Corpus de Português Língua Estrangeira/Língua Segunda – COPLE2[1]

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

Este trabalho visa apresentar o *Corpus de Português Língua Estrangeira/Língua Segunda* (COPLE2), em compilação na Faculdade de Letras da Universidade de Lisboa. Este *corpus*, constituído por materiais recolhidos no âmbito dos cursos de Português Língua Estrangeira do Instituto de Língua e Cultura Portuguesa (ICLP) e dos exames de acreditação do Centro de Avaliação de Português Língua Estrangeira (CAPLE), tem como objetivo servir de base para estudos linguísticos sobre a aprendizagem do português por falantes estrangeiros e para o desenvolvimento de aplicações na área do ensino do português.

A produção de *corpora* de aprendizagem das línguas tem conhecido um crescente interesse, embora a maioria dos recursos produzidos vise a língua inglesa. É o caso do *corpus* que constitui uma referência nesta área, o *International Corpus of Learner English* - ICLE (Granger, 2003), ou do *Longman Learner's Corpus (LLC)*, entre outros. No caso do português, podemos citar Leiria (2001), o *corpus Recolha de dados de PLE*[2], o *Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2)*[3] e o *Corpus de Aquisição de L2* (CAL2)[4].

O *corpus* COPLE2 inclui, por agora, um subconjunto de materiais escritos e orais que integram um acervo de aproximadamente 1000 textos produzidos por cerca de 500 alunos de PLE/L2 que frequentaram a FLUL (ICLP e CAPLE) em 1999-2011 e continuará, no futuro, a ser alimentado por novas produções de alunos. Nesta primeira fase, procedeu-se ao tratamento dos textos escritos, que foram organizados por (i) tipos de tarefas e níveis de língua (testes diagnósticos, testes intercalares, exames, etc.) e (ii) língua materna dos informantes. A primeira fase envolveu a digitalização dos originais e a sua transcrição em formato XML. Cada ficheiro de transcrição contém um cabeçalho com metadados detalhados em formato XML, que cobrem informação sobre o informante, o texto produzido e o seu tratamento. Foram estabelecidas normas de transcrição do original, que permitem dar fielmente conta de todas as modificações assinaladas pelo informante durante a produção do seu texto

[1] Este trabalho foi desenvolvido no Centro de Linguística da Universidade de Lisboa ((UID/LIN/00214/2013) e financiado pela Fundação Calouste Gulbenkian (Proc. n.º 134655), pela Fundação para a Ciência e Tecnologia e pela Associação para o Desenvolvimento da Faculdade de Letras da Universidade de Lisboa.

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

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

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

(apagamentos, adições, alternativas) e também das correcções ou comentários introduzidos pelo professor. As transcrições foram anonimizadas, bem como eventuais referências no interior dos textos produzidos. Tanto os metadados como a transcrição estão de acordo com as normas da *Text Encoding Initiative* (TEI). Além disso, cada transcrição dispõe ainda de uma versão TXT, com o texto que o informante pretendia ser o definitivo. Damos em (1) um excerto de uma transcrição em formato XML de um texto produzido por um falante cuja língua materna é o mandarim.

(1) A minha mãe <del hand="corrector">soube</del> <remarks>professor: marca</remarks> essas palavras<add hand="zh001">não</add> foram normais imediatamente e <add hand="corrector">"</add> abraçou <add hand="corrector">"</add> a bolsa dela.

Numa segunda fase, a transcrição das gravações de exames orais será acompanhada do alinhamento entre o texto e o sinal acústico. Todos os textos serão ainda lematizados e anotados ao nível morfossintáctico com ferramentas desenvolvidas para o português (Généreux et al., 2012). A partir dos dados compilados, será desenvolvida uma tipologia de erros que servirá de base a um esquema de anotação (Tono, 2003) a aplicar ao *corpus*, também em formato XML. Pretende-se que o COPLE2 possa ser pesquisado *online*, nos seus diversos níveis de anotação.

Assim, o COPLE2 visa fornecer dados acessíveis a professores e/ou investigadores que poderão aí (i) identificar erros gerais na aprendizagem de PLE/L2 e erros que possam resultar diretamente da interferência da língua materna ou de outras línguas estrangeiras previamente adquiridas; (ii) aproveitar a informação disponível para desenhar materiais e adequar estratégias de ensino a um público-alvo específico; (iii) usar materiais que ilustrem a interação escrita/oralidade, pouco frequentes no contexto de ensino de PLE/L2. Estão já em curso, ou serão em breve iniciados, estudos, baseados neste *corpus*, sobre a aquisição de vogais, da flexão verbal, das unidades multilexicais (idiomáticas ou não), das expressões com valor modal e, ainda, sobre construções relativas e infinitivas.

Referências bibliográficas

Granger, Sylviane (2003) The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition. *TESOL Quarterly* 37(3): 538–546.

Leiria, Isabel Maria Caetano (2001) *A aquisição por falantes de Português Europeu língua não materna dos aspectos verbais expressos pelos Pretéritos Perfeito e Imperfeito*, Dissertação de Mestrado em Linguística Portuguesa Descritiva, Faculdade de Letras da Universidade de Lisboa.

Tono, Yukio (2003) Learner corpora: Design, development and applications. In Archer, Dawn, Paul Rayson, Andrew Wilson and Tony McEnery (eds.). 2003. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University: University Centre for Computer Corpus Research on Language., 800–809.

Généreux, M.; Hendrickx, I. & Mendes, A. (2012). A Large Portuguese Corpus On-Line: Cleaning and Preprocessing. In Caseli, H. et al. (eds.) *Computational Processing of the Portuguese Language. Proceedings of the 10th International Conference PROPOR2012*. Berlin, Heidelberg: Springer-Verlag, pp. 113-120.