Apresentação do *Corpus de Português Língua Estrangeira/Língua Segunda – COPLE2*

Sandra Antunes, Amália Mendes, Anabela Gonçalves, Maarten Janssen, Nélia Alexandre, António Avelar, Adelina Castelo, Inês Duarte, Maria João Freitas, José Pascoal, Jorge Pinto
Centro de Linguística da Universidade de Lisboa
Instituto de Cultura e Língua Portuguesa
Centro de Avaliação de Português Língua Estrangeira

Abstract: In this article, 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. Our aim is to enhance the learning data of Portuguese, a less commonly taught language. We believe that 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; (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. We will briefly describe the work in progress regarding the constitution and linguistic annotation of this corpus.
Keywords/Palavras-chave: *Corpus* de aprendizagem, ensino e aquisição de português LE/L2, anotação do erro.
Learner corpus, second language acquisition, foreign language teaching, error annotation.

1. Introdução

O *Corpus de Português Língua Estrangeira/Língua Segunda – COPLE2* é um projeto em desenvolvimento na Faculdade de Letras da Universidade de Lisboa (FLUL)[1] que tem como objetivo a compilação e análise de materiais escritos e orais produzidos por alunos estrangeiros que estão a aprender português como língua estrangeira (LE) ou língua segunda (L2), bem como

[1] Este projeto tem o apoio financeiro da Fundação para a Ciência e a Tecnologia (UID/LIN/00214/2013), Fundação Calouste Gulbenkian (Projeto LeCIEPLE, Proc. nr. 134655) e Associação para o Desenvolvimento da Faculdade de Letras da Universidade de Lisboa.

por candidatos a exames de certificação de proficiência de nível de língua. A recolha dos materiais é feita na FLUL, no âmbito dos cursos de Português Língua Estrangeira, do Instituto de Cultura e Língua Portuguesa (ICLP), e dos exames de acreditação realizados pelo Centro de Avaliação de Português Língua Estrangeira (CAPLE).

A produção de *corpora* de aprendizagem das línguas tem conhecido um crescente interesse, embora a maioria dos recursos produzidos vise a língua inglesa[2]. É o caso dos *corpora* que constituem uma referência nesta área: o *Longman Learner's Corpus*[3], o *Cambridge Learner Corpus* (Nichols, 2003)[4] e **o** *International Corpus of Learner English* (ICLE)[5]. Este último *corpus* resultou de um projeto que teve início em 1990 e constitui uma referência pela sua extensão (3 milhões de palavras) e metodologia (Granger *et al*., 2009)**.** O ICLE contém produções escritas de aprendentes de vários países, de nível avançado, e conta com parcerias internacionais, tendo tido já várias edições. Um dos seus subcorpora foi desenvolvido no Brasil, na Universidade Católica de São Paulo, e contém produções escritas de brasileiros aprendentes da língua inglesa (Berber Sardinha, 2001). O *USP Multilingual Learner Corpus* é uma extensão desse trabalho, e abrange outras línguas, como o alemão e o espanhol. Existem ainda, para o inglês, o *corpus* oral *Louvain International Database of Spoken English Interlanguage* (LINDSEI)[6]**.** Têm, contudo, vindo a ser constituídos vários *corpora* que visam a aprendizagem de outras línguas, além do inglês, como os *corpora* de aprendizagem do francês (Delais-Roussarie & Yoo, 2010), do espanhol (Lozano, 2009) e do árabe (Abuhakema *et al*., 2008).

No caso do português, destacam-se vários *corpora* já compilados ou em curso. O *corpus Recolha de dados de Aprendizagem do Português Língua Estrangeira*[7], desenvolvido na

[2] A preponderância de recursos no que respeita ao estudo do inglês L2 pode ser observada na lista *Corpora Around the World*, disponibilizada no *site* da Universidade Católica de Louvain (http://www.uclouvain.be/en-cecl-lcworld.html).

[3] http://longmandictionariesusa.com/longman/corpus#aa

[4] http://www.cambridge.org/gb/cambridgeenglish/about-cambridge-english/cambridge-english-corpus

[5] https://www.uclouvain.be/en-cecl-icle.html

[6] https://www.uclouvain.be/en-cecl-lindsei.html

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

Faculdade de Letras da Universidade de Lisboa na sequência do trabalho realizado em Leiria (2001), é um *corpus* composto por 470 produções escritas de aprendentes de português língua estrangeira, num total de 70.500 palavras transcritas. A natureza dos materiais compilados, os seus metadados e as normas de transcrição são apresentados na página do *corpus*, que permite, ainda, descarregar a totalidade dos materiais em formato TXT. A mesma metodologia foi seguida no *Corpus de Produções Escritas de Aprendentes de PL2* (PEAPL2) [8], na Faculdade de Letras da Universidade de Coimbra, e que é constituído por um acervo de 516 produções escritas, num total de 119.381 palavras, que pode ser descarregado em formatos TXT e DOC. O *Corpus de Aquisição de L2* (CAL2)[9], da Universidade Nova de Lisboa, contém 281.301 palavras e difere dos anteriores por incluir produções escritas de falantes adultos e de crianças – embora o número de textos do primeiro tipo seja muito superior (1380 vs. 103) – e, ainda, por incluir produções orais (192 entrevistas de adultos e 95 entrevistas de crianças).

O presente projeto propõe-se, assim, a ampliar e atualizar os dados recolhidos na FLUL, produzidos no âmbito da aprendizagem de português LE/L2, de modo que possa servir de suporte à investigação e à formação de professores, uma vez que fornecerá dados empíricos que permitirão: (i) a identificação de erros comuns produzidos durante o processo de aquisição desta língua; (ii) a criação de um perfil linguístico dos aprendentes de PLE/L2, tendo em atenção a sua língua materna (L1), e o consequente aperfeiçoamento de materiais didáticos e adequação de estratégias de ensino a diferentes públicos; (iii) o desenvolvimento de materiais no âmbito da formação de professores, tendo em atenção a análise da correção dos professores. Os dados recolhidos e tratados até ao momento estão disponíveis para visualização e pesquisa na página do projeto**[10]**.

Neste artigo, descrever-se-á, nas secções 2 e 3, respetivamente, o trabalho referente às tarefas de constituição e transcrição do *corpus*. A secção 4 debruçar-se-á sobre a interface de

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

[9] http://cal2.clunl.edu.pt/. O *corpus* está acessível através de pedido direto aos seus autores.

[10] http://www.clul.ul.pt/pt/investigacao/547

visualização e de pesquisa e sobre o processo de anotação linguística e a secção 5 encerra este artigo, apontando-se as utilizações possíveis dos resultados.

2. Constituição do *corpus*

O COPLE2 é composto por um acervo de materiais escritos e orais produzidos por 483 alunos de PLE/L2 (424 para o escrito e 59 para o oral) que frequentaram a FLUL em cursos anuais ou de verão de Português Língua Estrangeira, ministrados pelo ICLP, ou que realizaram exames de certificação de proficiência de nível de língua no CAPLE, entre os anos de 2010-2014.

2.1. Constituição do *corpus* escrito

Devido à heterogeneidade dos aprendentes, exerceu-se um controlo rigoroso sobre as variáveis dos informantes e dos textos, tendo todos os metadados sido tratados numa base de dados. Deste modo, no que respeita ao perfil dos informantes, teve-se em consideração os seguintes campos e valores:

(i) idade

os informantes têm idade compreendida entre os 18 e os 40 anos (80% dos informantes têm idade entre os 18 e os 30 anos, enquanto apenas 20% apresentam idade entre os 31 e os 40 anos);

(ii) língua materna

foram consideradas 14 línguas maternas diferentes, com base no requisito mínimo de 6 informantes por L1 (cf. Quadro 1);

L1	N.º Informantes	L1	N.º Informantes
Chinês	129	Italiano	20
Inglês	65	Holandês	11
Espanhol	52	Tétum	9
Alemão	39	Árabe	8
Russo	25	Polaco	8
Francês	23	Coreano	6
Japonês	23	Romeno	6

Quadro 1: Distribuição dos informantes por L1

(iii) nacionalidade (relevante no caso de línguas que são faladas em vários países, como é o caso do inglês, que, no *corpus* em questão, é falado por informantes provenientes do Reino Unido (Inglaterra, Escócia e Irlanda do Norte), Estados Unidos da América, Austrália, Canadá, Índia e Nova Zelândia);
(iv) habilitações académicas;
(v) conhecimento de outras línguas estrangeiras;
(vi) proficiência (inicial, elementar, intermédio, avançado, superior – que correspondem, respetivamente, aos níveis A1, A2, B1, B2, C1, do *Quadro Europeu Comum de Referência para as Línguas*);
(vii) tipo de curso (anual ou de verão);
(viii) permanência em países lusófonos (onde, quando, durante quanto tempo);
(ix) tempo de estudo de português.

Relativamente ao perfil dos textos, foram considerados os seguintes dados:

(i) género textual (texto argumentativo, texto informativo, texto narrativo, carta pessoal, carta formal, diálogo, *e-mail*/mensagens, crítica literária);
(ii) tópico (a vida do informante em Portugal, carta para um amigo sobre as férias na praia, carta de reclamação para uma agência de viagens, mercado de trabalho no país do informante, etc.);
(iii) tipo de tarefa (teste diagnóstico, intercalar ou final, trabalho de casa, exame CAPLE);
(iv) condições da tarefa (com/sem limite de tempo para redigir);
(v) recurso ou não a materiais didáticos auxiliares (dicionários, gramáticas, outros);
(vi) número de palavras;
(vii) data.

Os ficheiros de transcrição (ver critérios de transcrição do manuscrito original na secção 3) foram nomeados de modo que se pudesse depreender facilmente o perfil do informante e do texto. Assim, os mesmos são compostos por:

- 5 caracteres correspondentes ao código do informante (2 letras relativas à L1, de acordo com os códigos ISO 639-1[11], e 3 dígitos);
- o tipo de curso: anual (CA) ou de verão (CV);
- o nível de proficiência: inicial (I), elementar (E), Intermédio (M), avançado (A) ou superior (S);
- o tipo de teste: teste diagnóstico (TD), teste intercalar (TI) ou teste final (TF).

Deste modo, um ficheiro com o nome fr010CVITD, por exemplo, corresponderá ao informante francês (fr010) que frequentou um curso de verão (CV) de nível inicial (I) e fez um teste diagnóstico (TD).

Os Quadros 2, 3 e 4 ilustram a constituição do *corpus* à data de publicação deste artigo e fornecem dados sobre o número de textos por nível de proficiência, género textual e língua materna.

L1	Inf.	Masc.	Fem.	Média idade	Testes	Textos	Palavras	Média palavras/texto
Chinês	129	33%	67%	22	277	323	57.377	178
Inglês	65	34%	66%	24	118	142	21.610	152
Espanhol	52	42%	58%	29	102	139	21.200	153
Alemão	39	38%	68%	27	69	76	12.548	165
Russo	25	8%	92%	25	52	70	9.697	139
Francês	23	26%	74%	29	40	43	7.808	181
Japonês	23	26%	74%	23	45	50	6.809	136
Italiano	20	30%	70%	25	28	34	5.875	172
Holandês	11	18%	82%	23	14	15	1.993	133
Tétum	9	56%	44%	31	19	22	3.163	144
Árabe	8	25%	75%	30	13	13	2.206	170
Polaco	8	25%	75%	26	16	22	2.810	128
Coreano	6	17%	83%	24	9	9	1.530	170
Romeno	6	0%	100%	26	8	8	2.057	257
TOTAL	**424**	**32%**	**68%**	**26**	**810**	**966**	**156.691**	**163**

Quadro 2: Constituição do *corpus* escrito

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

Proficiência	Chin.	Ing.	Esp.	Ale.	Rus.	Fr.	Jap.	It.	Hol.	Tét.	Ár.	Pol.	Cor.	Rom.	Total
Inicial	9	18	4	14	8	6	4	1	4	0	1	3	0	0	**72**
Elementar	107	57	79	27	13	16	27	13	5	17	4	12	3	2	**382**
Intermédio	109	43	43	20	29	14	11	13	4	4	5	2	5	3	**305**
Avançado	86	23	11	11	19	7	8	6	2	0	1	4	1	2	**181**
Superior	12	1	2	4	1	0	0	1	0	1	2	1	0	1	**26**
Total	**323**	**142**	**139**	**76**	**70**	**43**	**50**	**34**	**15**	**22**	**13**	**22**	**9**	**8**	**966**

Quadro 3: Distribuição dos textos por nível de proficiência dos informantes e por língua materna

Registo	Chin.	Ing.	Esp.	Ale.	Rus.	Fr.	Jap.	It.	Hol.	Tét.	Ár.	Pol.	Cor.	Rom.	Total
Dialogo	9	13	8	8	6	3	7	2	1	3	0	1	1	0	**62**
Carta formal	36	10	20	6	5	6	3	3	0	0	5	3	3	1	**101**
Carta pess.	28	22	18	6	8	7	9	7	4	5	2	2	0	3	**121**
Informativo	30	17	5	9	8	2	6	4	3	3	0	4	2	0	**93**
Mensagem	29	5	20	0	0	1	1	3	0	0	0	2	0	0	**61**
Opinião	148	45	36	28	34	16	13	9	2	2	4	3	2	1	**343**
Narração	43	29	32	19	9	8	11	6	5	9	2	7	1	2	**183**
Crítica lit.	0	1	0	0	0	0	0	0	0	0	0	0	0	1	**2**
Total	**323**	**142**	**139**	**76**	**70**	**43**	**50**	**34**	**15**	**22**	**13**	**22**	**9**	**8**	**966**

Quadro 4: Distribuição dos textos por género textual e por língua materna

Observando os dados estatísticos sobre a constituição do *corpus* escrito, verifica-se que:

- 30% dos aprendentes de português LE/L2 têm o chinês (que engloba aqui o mandarim e o cantonês) como L1;
- os textos foram produzidos maioritariamente por informantes do sexo feminino (68%);
- o nível médio de proficiência corresponde ao elementar (40%);
- o género textual mais produtivo é o de opinião (36%).

2.2. Constituição do *corpus* oral

O trabalho de compilação do *corpus* oral está ainda em curso. Em outubro e novembro de 2014, foram recolhidas 28 gravações, com 54 informantes, no âmbito dos exames de certificação da proficiência em Português Língua Estrangeira realizados pelo CAPLE, uma unidade da FLUL. Todos os informantes autorizaram a recolha das produções escritas e a gravação dos registos de voz para análise linguística no âmbito deste projeto. As gravações correspondem a exames feitos com o objetivo de obtenção de vários tipos de diplomas, dependendo do nível de proficiência da língua: CIPLE (certificado inicial), DEPLE (diploma elementar), DIPLE (diploma intermédio), DAPLE (diploma avançado), DUPLE (diploma universitário).

Os exames incluem conversas entre dois ou três candidatos, moderadas pelo avaliador. Os tópicos abordados dependem do grau de proficiência e podem incluir a apresentação dos candidatos, a simulação de situações comunicativas do quotidiano e do domínio laboral, e a apresentação de opiniões e de argumentos sobre determinados temas.

No que respeita aos metadados, estes foram adaptados para este tipo de registo e incluem informação respeitante aos seguintes campos, caraterísticos do oral:

(i) duração total da gravação, duração do excerto transcrito e localização do excerto transcrito (m's'');

(ii) qualidade acústica da gravação;

(iii) condições da gravação: gravador visível/escondido, investigador participante/não-participante;

(iv) interatividade: interativo (diálogos e conversas), não-interativo (monólogos) semi-interativo (monólogos pontuados por intervenções esporádicas por parte do interlocutor);

(v) planeamento: espontâneo (comunicação sem tópicos determinados previamente), semiespontâneo (tópicos orientados por um dos intervenientes), planeado (discurso planeado previamente em detalhe pelos intervenientes);

(vi) participação: elicitado (o investigador pede para produzir atos isolados, como frases, palavras, fonemas, etc.), não-elicitado (o investigador não intervém no discurso), não-observado (o gravador regista continuamente conversas num determinado local);

(vii) contexto social: familiar (entre membros de uma família), privado (entre amigos e colegas), público (entre pessoas desconhecidas ou pouco familiares), ambiente controlado (testes linguísticos);

(viii) canal: cara a cara (discurso espontâneo), ambiente experimental (testes linguísticos), radiodifusão (noticiários, reportagens, *talk shows*, entrevistas, etc.), formal (conferências, aulas, homilias, discursos e debates políticos, etc.), conversas telefónicas (discurso espontâneo, interação homem--máquina).

Das 28 gravações recolhidas, foram, até ao momento, transcritas 12, num total de 24 informantes (14 homens e 10 mulheres), com idades compreendidas entre os 20 e os 49 anos e, maioritariamente, com nível de proficiência inicial. Os informantes têm 8 línguas maternas diferentes, conforme se pode observar no Quadro 5.

L1	Inf.	Masc.	Fem.	Média idade	Palavras	Proficiência
Romeno	7	4	3	31	6.554	Inicial
Moldavo	5	3	2	32	2.688	Inicial
Russo	3	3	0	34	2.173	Inicial
Espanhol	3	1	2	27	3.910	Ini./Avan./Sup.
Ucraniano	2	2	0	40	1.646	Inicial
Chinês	2	0	2	31	2.364	Ini./Sup.
Inglês	1	1	0	--	554	Inicial
Grego	1	0	1	24	1.107	Avançado
TOTAL	**24**	**14**	**10**	**31**	**20.996**	**Inicial (83%)**

Quadro 5: Constituição do *corpus* oral

3. Transcrição dos dados

3.1. Transcrição do *corpus* escrito

Após a seleção e digitalização dos manuscritos, em papel, procedeu-se à sua transcrição em formato XML de acordo com as normas de transcrição estabelecidas pela *Text Encoding Initiative* – TEI (Burnard & Bauman, 2013). Cada ficheiro contém um cabeçalho com os metadados detalhados (importados da base de dados) e a transcrição do texto, na qual se encontram codificadas todas as intervenções presentes no manuscrito original (cf. Figura 1). Deste modo, cada transcrição inclui:

(i) as modificações assinaladas pelo informante durante a produção do texto (apagamentos, adições, segmentos alternativos, segmentos transpostos, sublinhados, etc.);

(ii) as correções e os comentários introduzidos pelo professor.

Procedeu-se, igualmente, à anonimização de todos os nomes e outros dados pessoais através da inserção de códigos na transcrição (Hinrichs, 2006).

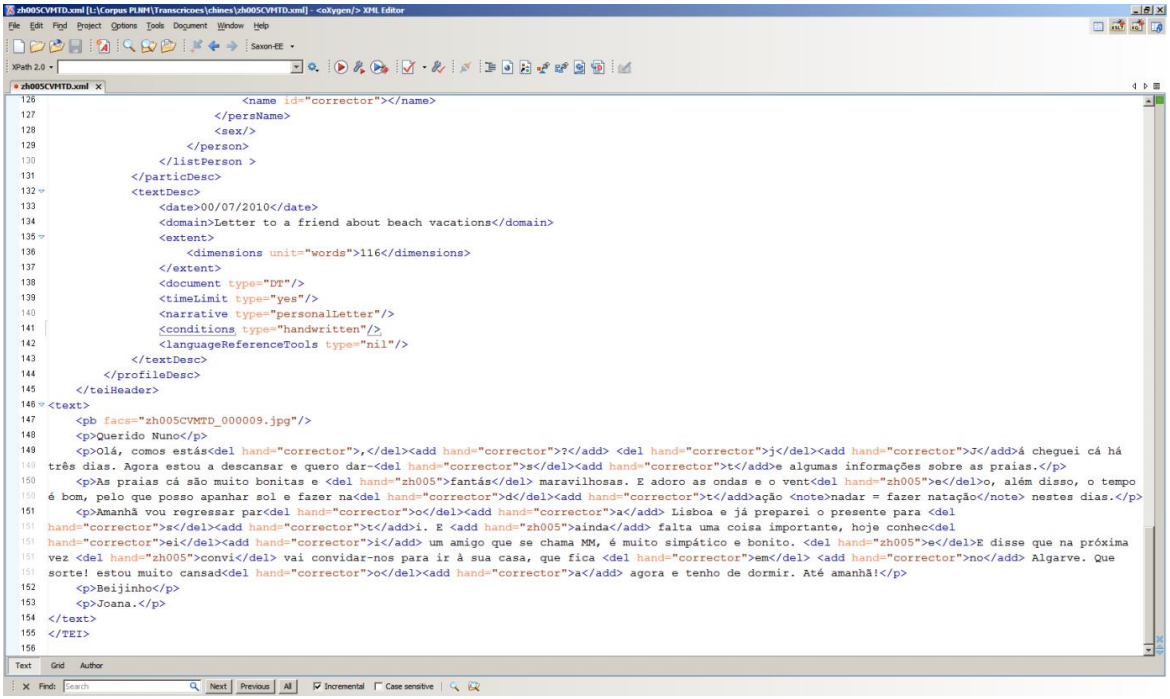

Figura 1: Transcrição em formato XML

Para cada transcrição existe, igualmente, uma versão TXT, que não inclui os metadados nem a correção do professor, representando, apenas, a versão final e limpa do texto escrito pelo informante, conforme se pode observar na Figura 2, que reproduz o mesmo texto exemplificado na Figura 1.

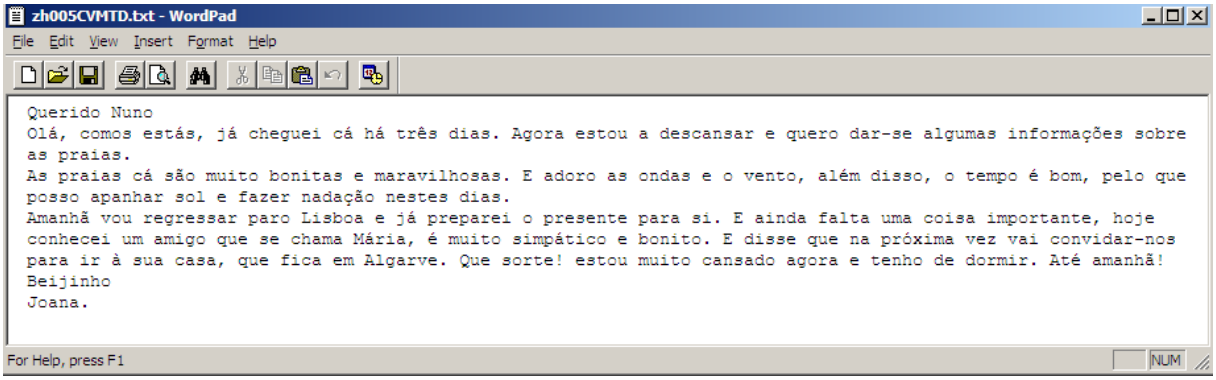

Figura 2: Transcrição em formato TXT

3.2. Transcrição do *corpus* oral

As transcrições do *corpus* oral foram feitas de acordo com as normas de transcrição adotadas pelo grupo Anagrama, do Centro de Linguística da Universidade de Lisboa, e baseiam-se nas convenções de transcrição utilizadas nos projetos CHILDES (MacWhinney, 2000) e C-ORAL-ROM (Cresti and Moneglia, 2005), que privilegiam uma transcrição baseada na prosódia. Assim, em detrimento de sinais de pontuação, caraterísticos do registo escrito, são usados símbolos que representam a entoação prosódica, como '/' (que assinala uma quebra prosódica breve no interior do discurso) ou '//' (que assinala uma quebra prosódica no fim do discurso). O símbolo '?' mantém-se nos contextos interrogativos. São igualmente transcritos todos os fenómenos de disfluência caraterísticos dos *corpora* orais, tais como as pausas preenchidas, as repetições, as reformulações, as palavras e os enunciados interrompidos ou abandonados e todos os outros eventos que desempenham um papel estrutural na organização discursiva.

Foi utilizado o *software* EXMARaLDA (Schmidt, 2012), que permite o alinhamento entre o texto e o sinal acústico (cf. Figura 3).

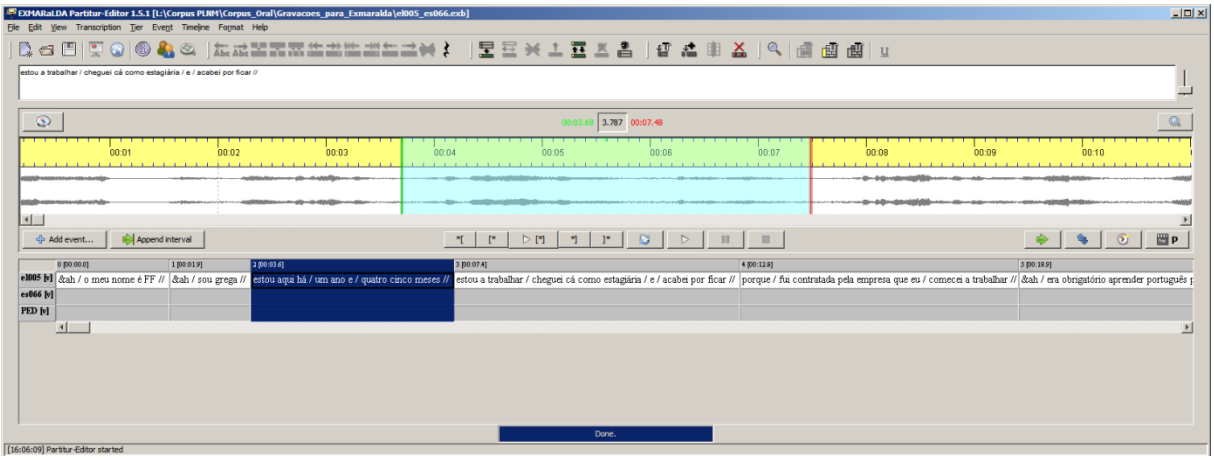

Figura 3: Transcrição e alinhamento texto-som no *software* EXMARaLDA

Tal como no *corpus* escrito, também neste caso se procedeu à anonimização de todos os nomes e dados pessoais, quer através da inserção de códigos na transcrição, quer através da inserção de silêncio no ficheiro de som. Todas as transcrições foram posteriormente convertidas para o formato TEI.

4. Interface de acesso e anotação linguística

No âmbito do projeto que se descreve neste trabalho, está a ser utilizada a plataforma TEITOK (*The Tokenized TEI Environment*)[12], com interface gráfico, que permite a importação dos ficheiros XML e o seu tratamento linguístico. Esta plataforma oferece, num único ambiente, várias funcionalidades:

(i) diferentes visualizações dos textos do *corpus* escrito: o código XML, a interpretação gráfica dos códigos XML (que produz uma versão digital equivalente ao manuscrito original, com todas as modificações do texto feitas pelos alunos, bem como as correções do professor) e a versão final do aluno, sem as suas reformulações e sem as correções do professor (cf. Figura 4). No que respeita ao *corpus* oral, é possível

[12] A plataforma TEITOK (http://alfclul.clul.ul.pt/teitok/site/index.php?action=about) está a ser desenvolvida no Centro de Linguística da Universidade de Lisboa pelo Investigador FCT Maarten Janssen.

visualizar o código XML e o alinhamento entre o texto e o sinal acústico, em formato TEI (cf. Figura 5);

(ii) tokenização dos ficheiros (que consiste na identificação de *tokens* simples (unidades gráficas isoladas por espaço) e de *tokens* que correspondem a formas contraídas (*do*, *no*, *naquelas*, etc.);

(iii) diferentes níveis de informação linguística:

- normalização ortográfica, que permite corrigir a ortografia de formas erradamente produzidas pelos alunos;
- lematização;
- anotação morfossintática (em curso);
- codificação do erro a partir de um esquema tipológico (em curso).

(iv) opções de pesquisa sobre os textos do *corpus*, nos diversos níveis de anotação, baseadas no sistema CQP. A pesquisa pode ser feita pela forma ortográfica (*word*), pelo lema (*lemma*), pela classe de palavra (*pos tag*), por expressões regulares, por combinações destas opções, pelos metadados – nacionalidade, L1, L2, proficiência, etc. – (*advanced search*) ou pelo tipo de erro (em curso).

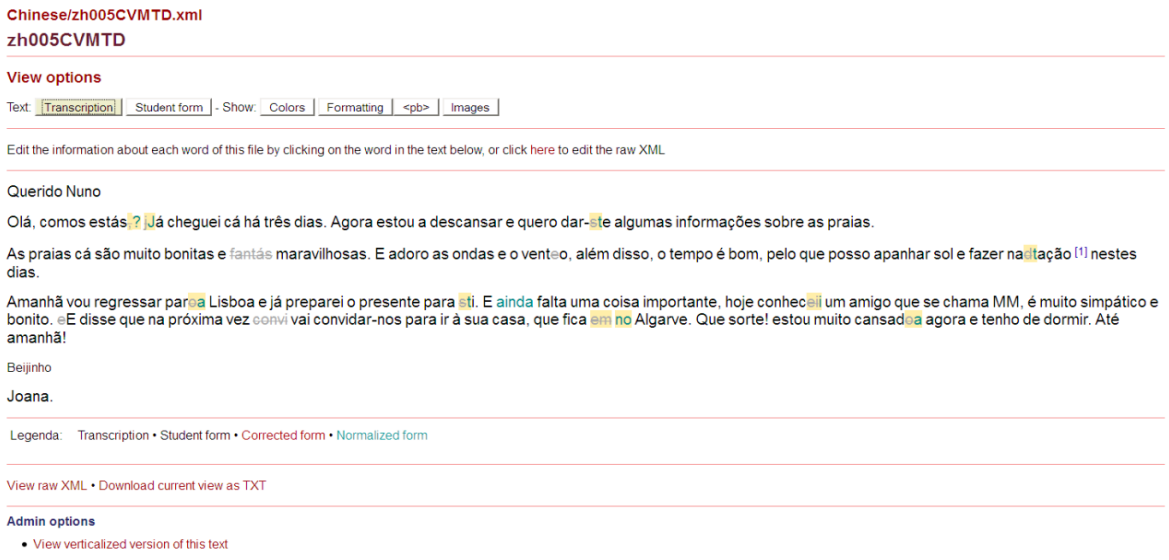

Figura 4: Ambiente gráfico da plataforma TEITOK para o *corpus* escrito

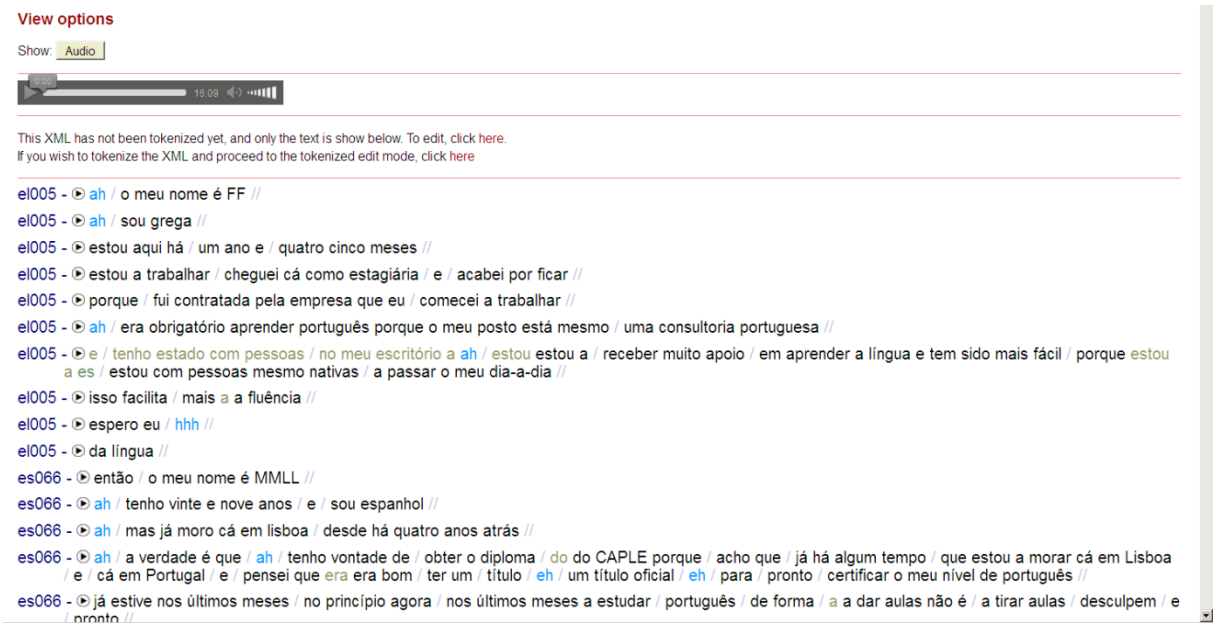

Figura 5: Ambiente gráfico da plataforma TEITOK para o *corpus* oral

No que respeita à anotação linguística dos dados, a lematização e anotação morfossintática são feitas de modo automático (após treino do anotador). A anotação morfossintática inclui as etiquetas das categorias gramaticais (acrónimos das categorias, em inglês), seguidas das etiquetas de traços nominais (género e número), verbais (tempo, modo, pessoa e número) e da flexão do infinitivo (flexionado ou não flexionado), de acordo com a anotação adotada para o *Corpus de Referência do Português Contemporâneo*[13] (Mendes *et al*., 2014). O exemplo 1 ilustra a anotação de uma frase do *corpus* (zh005CVMTD).

(1) As/A/DA#fp praias/PRAIA/CN#fp cá/CÁ/ADV são/SER/V#pi-3p muito/MUITO/ADV bonitas/BONITO/ADJ#fp e/E/CJ maravilhosas/MARAVILHOSO/ADJ#fp ./PNT

Relativamente aos processos de normalização ortográfica e codificação do erro, estes serão feitos manualmente. Tendo em consideração que a eficácia do ensino de uma língua estrangeira

[13] http://alfclul.clul.ul.pt/CQPweb/

depende da consciência, por parte dos formadores, das principais dificuldades enfrentadas pelos aprendentes, este *corpus* – a par dos principais *corpora* de aprendizagem existentes para outras línguas – pretende anotar todos os erros encontrados nos diversos níveis de análise linguística (ortografia, léxico, morfologia flexional e derivacional, sintaxe, semântica, questões estilísticas, pontuação) e corrigi-los, sempre que possível.

Para tal, faz parte do plano de trabalho:

(i) a elaboração de um esquema tipológico de etiquetas de erro, baseado em trabalhos como os de Tono (2003), Nicholls (2003), Dagneaux *et al*. (2005) e Rosen *et al*., (2013);

(ii) a elaboração de um manual de correção do *corpus*;

(iii) a inserção manual das etiquetas de erro e respetiva correção.

A tipologia de erros incluirá, pelo menos, dois níveis de anotação: (i) o nível linguístico (léxico, sintaxe, semântica, etc.); (ii) o tipo de erro (erro ortográfico, erro de concordância, erro na seleção do verbo auxiliar, erro na seleção do verbo copulativo, erro na seleção da preposição, erro na ordem de palavras, palavras em falta ou redundantes, etc.). Os exemplos 2-7 ilustram alguns erros encontrados no *corpus*.

(2) Detasto ladrão (zh001CVETF)

(3) No meu país as pessoas são menos social (nl001CVETF)

(4) Às vezes, acho que nós estamos cruéis para os animais (zh059CAATI)

(5) quanto custa o alugamento? (es003CVETD)

(6) por isso fiquei na casa (ru001CVMTD)

(7) as ondas são grandes, mas eles não perigosos porque esta praia tem permanentemente vigilancia dos nadadores-salvadores (de002CVMTD)

5. Aplicações

O COPLE2 constituirá um recurso importante para o estudo do ensino e da aprendizagem do português LE/L2, na medida em que inclui uma grande variedade de L1s, o que permite a

realização de estudos com base na Análise Contrastiva Interlíngua (Granger, 1996). Para uma determinada língua (neste caso, o português), esta análise consiste em comparar: (i) dados produzidos por informantes nativos e não-nativos (L1 vs. L2), de modo que se possa observar, mais facilmente, quaisquer produções tipicamente não-nativas produzidas pelos aprendentes (no que respeita aos dados de português L1, é utilizado o *corpus* CRPC); (ii) dados produzidos por falantes não-nativos com diversas línguas maternas (L2 vs. L2), de forma a poder-se avaliar se alguns dados são influenciados pela L1 dos informantes ou se são produzidos por outros aprendentes, em geral, independentemente da L1, o que permitirá despistar falsos fenómenos de transferência (Jarvis, 2000; Paquot, 2013).

Assim, este *corpus* visa fornecer dados acessíveis a professores e/ou investigadores que permitam realizar trabalhos de natureza linguística variada, como a identificação de erros comuns na aprendizagem de PLE/L2 e de erros que possam resultar de transferências da língua materna ou de outras línguas estrangeiras previamente adquiridas. Neste sentido, estão já em curso estudos baseados neste *corpus* sobre a aquisição de vogais (Castelo *et al.*, em prep.), as unidades multilexicais (Antunes & Mendes, em prep.), as construções relativas (Alexandre & Pinto, 2014), as construções copulativas (Alexandre & Gonçalves, 2015) e o papel da L1 e da L2 na aquisição lexical de português L3 (Pinto, em prep.).

Estudos desta natureza possibilitarão o desenvolvimento de aplicações e materiais didáticos na área do ensino de PLE/L2, adequando estratégias de ensino a um público-alvo específico. A observação dos dados sobre a intervenção do professor na correção dos textos auxiliará o desenvolvimento de materiais no âmbito da formação de professores, apelando à sensibilização e consciencialização de erros comuns dos aprendentes em geral, ou de falantes de uma língua em particular. Este projeto constituirá, igualmente, um recurso importante no que respeita ao acesso a materiais que ilustrem a interação escrita/oralidade, pouco frequentes no contexto de ensino de PLE/L2.

Referências

Abuhakema, G., R. Faraj, A. Feldman & E. Fitzpatrick (2008) "Annotating an Arabic Learner Corpus for Error". *Proceedings of the Sixth International Language Resources and Evaluation* (LREC 2008). Pp. 1347-1350.

Alexandre, N. & J. Pinto (2014). "Aspects of relative clauses in Portuguese as a foreign language by Chinese learners". *20th Conference of the European Association for Chinese Studies*, 22-26 julho, Braga/Coimbra. Disponível em http://www.clul.ul.pt/research-teams/547?lang=en.

Alexandre, N. & A. Gonçalves (2015) "Copular constructions in Portuguese as a second language (PL2) by Chinese learners: Do typological differences matter?". Comunicação apresentada no *Workshop on Copulas across Language*s, 18-19 de junho, University of Greenwich, London, Inglaterra.

Antunes, S. & A. Mendes (em prep.) "Portuguese Multiword Expressions: data from a learner corpus". Póster aceite para apresentação na *LCR2015: Third Learner Corpus Research Conference*, 11-13 de setembro de 2015, Radboud University, Nijmegen, Holanda.

Berber Sardinha T. (2001) "O Corpus de Aprendiz Br-Icle". http://www2.lael.pucsp.br/~tony/temp/publications/2001bricle-interc.pdf

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

Castelo, A., R. Santos & M. J. Freitas (em prep.) "O uso de vogais ortográficas por aprendentes de Português como língua estrangeira: unidade na diversidade". Comunicação aceite para apresentação no Congresso *Língua Portuguesa: Unidade na diversidade – Cultura, Literatura, História, Linguística, Tradução e Ensino*, 5-6 de novembro de 2015, Lublin, Polónia.

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

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

Delais-Roussarie E. & H. Yoo (2010) "The COREIL corpus: a learner corpus designed for studying phrasal phonology and intonation". In: Dziubalska-Kołaczyk K., M. Wrembel & 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 & 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., E. Dagneaux, F. Meunier & M. Paquot (eds.) (2009) *International Corpus of Learner English*. Version 2. UCL: Presses Universitaires de Louvain.

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

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

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

Leiria, I. (2001) *Léxico – aquisição e ensino do Português Europeu língua não materna*, Dissertação de Doutoramento em Linguística Aplicada. Faculdade de Letras da Universidade de Lisboa.

Lozano, C. (2009) "CEDEL2: Corpus Escrito del Español L2". In: Bretones Callejas, C. M. *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.

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

Nicholls, D. (2003) "The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT". In D. Archer, P. Rayson, A. Wilson, & T. McEnery (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.

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

Pinto, J. (em prep.) "O papel da L1 e da L2 na aquisição lexical de português L3". Comunicação aceite para apresentação no *Congreso Internacional de Literatura, Lengua y Traducción "liLETRAd" 2015*, 7-8 de julho de 2015, Universidade de Sevilha, Espanha.

Rosen, A., J. Hana, B. Štindlová & A. Feldman (2013) "Evaluating and automating the annotation of a learner corpus". In *Languages Resources and Evaluation 48(1)*. Pp.1-23.

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., P. Rayson, A. Wilson, & T. McEnery (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.