Corpus de Português Língua Estrangeira / Língua Segunda – COPLE2

Amália Mendes, Sandra Antunes, Nélia Alexandre, António Avelar, Adelina Castelo, Inês Duarte, Maria João Freitas, Anabela Gonçalves, José Pascoal, Jorge Pinto, Maarten Janssen

Centro de Linguística da Universidade de Lisboa
Instituto de Cultura e Língua Portuguesa
Centro de Avaliação de Português Língua Estrangeira

1. Apresentação

O *Corpus de Português Língua Estrangeira/Língua Segunda – COPLE2* é um projeto em desenvolvimento na Faculdade de Letras da Universidade de Lisboa (FLUL), sendo, atualmente, financiado pela Fundação Calouste Gulbenkian, através do projeto LeCIEPLE. Consiste na compilação de materiais produzidos no âmbito dos cursos de Português Língua Estrangeira do Instituto de Cultura e Língua Portuguesa (ICLP) e dos exames de acreditação do Centro de Avaliação de Português Língua Estrangeira (CAPLE).

2. Constituição do *corpus*

O COPLE2 incluirá, numa primeira fase, um acervo de aproximadamente **1000 textos escritos** produzidos por cerca de **500 alunos** de PLE/L2 que frequentaram a FLUL (ICLP e CAPLE) entre os anos de **2010-2012.**

Numa segunda fase, proceder-se-á à transcrição de cerca de **300 gravações**, que será ainda acompanhada do **alinhamento entre o texto e o sinal acústico**.

Devido à heterogeneidade dos aprendentes, exerceu-se um controlo rigoroso sobre as variáveis dos informantes e dos textos para tornar mais transparente a interpretação dos dados.

2.1. Perfil dos informantes

➢ idade entre os 18 e os 40 anos (80% 18-30 anos; 20% 31-40 anos);

➢ 14 L1 diferentes (mínimo de 6 informantes por L1, para o registo escrito);

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	27	Polaco	8
Japonês	23	Coreano	6
Francês	23	Romeno	6

➢ nacionalidade, relevante no caso de línguas que são faladas em vários países (ex. chinês: China; Macau; Hong Kong; Taiwan);

➢ habilitações académicas;

➢ conhecimento de outras línguas estrangeiras;

➢ proficiência (inicial, elementar, intermédio, avançado, superior);

➢ tipo de curso (anual ou de verão);

➢ anos de estudo de português.

2.2. Perfil dos textos

➢ registo (argumentativo, informativo, narrativo, carta pessoal, carta formal, diálogo, *e-mail*, crítica de uma obra);

➢ tópico;

➢ tipo de tarefa (teste diagnóstico, intercalar ou final, trabalho de casa, exame CAPLE);

➢ condições da tarefa (limite de tempo para redigir);

➢ recurso a materiais didáticos auxiliares (dicionários, gramáticas, outros).

3. Transcrição dos dados

Após a seleção e digitalização dos textos, procedeu-se à sua transcrição em **formato XML** de acordo com as normas de transcrição estabelecidas pela ***Text Encoding Initiative*** **(TEI)**. Cada ficheiro contém um cabeçalho com os metadados detalhados, o texto produzido pelo informante e a sua codificação.

As transcrições têm como principais objetivos:

❖ dar fielmente conta de todas as modificações assinaladas pelo informante durante a produção do texto (apagamentos, adições, alternativas, etc.);

❖ assinalar as correções ou comentários introduzidos pelo professor;

❖ anonimizar todos os dados pessoais, bem como eventuais referências críticas no interior dos textos.

Ficheiro XML: <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.

Versão definitiva do aluno: 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.

4. Normalização, lematização, anotação PoS e codificação do erro

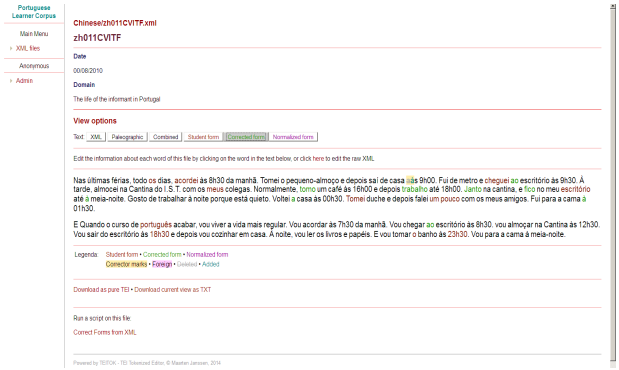
Será possível visualizar diferentes versões dos textos na plataforma TEITOK:

➢ código XML; paleográfica; produzida pelo aluno; corrigida pelo professor (já disponíveis);

➢ normalizada quanto à ortografia;

➢ lematizada e anotada morfossintaticamente (o texto produzido pelo aluno);

➢ codificação do erro a partir de um esquema tipológico.



Pretende-se que o COPLE2 possa ser pesquisado através do sistema CQP nos seus diversos níveis de anotação.

5. Aplicações

O COPLE2 visa fornecer dados acessíveis a professores e/ou investigadores que permitam realizar estudos de natureza linguística variada:

➢ identificar erros gerais na aprendizagem de PLE/L2 e erros que possam resultar de transferências da língua materna ou de outras línguas estrangeiras previamente adquiridas;

➢ desenvolver aplicações e materiais didáticos na área do ensino do português, adequando estratégias de ensino a um público-alvo específico;

➢ fornecer dados sobre a intervenção do professor na correção dos textos;

➢ usar materiais que ilustrem a interação escrita/oralidade, pouco frequentes no contexto de ensino de PLE/L2.

Projeto UID/LIN/00214/2013

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA
FUNDAÇÃO CALOUSTE GULBENKIAN

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