data-informed selection of language features for each proficiency level.

Keywords: learner corpus, corpus compilation, language learning/teaching

1. Introduction

The COPLE2 corpus[1] is a written and spoken learner corpus (i.e., a corpus of productions by foreign or second language learners (Leech, 1998)) for Portuguese that aims at providing empirical data for the teaching and learning of Portuguese as a second language (L2) or foreign language (FL). Over the past few years we are seeing a substantial growth in the area of learner corpus research applied to other languages besides English. The COPLE2 corpus will contribute to broaden this emerging domain by providing data for the global Romance language that is Portuguese. The corpus includes 978 learner productions and a total of 182,474 tokens. By including learners of Portuguese with different mother tongues (L1), it furthermore provides a resource for the development of interlanguage studies and teaching materials that target specific L1s. The corpus is available online using the TEITOK environment that offers a rich set of functionalities for the corpus administrators as well as users/visitors[2].

English learner corpora, such as the International Corpus of Learner English (Granger et al., 2009), the Longman Learner's Corpus, or the Cambridge Learner Corpus (Nicholls, 2003), have been developed in the recent years for different L1. The importance of such empirical data has been increasingly recognised for studies in Second Language Acquisition (SLA) and language teaching/learning, although they are far from achieving their full potential impact, in part due to the lack of resources for languages besides English. Initiatives such as the compilation of corpora for French (Delais-Roussarie & Yoo, 2010) and Spanish (Lozano, 2009) are a few examples of resources that address this shortcoming. Some few recent corpora have been devised to illustrate the Common European Framework of Reference for Languages – CEFR (Council of Europe, 2001) with learner texts produced in CEFR-related tests. It is the case, for instance, of the Cambridge Learner Corpus with the adherent English Profile Project (Hawkins and Filipović, 2012), and of the MERLIN corpus, that addresses Czech, German and Italian (Boyd et al., 2014).

In the case of the Portuguese language, there are some initiatives in the compilation of learner corpora. The corpus Recolha de dados de Aprendizagem do Português Língua Estrangeira[3], that follows the precursor work developed in Leiria (2001), was compiled at the Faculty of Arts of the University of Lisbon (FLUL) and includes 470 texts and 70.500 tokens. The same methodology was applied to the Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2)[4], at the University of Coimbra (CELGA), with 516 texts and 119.381 tokens. The recently compiled Corpus de Aquisição de L2 (CAL2)[5], at the New University of Lisbon (CLUNL), contains 281.301 words, and includes texts produced by adults and children, as well as a spoken subset. The COPLE2 corpus not only follows previous efforts in this domain by making use of the rich set of learner texts available at FLUL, but furthermore provides rich TEI annotation of the actual writing and error corrections. It also provides POS tags, as well as powerful multilayer query options. The categorization of the corpus texts in CEFR scales is furthermore a tool to produce an empirically motivated selection of language features for each proficiency level, together with cross-language comparisons of the L1s.

2. The Data

The COPLE2 corpus includes written and spoken materials. The written subpart currently contains 966 free essays, produced by 424 students between 2010 and 2012, in two different types of situation: as regular tests in the classes of Portuguese as Foreign Language at the *Instituto*

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

[2] http://alfclul.clul.ul.pt/teitok/learnercorpus

[3] http://www.clul.ul.pt/pt/recursos/314-corpora-of-ple

[4] http://www.uc.pt/fluc/rcpl2/

[5] http://cal2.clunl.edu.pt/

de Cultura e Língua Portuguesa (ICLP), or as accreditation exams taken at the *Centro de Avaliação de Português Língua Estrangeira* (CAPLE), both at FLUL. Students are aged between 18 and 40 years (80% aged 18-30) and represent 14 different mother tongues (we provide in brackets the number of texts per first language): Chinese (323), English (142), Spanish (139), German (76), Russian (70), French (43), Japanese (50), Italian (34), Dutch (15), Tetum (22), Arabic (13), Polish (22), Korean (9) and Romanian (8). We selected L1 that had a minimum of 6 texts in our initial data set. Texts are classified into 5 levels of proficiency that match the levels of the CEFR: Beginner (A1), Elementary (A2), Intermediate (B1), Upper Intermediate (B2), and Advanced (C1). The students are asked to perform different tasks that fall into the following genres: dialogue, formal and personal letters, informative, message/e-mail, argumentative, recount, book review and retell a story (argumentative genre accounts for 36% of the texts).

The compilation of the spoken subpart started recently and currently includes 12 transcriptions of accreditation exams of Portuguese as a foreign language (FL) at CAPLE to obtain different levels of proficiency, from A1 to C1. The exams involve a conversation between two candidates and are moderated by the evaluator. Topics are selected according to the level of proficiency and include introducing themselves, simulating different situations of daily or professional life, and presenting their opinion along with arguments to support it. The 12 transcriptions involve 24 informants, aged between 20 and 49, mostly from A1 level, representing the following L1 (number of informants per L1 is provided between brackets): Romanian (7), Moldavian (5), Russian (3), Spanish (3), Ukranian (2), Chinese (2), English (1), Greek (1). All informants have signed an informed consent that covers both sound and transcription.

We provide in Table 1 information on the composition of the COPLE2 corpus by proficiency level and modality, with the total number of texts and tokens per level.

Level	Written		Spoken		Total	
	Texts	Tokens	Texts	Tokens	Texts	Tokens
A1	72	6,438	10	18,803	82	25,241
A2	382	49,761	0	0	382	49,761
B1	305	53,042	0	0	305	53,042
B2	181	39,665	1	3,010	182	42,675
C1	26	7,785	1	3,970	27	11,755
Total	966	156,691	12	25,783	978	182,474

Table 1: COPLE2 constitution

3. The Metadata

The detailed metadata are encoded in each file in a TEI-compliant header in XML format (Burnard and Bauman, 2013). The profile of the candidate is described in 20 fields, while the task and the text are described in 14 fields. The metadata of the informants were recovered from their registration form at the course of Portuguese as foreign language. However, the form was frequently incomplete since it was filled in paper format and no strong requirement was applied. Consequently, we established a set of 7 fields that were required for the productions of the informant to be included in COPLE2: name, age, nationality, gender, mother tongue, knowledge of other second or foreign languages, period of time studying Portuguese. In practice, the information on the knowledge of other languages was frequently missing and we ended up by eliminating it from the list of required fields. This shortcoming has been recently addressed through a new online registration system that requires the student to fully fill in the set of data on his/her learner profile (for instance, information such as the level of proficiency, stays in Portuguese speaking countries (where, when, how long), education, mother tongue of mother/father, language spoken at home).

In the written subpart, the metadata for the text profile include fields on: genre, topic, task description (diagnostic test, mid-term or final test, homework, CAPLE accreditation exam), timebound or not, with access to reference books or not, number of tokens, date. The metadata of the spoken subpart encode information on the recording situation: total time of recording; total time of the segment that is transcribed; acoustic quality; hidden or visible recording; involvement of the evaluator; interactive (dialogue), non interactive (monologue) or semi-interactive (monologue with few interactions); spontaneous or planned; elicitation or non elicitation; social context (family, private, public, controlled environment); channel (face to face, experimental, media, phone conversations, etc.).

The name of the transcription files enables the identification of informant, proficiency level and exam type by using a regular pattern:

- 5 characters match the code of the informant: 2 letters refer to the first language (ISO 639-16) and 3 digits identify each informant;
- course type: annual course (CA) or summer course (CV);
- proficiency level: beginner (I), elementary (E), intermediate (M), upper intermediate (A) and advanced (S);
- exam type: diagnostic (TD), mid-term (TI) or final (TF).

Example: fr010CVITD: French speaking informant number 10, who took a summer course (CV), with a "beginner" level of proficiency, and produced the text while taking a diagnostic test (TD).

The national variant of the students (such as British English or Canadian French) is not stored explicitly, but can be inferred from crossing the *nationality* with the field *mother tongue*. In specific cases, we also faced the problem of identifying the first language of the informants correctly: for instance, many students indicated Chinese as the mother tongue, but this did not always seem to indicate Mandarin, but often rather Cantonese.

[6] http://www.loc.gov/standards/iso639-2/php/code_list.php

4. Data Preparation

The hand-written essays were first scanned and saved in pdf format, and then manually transcribed. The attempt to automatically obtain digital files through OCR was not successful due both to the handwritten nature of the text and to the number of deletions, insertions and corrections visible on the paper file. Overall, the scanned essays have a good quality which we can rely on to perform the manual transcriptions. The major problem that has arisen during the transcription process had to do with the student's handwriting. Some letters can be easily mixed up (such as 'a' and 'o', or 'm' and 'n'), and any transcriber's error is likely to influence the identification and analysis of the student's errors.

Our transcription is very close to the original document: all the changes made by the student during the writing process, such as deletions, additions, transposition of segments were encoded in TEI compliant XML. The inclusion of these elements is useful to the analysis of the student's difficulties, as well as the cognitive processes during L2 learning stages: how was the discourse firstly structured; to what extent are words confused with homophone words; etc. All the corrections and comments made by the teacher were also transcribed. Line breaks and word breaks were not marked, but paragraphs were kept. All personal information, such as names, addresses, phone numbers, were anonymised, as illustrated in examples (1) and (2), where a name of a company is replaced by "XX". For each text there is also a clean version in txt format that corresponds to the final version intended by the student.

The transcriptions of the spoken subpart follow the CHILDES (MacWhinney, 2000) and C-ORAL-ROM (Cresti and Moneglia, 2005) conventions, which are based on prosodic segmentation and mark all disfluencies. For the text-audio alignment, we used EXMARaLDA (Schmidt, 2012). Transcribing learner speech involved the challenge of deciding how to express cases of incorrect word pronunciation. A faithful transcription would lead to a multiplication of different writing options and to many inconsistencies depending on the transcriber. We consequently decided to normalize cases of pronunciation according to the Portuguese writing norm and may later add a specific layer of phonetic transcription for specific segments. A different option was taken regarding non normative productions involving morphology, the lexicon and syntax, so that a wrong segment of a word is reproduced as such, and so are syntactic constructions. But in some cases it is far from clearcut whether it is a matter of lexical error or of incorrect pronunciation.

An example of a XML transcription of a hand-written essay produced by a Chinese speaker, and the final version (removing information of deletions, additions, etc.) are given below, in (1) and (2) respectively. Deletions are marked as <del> and insertions as <add> and both may receive an attribute *hand* to identify the author of the change (student or teacher). Any marks made over a word or segment of the text are encoded as <hi> and can be further described as to authorship (*hand*) and to the type of mark on the attribute (*rend*) : either underlined, circled or crossed. These cases are all exemplified in the XML version in (1).

(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.

Translation: Usually, I wake up at eight o'clock in the morning, and I take a shower and breakfast. I leave home and catch the metro to the university, I arrive to the office of XX at nine in the morning. I write the engineer books, or I have class.

5. Annotation, visualization and queries

After the transcription was complete, the XML files were imported into the Tokenized TEI Environment (TEITOK) for visualization, annotation and search functions (Janssen, 2015). TEITOK interprets the XML encoding to enable the visualization of different versions of the written text: the XML version; the transcription version (visualization close to the full information of the original document, cf. Fig. 1); student form, with the final version intended by the student; corrected form, marking the teacher's corrections; the image of the handwritten essay (on request). The spoken transcriptions are visualized as speech turns with a link to the audio sequence.

Since the corpus contains both written and spoken files, all files in the corpus need to have the same format. For that, all EXMARaLDA files were converted to TEI format, where segments were converted to utterances, and tiers were merged, sorting utterances by time index, and associating each utterance with the speaker (tier) it originated from. This leads to a visual display of the spoken texts that resembles a transcribed interview, with each utterance time-aligned with the sound file. CSS makes sure that the TEI representation of the transcription looks similar to the original CHILDES style transcription.

The corpus is automatically tokenized and, for each token, a normalized version may be provided as an attribute for each token. The automatic POS annotation and lemmatisation was performed in the TEITOK environment, using the Neotag tagger (Janssen, 2012), trained over the CRPC – Reference Corpus of Contemporary Portuguese (Mendes et al., 2014).

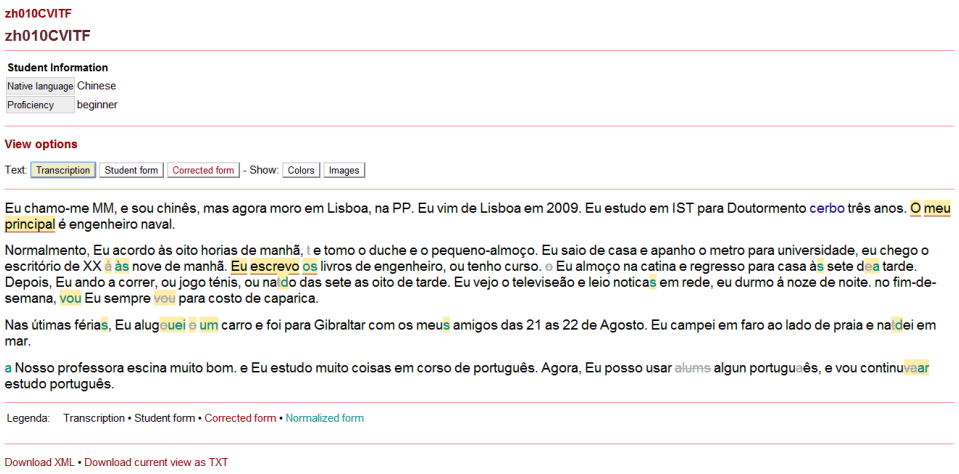

Figure 1: The TEITOK environment

The TEITOK environment provides corpus search facilities using CQP (Christ et al., 1999). In the creation of the CQP corpus, various types of information are exported: text-level data such as information about the student, token-level information including POS, lemma, original orthography and normalized orthography, and segment-based information such as error annotations.

In the TEITOK corpus, tokens have two level of annotation: orthographic words and grammatical words. For most words, those levels coincide, but for contraction and MWE, they do not. In the case where a student writes *du* and means to write *do*, the corpus contains a single orthographic token, with a written form and a normalized form, which contains two grammatical words: *de* (preposition) and *o* (article). In the CQP corpus, the grammatical words are exported. But in the KWIC results, the original XML is displayed, meaning that the original orthographic word will be displayed, along with the information about the grammatical words inside.

Searches can combine different types of information, making it possible to perform complex search queries. For instance, a frequent but difficult ending in Portuguese is *–ão*, and by comparing the written form with the normalized form, it is possible to search for all words that should have been written with *–ão* but were written with say *–am* (pronounced identically) or *–ao* (visually similar), and then get a distribution over the native tongue of the student to see whether these errors are specific for a certain group of native speakers.

The next step will be to label the data following a typological scheme for error annotation (Tono, 2003; Nicholls, 2003; Dagneaux et al., 2005). Error annotation will be done at two level: word-internal errors such as orthographic errors; morphological errors and lexical errors will be modelled over the token, by adding the error code in the same fashion as the POS and lemma, as shown in Figure 2. Segment based errors, such as word order errors, agreement errors, errors to idiomatic expressions, etc., will be modelled in a stand-off fashion using a module of TEITOK specifically built for this purpose, where error codes are related to sequences of tokens, in a way similar to for instance the UAM Corpus Tool

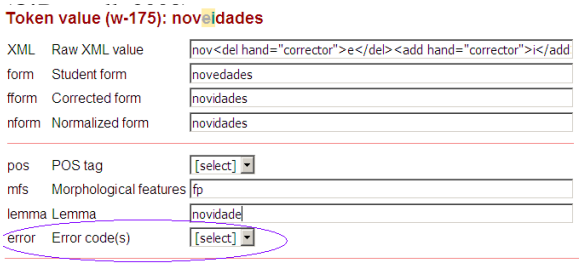

Figure 2: Error annotation scheme at the token level (in preparation)

6. Two Case Studies

The COPLE2 corpus is meant as a source of data for SLA research. As an example of such research, this section will briefly describe two studies based on the COPLE2 corpus, both performed by members of the project team: the first on copular predicative constructions and the second on multiword expressions.

6.1 Copular Predicative Constructions

Alexandre and Gonçalves (2015) focus on the acquisition of copular predicative constructions by Chinese learners of European Portuguese (EP) as L2/FL in order to answer the following question: do typological differences in the syntactic encoding of the copula in L1 affect L2 acquisition? EP and Mandarin Chinese (MC), two typologically distinct languages, vary in the way they syntactically encode the predication: EP always requires an overt copular verb (see 1), whereas MC may omit it under certain conditions, specifically when the predicate is adjectival (see (2) vs. (3)):

(1) a. A Maria **é/está** feliz.
the Maria be happy
'Mary is happy'.
b. *A Maria **Ø** feliz.

(2) a. Zhāng sān **Ø** gāo-xìng le.
Name Name high-excite CRS
'Zhang San is (now) happy'. (Sun 2006: 151)

(3) a. Zhang san **shì** zhongguó rén
Name Name be Chinese person
'Zhang San is Chinese'. (Sun 2006: 151)

b. Tā **zài** xué -xiào.
3rd be-at learn -school
'He is at school'. (Sun 2006: 157)

The examples show that the same kind of predication is involved in the two languages; however, the way they syntactically encode it varies, which may constitute a problem for MC speakers learning EP. Assuming, in line with White (2013), that in the first stages of L2 acquisition learners are constrained by the specifications in their L1, we would predict that Chinese learners of EP exhibit lower rates of errors in nominal and prepositional predicative constructions and higher rates of errors in adjectival predicative constructions. In order to test these predictions, Alexandre and Gonçalves (2015) restricted the analysis to the Chinese subcorpus of COPLE 2 and used a section of 100 written texts (out of 323), from elementary (50) and intermediate (50) levels.

The data of those learners only partially confirm the predictions above. In both levels, the omission of the copular verb is infrequent (3% in elementary; 5.2% in intermediate). At the elementary level, the omission of the copular verb was not found with prepositional predicates (see 4), but it was found with adjectival predicates (see 5), as expected:

(4) Agora eu **estou** em Lisboa.
now I am in Lisbon
'Now I am in Lisbon'.

(5) Eles são casal que **Ø** muitos simpáticos.
they are couple that very nice
'They are a very nice couple'.

However, contrary to what was expected, omission of the copula occurs in contexts where MC precludes it (i.e. with nominal predicates; see (6)), and there were in fact more omissions with nominal predicates than with adjectival ones:

(6) Acho que **Ø** uma apresentação boa.
think.1sg that a presentation good
'I think it was a good presentation'.

As for the intermediate level, the data are very similar. Besides the low frequency of errors, the copular verb was not omitted with prepositional predicates (see (7)), as expected, and the omission of the copula was observable in contexts where MC precludes it (i.e. with nominal predicates; see (8), a cleft sentence), contrary to what was expected:

(7) Estou na praia.
be.1sg in.the beach
'I am in the beach'.

(8) O quatro **Ø** que eu gostei mais no viagem.
the bedroom what I like the.most in.the travel
'The bedroom was what I like the most in the travel'.

However, more omissions with adjectival predicates than with nominal predicates were observable, as opposed to what happened at the elementary level:

(9) A praia **Ø** fatástica, ondas **Ø** boas e pessoas **Ø** alegres e simpáticas.
the beach fine waves good and people happy and nice
'The beach is fine, the waves are good and people are happy and nice'.

Considering these data, Alexandre and Gonçalves (2015) suggest that: (i) when the formal features of the L1 and L2 are similar, speakers attain more target-grammar productions, even in the first stages of L2 acquisition; (ii) the development of linguistic awareness and explicit knowledge highlights the differences between L1 and L2, which may result in a high degree of variability, maybe due to over and undergeneralization (Prévost and White, 2000; White, 2003; Cabrera and Zubizarreta, 2005), which explains the unexpected behavior of learners in the contexts that are also precluded in their L1; (iii) regressions in the higher level is in line with studies on development of L1 explicit knowledge (Afonso, et al., 2014).

6.2 Multiword Expressions

The study of multiword expressions (MWE) in learner corpora is especially important because L2 learners frequently struggle to choose the right combination of words and eventually produce errors related to the lexical-grammatical, semantic or stylistic aspects of MWE (Nesselhauf, 2004; Paquot, 2013).

Antunes and Mendes (2015) analyze which types of Portuguese MWE are particularly difficult to the students: idiomatic expressions are idiosyncratic and must be learned as a chunk, whereas compositional expressions may pose degrees of lexical and syntactic restrictions that are not easily acquired. It is also important to observe the strategies that are employed by the students to deal with unknown MWE. Since L1 and other L2s play a significant role in the students' productions, the results are also analyzed from the point of view of language transfer.

The authors consider a MWE to be a grammatically complete sequence of words that suffer some process of lexicalization. As a result, the sequence shows some lexical, syntactic and semantic cohesion (even though they don't necessarily have an idiomatic meaning). For the analysis, the authors considered the following typology of MWE, informed by proposals of different authors (Benson et al., 1986; Hausmann, 1989; Sinclair, 1991; Cowie, 1998; Mel'čuk, 1998): (i) **formulae**, like greetings, compliments, etc. (*com os melhores cumprimentos* 'best regards'); (ii) **collocations**, compositional expressions that reveal a tendency to co-occur in certain contexts (razão especial 'special reason'), or that present restricted collocability (*contrair uma doença* 'to contract a disease'); (iii) **compound nouns**, compositional or idiomatic expressions with strict constraints on lexical and syntactic variation (*guitarra elétrica* 'electric guitar'); (iv) **light verb constructions**, V + N expressions where the verb meaning appears to be partially bleached (*tirar uma fotografia* 'take a picture'); (v) **grammatical combinations**, expressions that function as text organizers or discourse markers (*por outro lado* 'on the other hand'); (vi) **idioms**, expressions where the meaning does not correspond to the meaning of

their elements (*não há bela sem senão* 'every rose has its thorn').

This study was based on the written subpart of COPLE2, and the analysis was restricted to learners of Portuguese with Spanish, English and Chinese as L1, three languages of different language families: Romance, Germanic (both indo-European) and Sinitic, respectively. The analyzed units were checked against the Portuguese reference corpus CRPC[7]. Table 2 presents information on the 3 subcorpora. The size of the subcorpora varies since it is dependent on the number of informants for each L1:

L1	Inf.	Age	Texts	Total Words	Words/Text
Chinese	129	21.9	323	57.385	178
English	65	24.5	142	21.610	152
Spanish	52	28.3	139	21.200	153
TOTAL	**246**	**24.9**	**604**	**100.195**	**161**

Table 2: COPLE2 subcorpus

The data were analyzed firstly by L1, secondly by the type of MWE and thirdly by the type of error. Table 3 presents the number of errors per type of MWE in each subcorpus. Collocations and grammatical combinations account for most of the MWE errors encountered in the 3 subcorpora (e.g. #*escritura formal* 'formal deed' vs. *escrita formal* 'formal writing'; English L1). Apart from these two categories, the Chinese L1 subcorpus shows a high frequency of errors in what concerns the categories compound nouns and light verbs constructions (e.g. #*dar muita confusão* 'to give a lot of confusion' vs. *fazer muita confusão* 'to make a lot of confusion'; Chinese L1).

MWE type	L1 Chinese	L1 English	L1 Spanish
Collocations	**97 (45%)**	**34 (53%)**	**12 (36%)**
Gram. Comb.	**33 (15%)**	**17 (27%)**	**11 (33%)**
Compounds	**33 (15%)**	2 (3%)	3 (9%)
Formulae	14 (6%)	0 (0%)	3 (9%)
Idioms	6 (3%)	5 (8%)	1 (3%)
Light verbs + pred.	**35 (16%)**	5 (8%)	3 (9%)
Total of errors	218	63	33
Size of subcorpus	57,385	21,610	21,200
Error per 1000 tok.	3.81	2.96	1.55

Table 3: CIA – errors per MWE type

The subcorpora of Spanish and English as L1 are comparable in terms of number of texts and total number of words (see Table 2). However, the data in Table 3 (normalized to corpus size per 1000 tokens) show that English students produce twice as many MWE errors as the Spanish ones, which is presumably related to the high similarity between Spanish and Portuguese. Chinese students are clearly the ones producing more errors. However, when compared to the English students, we see that the increase in errors is not proportional to the increase in number of words in the corpus.

Table 4 shows the type of errors that are produced by each subset of learners. In this table, there are indications of the number of errors that appear influenced by either the L1 or another L2 of the student, so amongst the 75 lexical mismatch errors by Chinese students, there were 13 that seemed influenced by their L1, and 8 that seemed influenced by a L2. The type of MWE error that seems most influenced by transfer is lexical mismatch. Example (10) shows a production from a Chinese student with a direct translation of the Chinese expression to Portuguese:

(10) *quero **abrir os meus olhos** e descobrir novos valores*
('I want **to open my eyes** and find new values')

instead of:

*quero **alargar os meus horizontes** e descobrir novos valores* ('I want **to broaden the horizon** and find new values').

Example (11) shows a production from a Chinese student with transfer from English, his L2 (the English word 'balance' has a similar morphology than the Portuguese word *balança*, which means 'scale'):

(11) *a destruição biológica pode afectar **a balança da natureza*** ('biological destruction may affect **the scale of the nature**')

instead of:

*a destruição biológica pode afectar **o equilíbrio da natureza*** ('biological destruction may affect the **balance of the nature**').

This overview of mismatches in the production of MWE in the 3 subcorpora provides input for the teaching and learning of Portuguese L2. Our data show that collocations are especially difficult because they pose degrees of restrictions that are not easily acquired, and there is little information available in Portuguese dictionaries (comparing with resources for English). There are few cases of idiomatic expressions in our corpus probably because the learners have elementary proficiency and are not yet familiarized with them. To target this subtype, other methods, such as translations or elicitation tests, would be required.

The data provides ground for an approach that is "tailored" for each language in what concerns teaching Portuguese as L2: for instance, speakers of Chinese L1 require special input in dealing with light verb selection; special attention to categories that do not have an equivalence in L1 may improve the correct production of MWE, as in the case of the articles and nominal/verbal agreement with Chinese speakers.

We plan in the future to contrast our findings with cases of correct production of MWE in the 3 corpora and to categorize our data according to the proficiency level.

7. Final Remarks

The COPLE2 corpus is a new learner corpus for Portuguese, addressing both written and spoken modality, with detailed metadata and a rich XML encoding of the hand-written originals and of the audio transcriptions. It is automatically annotated with POS and lemma information and provides a full set of functionalities in visualization and search in the TEITOK environment. The upcoming annotation at error level will further improve this resource.

[7] http://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc

Type of error	L1 Chinese	L1 English	L1 Spanish
Lexical mismatch	75 (34%) **L1:13 L2: 8**	20 (31%) **L1: 8 L2: 2**	13 (39%) **L1: 4**
Error on preposition	20 (9%)	18 (28%) **L1: 4**	5 (15%)
Morphology	19 (9%) **L1:16**	3 (5%)	4 (12%) **L1: 4**
Imposs. of substitution for synonyms	38 (17%) **L1:10**	9 (14%) **L1: 4**	3 (9%) **L1: 1**
Light verbs	20 (9%) **L1: 2 L2:1**	5 (8%)	3 (9%)
Phonology	7 (3%)	4 (6%)	2 (6%)
Requirement of article	1 (0.5%)	1 (2%)	1 (3%) **L1: 1**
Impossibility of insertion of article	10 (5%) **L1: 10**	1 (2%) **L1: 1**	1
Choice of article	0 (0%)	1 (2%)	0
Word order	8 (4%) **L1: 4**	0	1 (3%)
Transposition of semantic relation	0 (0%)	1 (2%)	0
Manipulation	1 (0,5%)	0	0
Periphrasis	6 (3%) **L1: 2**	0	0
Impossibility of singular/plural form	10 (5%) **L1: 10**	0	0
Error on complement	2 (0.5%)	0	0
Transposition of foreign word	1 (0.5%) **L1:1 L2:1**	0	0
TOTAL	**218 (100%)**	**63 (100%)**	**33 (100%)**

Table 4: CIA – errors per type of error

The corpus provides learner data for the study of teaching and learning of L2 Portuguese, and the rich set of L1 included enables a large array of cross-language studies (Granger, 1996). This is especially useful to rightly identify cases of L1 transfer (Jarvis, 2000). Some specific topics have already been informed by empirical data provided by COPLE2, such as the interlanguage contrastive analysis of copular verbs (Alexandre and Gonçalves, 2015), formulaic language (Antunes and Mendes, 2015), L1 and L2 lexical transfer (Pinto, 2015), relative clauses (Alexandre and Pinto, 2014) and the orthographic use of vowels (Castelo et al., 2015).

The analysis of the input from the teacher (corrections and comments) could be useful to the development of teaching material, which could easily focus not only on common errors made by students in general, but also on particular errors made by the students from a specific L1.

The corpus data provide a direct contribute to didactic applications and resources for Portuguese learning and may illustrate the relevant features for the CEFR levels for L2 Portuguese.

8. Acknowledgements

This work was partially supported by Fundação Calouste Gulbenkian (Proc. nr. 134655), Fundação para a Ciência e a Tecnologia (project PEst-OE/LIN/UI0214/2013) and Associação para o Desenvolvimento da Faculdade de Letras da Universidade de Lisboa. The authors wish to thank the anonymous reviewers for their comments and suggestions.

9. Bibliographical References

Afonso, C., Gonçalves, A. and Freitas, M. J. (2014). A princesa ficou *adormir ou a dormir? Dados sobre a consciência da unidade palavra em Português europeu. *Linguística: Revista de Estudos Linguísticos da Universidade do Porto 9*, pp. 35--58.

Alexandre, N. and Pinto, J. (2014). Aspects of relative clauses in Portuguese as a foreign language by Chinese learners. In *20th Conference of the European Association for Chinese Studies*. July 22-26, Braga, Coimbra.

Alexandre, N. and Gonçalves, A. (2015). Copular constructions in Portuguese as a second language (PL2) by Chinese learners: Do typological differences matter? In *Workshop on Copulas across Languages*. June 18-19, University of Greenwich, London, England.

Antunes, S. and Mendes, A. (2015). Portuguese Multiword Expressions: data from a learner corpus. Poster presented at *LCR2015: Third Learner Corpus Research Conference*. September 11-13, Radboud University, Nijmegen, The Netherlands.

Benson, M., Benson, E. and Ilson, R. (1986). *The BBI Combinatory Dictionary of English a guide to word combination*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B. and Vettori, C. (2014). The MERLIN corpus: Learner Language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp.1281--1288.

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

Cabrera, M. and Zubizarreta, M. L. (2005). Overgeneralization of Causatives and Transfer in L2 Spanish and L2 English. In D. Eddington (ed.). *Selected Proceedings of the 6th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*. Somerville, MA: Cascadilla Proceedings Project, pp. 15--30.

Castelo, A., Santos, R. and Freitas, M. J. (2015). O uso de vogais ortográficas por aprendentes de Português como língua estrangeira: unidade na diversidade. In *Língua Portuguesa: Unidade na diversidade – Cultura, Literatura, História, Linguística, Tradução e Ensino*. November 5-6, Lublin, Poland.

Christ, Oliver, Schulze, B., Hofmann, A. and Koenig, E. (1999). *The IMS Corpus Workbench: Corpus Query*

Processor (CQP): User's Manual. Institute for Natural Language Processing. University of Stuttgart. (CQP V2.2).

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K: Press Syndicate of the University of Cambridge.

Cowie, A. P. (ed.) (1998). *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press.

Cresti, E. and Moneglia, M. (Eds.) (2005). *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

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.

Delais-Roussarie E. and Yoo, H. (2010). The COREIL corpus: a learner corpus designed for studying phrasal phonology and intonation. In K. Dziubalska-Kołaczyk, M. Wrembel and M. Kul (Eds), *Proceedings of New Sound* 2010. Poznan, Pologne, pp. 100--105.

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.

Hawkins, J. A. and Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge: Cambridge University Press.

Hausmann, F. J. (1989). Le dictionnaire de collocations. *HSK 5.1*, pp. 1010--1019.

Janssen, M. (2012). NeoTag: a POS Tagger for Grammatical Neologism Detection. In *Proceedings of LREC 2012*, Istanbul, Turkey.

Janssen, M. (2015) Multi-level manuscript transcription: TEITOK. Presented at *Congresso de Humanidades Digitais em Portugal*. October 8-9, Lisbon.

Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon. *Language Learning 50(2)*, pp. 245--309.

Leech, G. (1998). Preface. In Granger, S. (ed.). *Learner English on Computer*. London: Addison Wesley Longman Limited, pp. xiv--xx.

Leiria, I. (2001). *Léxico – aquisição e ensino do Português Europeu língua não materna*. PhD Dissertation. Faculdade de Letras da Universidade de Lisboa.

Lozano, C. (2009). CEDEL2: Corpus Escrito del Español L2. In C. M. Bretones Callejas et al. (Eds). *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería, pp. 197--212.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mel'čuk, I. (1998). Collocations and Lexical Functions. In Cowie, A. P. (ed.). *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press, pp. 23--53.

Mendes, A., Généreux, M., Hendricks, I. (2014). *Manual for the CRPC on the CQPweb interface*. Manual 1.3. http://alfclul.clul.ul.pt/CQPweb/doc/CRPCmanual.v1_2_en.pdf.

Nesselhauf, N. (2004). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins Publishing Company.

Nicholls, D. (2003). The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson and T. McEnery (Eds.). *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University, pp. 572--581.

O'Donnell, M. (2008). The UAM CorpusTool: Software for corpus annotation and exploration. In C. M Bretones Callejas et al. (Eds.). *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería, pp. 1433-1447.

Paquot, M. (2013). Lexical bundles and L1 transfer effects. *Language Learning and Technology 14(2)*, pp. 30--49.

Pinto, J. (2015). O papel da L1 e da L2 na aquisição lexical de português L3. *Revista liLETRAd, 1*, pp. 299--310.

Prévost, P. and White, L. (2000). Missing surface inflection or impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research 16(2),* pp. 103--133.

Sag, I., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. Gelbukh (ed.). *Proceedings of CICLing-2002*. Vol. 2276- pp. 1--15.

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

Sinclair, J. (1991). *Corpus, Concordance and Collocation.* Oxford: Oxford University Press.

Sun, C. (2006). *Chinese: a Linguistic introduction*. Cambridge: Cambridge University Press.

Tono, Y (2003). Learner corpora: Design, development and applications. In D. Archer, P. Rayson, A. Wilson and T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University, pp. 800--809.

White, L. (2003). *Second Language Acquisition and Universal Grammar*. Cambridge: Cambridge University Press.

10. Language Resource References

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