CEPLEXicon
A lexicon of Child European Portuguese

Ana Lúcia Santos, Maria João Freitas, Aida Cardoso
Universidade de Lisboa (FLUL / CLUL)

als@letras.ulisboa.pt, joaofreitas@letras.ulisboa.pt, aidacard@gmail.com

Introduction

Main goals: offering the community a resource that compiles the lexical information contained in two different corpora and presenting it in a format that enables research in different areas (e.g., linguistics, speech therapy, education).

CEPLEXicon is registered (ISLRN: 408-817-203-152-3, ELRA ID: ELRA-L0094) and available through the ELRA catalogue. This is a free resource distributed by ELRA.

Building the Lexicon

Corpora

CEPLEXicon is based on two different corpora of child and child-directed speech: the **Santos corpus** (Santos, 2006; Santos et al., 2014) and the ***Child – Adult Interaction* corpus** (Freitas et al., 2012; built from the Freitas corpus (Freitas, 1997)).

This lexicon results from the automatic tagging of those two corpora, which include the speech produced by (i) **7 monolingual Portuguese children** aged between 1;02.00 and 3;11.12; (ii) **114 transcription files**, each corresponding to 40-50 minutes of child-adult interaction in a naturalistic setting (**86 hours of spontaneous speech**).

Santos database: transcribed according to the CHILDES (Child Language Data Exchange System) system and using the CLAN software (MacWhinney, 2000, http://childes.psy.cmu.edu/); this database is available in the CHILDES database (http://childes.talkbank.org/data/Romance/Portuguese/).

Freitas database: orthographically transcribed using EXMARaLDA (http://www.exmaralda.org/), according to transcription rules based on the CHILDES norms (Freitas et al., 2012).

Morphosyntactic Annotation

I. Automatic Tagging

POS-tagger: statistically trained on written corpora (644K tokens) and produced in the research unit ANAGRAMA (Centro de Linguística da Universidade de Lisboa – CLUL) (Généreux, Hendrickx & Mendes, 2012), using a set of 80 POS-tag labels. The same POS-tag labels were used when tagging the child and child-directed speech corpora to ensure adequacy and uniformity between corpora.

The tagger automatically generates a morphosyntactic tier in which a lemma and a POS-tag is assigned to each word of the transcription tier (1):

(1) *MAE: e mais?
%xmor: CJ|e ADV|mais ?
*TOM: e pa(ra) a praia.
%xmor: CJ|e PREP|para DA|a CN|praia . [Tomás 2;4.0]

a) Morphosyntactic categories: subset of the POS-tags previously used in the annotation of other corpora produced by CLUL, such as CRPC (Généreux, Hendrickx & Mendes, 2012; http://alfclul.clul.ul.pt/CQPweb/doc/CRPCmanual.v1_en.pdf).

b) lexicon's dimension: **98200 words**, from which **2201 lemas** were extracted.

c) Evaluation: **94.9%** of precision for the **POS-tagger** and **98%** of precision for the **lemmatizer** (Santos *et al*., 2014).

II. Manual Revision

The automatic tagging was submitted to a partial manual revision. The main corrections can be grouped in three categories.

a) clear cases of errors concerning the lemma and/or the POS-tag;

E.g.: "V|aleija" ('hurt'$_{PRESENT}$) → "V|aleijar" ('hurt'$_{INFINITIVE}$);
"ADJ|banheira" ('bathtub' tagged as an adjective) → "CN|banheira"
('bathtub' tagged as a common noun);

b) cases of ambiguity were verified against the transcription;

E.g.: "colar" is ambiguous between the common noun 'necklace' ("CN|colar") and the infinitive form of the verb 'to glue' ("V|colar").

c) words associated to POS-tags that may be expected to be infrequent in child speech before four years were manually verified against the transcription.

E.g.: all the occurrences of "REL|qual" ('which' tagged as a relative pronoun) were manually checked.

Structure of the Lexicon

- CEPLEXicon is available in .xls format and provides the following information:

(i) list of words (lemmas) produced by seven children, displayed in alphabetical order;

(ii) POS-tag corresponding to each lemma;

(iii) number (N) of occurrences of each lemma in three different age periods: <2 years; ≥ 2 and < 3 years; ≥ 3 years;

(iv) frequency (%) of each lemma in each age period: <2 years; ≥ 2 and < 3 years; ≥ 3 years;

(v) age of the first occurrence of each lemma for each child (year, month and day);

CEPLEXicon contains **2201 lemmas**, including **1043 common nouns**, **130 adjectives** and **303 verbs**.

➲ CEPLEXicon can be a relevant resource in different areas: (i) development of assessment and intervention resources in clinical contexts (e.g., speech therapy); (ii) development of didactic materials to be used by pre-school teachers in the classroom; (iii) development of educational games (e.g., children's books, software).

➲ CEPLEXicon was already used as a baseline reference in different projects.

- *Tracking Studies and Validation of the MacArthur-Bates Communicative Development Inventories for European Portuguese* (PTDC/MHC-PED/4725/2012, FCT, COMPETE e FEDER).
- *Cross-linguistic Child Phonology Project – EP* (IGAC 67/2014), under the Cross-linguistic Child Phonology Project - funding Conseil de Recherches en Sciences Humaines du Canada (#410-2009-0348214/2013 and UID/LIN/00214/2013);
- phonological assessment tool developed by Ramalho, Almeida & Freitas (2014) (SFRH/BD/88966/2012, Pest-OE/LIN/UI0).
- Afonso (2015): CEPLEXicon was used to validate the lexical stimuli included in the phonological awareness assessment tools proposed by the author in her PhD project.

Acknowledgments

The present work was developed within the **FCT funded project *Complement Clauses in the Acquisition of Portuguese*** **(PTDC/CLE-LIN/120897/2010)**, developed at Centro de Linguística da Universidade de Lisboa.

References: **Afonso, C. (2015)**. Complexidade Fonológica – Tarefa de Consciência Fonológica em Crianças do 1.º Ano do Ensino Básico. Ph.D. Dissertation. Universidade de Lisboa. ● **Généreux, M., I. Hendrickx & A. Mendes (2012)**. Introducing the Reference Corpus of Contemporary Portuguese On-Line. In *Proceedings of the 8th International Conference on Language Resources and Evaluation - LREC 2012*. ELRA, 2237-2244. ● **Freitas, M.J. (1997)**. *Aquisição da estrutura silábica do Português Europeu*. Ph.D. Dissertation. Universidade de Lisboa. ● **Freitas, M. J., A. Tanganho, M. Rocha & P. Oliveira (2012)**. Child-Adult Interaction: A Database on European Portuguese, CLUL, Anagrama. ● **MacWhinney, B. (2000)**. *The CHILDES Project: Tools for Analyzing Talk.* Mahwah / New Jersey: Lawrence Erlbaum Associates, 3rd Edition. ● **Ramalho, A. M., L. Almeida & M. J. Freitas (2014).** *Cross-linguistic Child Phonology Project – European Portuguese.* University of British Columbia, CLUL, IGAC registration number: 67/2014. ● **Santos, A. L. (2006)**. *Minimal Answers. Ellipsis, Syntax and Discourse in the Acquisition of European Portuguese*. Ph.D. Dissertation. Universidade de Lisboa. (Published 2009, Amsterdam/Philadelphia: John Benjamins). ● **Santos, A. L., M. Génereux, A. Cardoso, C. Agostinho & S. Abalada (2014)**. A corpus of European Portuguese child and child-directed speech. In *Proceedings of the 9th Conference on Language Resources and Evaluation – LREC 2014.* ELRA.

Av. Professor Gama Pinto, 2, 1649-003 Lisboa, Portugal

Tel.: +351 999 999 999 | URL: www.clul.ul.pt