CEPLEXicon

A Lexicon of Child European Portuguese

Ana Lúcia Santos, Maria João Freitas, Aida Cardoso
Faculdade de Letras da Universidade de Lisboa / CLUL
Universidade de Lisboa
Alameda da Universidade, 1600-214 Lisboa
E-mail: als@letras.ulisboa.pt, joaofreitas@letras.ulisboa.pt, aidacard@gmail.com

Abstract

CEPLEXicon (version 1.1) is a child lexicon resulting from the automatic tagging of two child corpora: the corpus Santos (Santos, 2006; Santos et al. 2014) and the corpus *Child – Adult Interaction* (Freitas et al. 2012), which integrates information from the corpus Freitas (Freitas, 1997). This lexicon includes spontaneous speech produced by seven children (1;02.00 to 3;11.12) during approximately 86h of child-adult interaction. The automatic tagging comprised the lemmatization and morphosyntactic classification of the speech produced by the seven children included in the two child corpora; the lexicon contains information pertaining to lemmas and syntactic categories as well as absolute number of occurrences and frequencies in three age intervals: < 2 years; ≥ 2 years and < 3 years; ≥ 3 years. The information included in this lexicon and the format in which it is presented enables research in different areas and allows researchers to obtain measures of lexical growth. CEPLEXicon is available through the ELRA catalogue.

Keywords: acquisition, lexicon, European Portuguese

1. Introduction

In this paper, we present a new lexicon for European Portuguese, representative of the lexicon of children aged 1;02 to 3;11[1]: CEPLEXicon. This lexicon contains 2 201 lemmas, which were based on the automatic tagging of 98 200 words. It is already registered (ISLRN: 408-817-203-152-3, ELRA ID: ELRA-L0094) and available through the ELRA catalogue[2].

The main goal of this work is to provide lexical information about child speech produced in a naturalistic setting. This child lexicon was built from a set of longitudinal data collected in the 1990s by researchers of the Faculdade de Letras da Universidade de Lisboa, namely the Santos corpus (Santos, 2006; Santos et al., 2014) and the corpus *Child – Adult Interaction* (Freitas et al., 2012).

Longitudinal data is an important source of information about the development of language in L1 acquisition. Hence, corpora built from longitudinal data, as the two corpora previously mentioned, play an important role in the evaluation of the acquisition process, since they help determining stages in linguistic development and they contribute to the study of the emergence and stabilization of specific linguistic structures.

Hence, CEPLEXicon was built taking into account the contribution of longitudinal data towards a better understanding of lexical development and the importance of making available more information about the acquisition of the lexicon in an accessible format. Since many of the corpora available through formats such as CHILDES and PHONBANK are heavily coded (with phonetic, syntactic, morphological information), our goal was to offer the community a resource that compiles the lexical information contained in two different corpora and to present this information in a format that enables research in different areas, such as linguistics, speech therapy, education, among others.

2. Building the Lexicon

In this section, we describe the corpora which were the basis for the lexicon now presented. In addition, we describe the process of construction of the lexicon, which included a process of automatic tagging and posterior manual revision.

2.1 Corpora

As it was previously mentioned, 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), the latter built from the Freitas corpus (Freitas, 1997)[3]. This lexicon results from the automatic tagging of these two corpora, which include the speech produced by seven monolingual Portuguese children aged between 1;02.00 and 3;11.12. This amounts to a total of 114 files, each corresponding to 40-50 minutes of child-adult interaction in a naturalistic setting (in a total of approximately 86 hours of spontaneous speech). In table 1, we summarize the information concerning the different speech samples which were the basis for the lexicon.

[1] We use the following convention to indicate children's age: yy;mm.day

[2] http://catalog.elra.info/product_info.php?products_id=1244.

[3] For a detailed description of these corpora, see Santos (2006), Santos et al. (2014), Freitas (1997) and Freitas et al. (2012).

Children	Age
Inês I.	1;06.06 – 3;11.12
Inês M.	1;05.09 – 2;09.03
Tomás	1;06.18 – 3;10.16
Laura	2;02.30 – 3;03.10
Marta	1;02.00 – 2;02.17
Pedro	2;07.00 – 3;07.24
Raquel	1;10.02 – 2;11.21

Table 1: Speech samples included in the lexicon, defined according to age.

Regarding the transcription of the corpora, the Santos database was originally 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/); the files from the Freitas database included here are currently orthographically transcribed using EXMARaLDA (http://www.exmaralda.org/), according to transcription rules largely based on the CHILDES norms (see Freitas et al., 2012 for a detailed description).

2.2 Automatic Tagging

This lexicon results from the automatic tagging of the two corpora previously mentioned, which comprised the lemmatization and morphosyntactic classification of each word in the corpora, in a total of about 98 200 words. The first experience in the automatic tagging of these corpora was done with the Santos corpus and is described in Santos et al. (2014). The tagger used for this task was trained on written corpora, which was produced in the research unit ANAGRAMA (Centro de Linguística da Universidade de Lisboa – CLUL) (Généreux, Hendrickx & Mendes, 2012). The POS-tagger was statistically trained on 644K tokens from a written corpus using a set of 80 POS-tag labels (these labels were used to tag different corpora produced at CLUL, namely the CRPC corpus[4] – Généreux, Hendrickx & Mendes, 2012). The same POS-tag labels were used when tagging the child and child-directed speech corpora, in order to ensure adequacy and uniformity between corpora (we will get back to POS-tag labels in section 2.4).

However, as orthographic transcriptions of speech and especially child and child-directed speech represent a challenge for any system statistically trained on written material, it was necessary to adapt the lemmatizer-tagger to the specificities of this particular type of data, through hand-crafted rules; the results obtained achieved 94.9% of precision for the POS-tagger and 98% of precision for the lemmatizer (a detailed description is found in Santos et al., 2014). The automatically tagged version of the Santos corpus is now available online in the CHILDES database. Given the good results obtained with the adaptation of the automatic tagger to child and child-directed speech, the same process was applied to the Freitas database.

As a result of the POS-tagging, each transcription file of both corpora has a morphosyntactic tier, which is the output generated by the tagger. In this tier, a lemma and a POS-tag is assigned to each word of the transcription tier, as illustrated in (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]

As the example shows, each word is assigned a POS-tag corresponding to a morphosyntactic category (e.g., "CN" indicates that a word is a common noun), followed by a vertical bar and a lemma (e.g., praia 'beach'). Usually the lemma of each word is the masculine singular form (in the case of nouns or adjectives, for instance) or the infinitive (in the case of verbs). During the tagging process, some specific annotations and metadata introduced during the transcription process were either removed or by-passed. For example, symbols like "xxx", denoting unintelligible speech, were disregarded (Santos et al., 2014), as shown in (2).

(2) CHILD: xxx quer bo(n)eca.
%xmor: V|querer CN|boneco .
[Inês 2;3.22]

2.3 Partial Manual Revision

Following the process of automatic annotation, all the words included in the speech of children in the two corpora were extracted, along with the information on lemmas and morphosyntactic category. The lemmas and POS-tags resulting from the automatic tagging were then submitted to a partial manual revision.

The main corrections resulting from the revision task can be described as follows. Firstly, clear cases of errors (concerning the lemma and/or the POS-tag) produced by the tagger were corrected and the necessary changes were introduced both in the lexicon file and in the corresponding transcription file (in the morphosyntactic tier). For example, cases such as "V|laleija" ('hurt'$_{\text{PRESENT}}$) were changed to "V|laleijar" ('hurt'$_{\text{INFINITIVE}}$), since the lemma of a verb must be the infinitive form. In the same sense, an occurrence such as "ADJ|banheira" ('bathtub' tagged as an adjective) was changed to "CN|banheira" ('bathtub' tagged as a common noun), to correct the error in the automatic attribution of the POS-tag.

Secondly, cases of ambiguity (for example, between the morphosyntactic categories verb/noun or noun/adjective) were verified against the transcription and the POS-tag and/or lemma were corrected according to the context. For instance, "colar" is ambiguous given that it may correspond to the common noun 'necklace' (in which case the word should be tagged as "CN|colar") or to the infinitive form of the verb 'to glue' (in which case it

[4] http://www.clul.ul.pt/en/resources/183-crpc.

should be tagged as "V|colar"). This type of ambiguity is relatively frequent and, for this reason, it was important to manually check such cases.

Finally, cases of words associated to POS-tags that may be expected to be infrequent in child speech before four years were equally manually verified against the corresponding transcription file. For instance, we expect a child younger than four to produce words such as *qual* 'which' or *onde* 'where' as interrogative pronouns, but is less likely that the same child would produce these forms as relative pronouns. For this reason, all the occurrences of "REL|qual" ("qual" tagged as a relative pronoun) and of "REL|onde ("onde" tagged as a relative pronoun) were manually checked.

Nevertheless, it is important to point out that this was a partial revision and that any automatic annotation implies an error rate. Hence, the tags of very frequent words, such as *a*, which can be ambiguous between the feminine singular form of the definite article 'the', the feminine singular of the accusative clitic pronoun, or the preposition 'to', were not exhaustively verified, since this would be an excessively time consuming task. Nevertheless, at this point it is worth remembering the evaluation performed on the results of the tagger for this type of data, which allowed us to expect high accuracy in this type of frequent ambiguous words (Santos et al., 2014).

2.4 Lemmas and Tags Used in the Lexicon

As already stated, the set of POS-tags used in the automatic tagging of the data was previously used in the annotation of other corpora produced by CLUL, such as CRPC. Nevertheless, some decisions regarding the lexicon presentation were made concerning e.g. verbs or closed class categories, in order to standardize the way in which data were presented in the lexicon file. These decisions can be summarized as follows.

In the case of closed class categories, the automatic lemmatizer used here assumes different lemmas corresponding to the masculine and the feminine forms. In order to keep consistency, this lemmatization rule was not changed in the morphosyntactic tier in the transcription files, but in the presentation of the lexicon the occurrences of masculine and feminine forms in closed classes were grouped under the masculine singular form. For example, the indefinite pronoun *outro* 'other$_{MASC-SG}$' and *outra* 'other$_{FEM-SG}$' were grouped under the lemma *outro* 'other$_{MASC-SG}$'. The only exception to this rule is the lemma of possessive pronouns, because the feminine forms of possessive pronouns are irregular (e.g., *meu* 'mine$_{MASC-SG}$' and *minha* 'mine$_{FEM-SG}$'). For this reason, the occurrences of the masculine and the feminine forms were kept separately under different lemmas, thus following the same general rule applied to all irregular feminine forms of nominal classes.

In the case of verbs, the automatic tagger assigned different POS-tags to different verb forms: (i) the tag "V" identifies a tensed verb form of a main verb; (ii) "VAUX" is assigned to the auxiliary verbs in compound tenses; (iii) "INF" identifies an infinitive verb form; (iii) "GER" is assigned to gerunds; (iv) "PPA" identifies a past participle form; and (v) "PPT" is assigned to past participles in compound tenses. Although these tags were kept in the morphosyntactic tier of the original transcription files, some of these categories were merged in the lexicon. In this sense, we kept the tags "V", "VAUX" and "PPA", thus allowing for a distinction between main and auxiliary verbs, on the one hand, and also between past participle forms (not in compound tenses) and other verb forms. Every occurrence of verb forms originally assigned other tags ("INF", "GER", and "PPT") were grouped under the tag "V".

Finally, certain compound nouns (usually tagged as proper nouns – "PNM"), for instance *Branca de Neve* 'Snow White' or *Aquário Vasco da Gama* 'Vasco da Gama Aquarium', were considered as a single lemma. This option allowed us to calculate the frequency of these proper nouns as a unit and prevented the erroneous inclusion in the lexicon of words that are part of these nouns, such as prepositions (for example *de* 'of', in *Branca de Neve* 'Snow White').

The list of resulting POS-tags conserved in the lexicon, with examples of words included in each morphosyntactic category, is presented in the appendix section of this paper (see table 2). We should insist on the fact that, apart from blending certain categories along the lines described in this section, no other changes were made to the POS-tag classes generated by the automatic tagger. Therefore, what is presented in table 2 corresponds to the subset of the POS-tag list originally used by the automatic tagger which was maintained in CEPLEXicon.

3. Structure of the Lexicon

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

1) List of words (lemmas) produced by seven children, displayed in alphabetical order.
2) POS-tag corresponding to each lemma.
3) Number (N) of occurrences of each lemma in three different age periods: <2 years; ≥ 2 and < 3 years; ≥ 3 years.
4) Frequency (%) of each lemma in each age period: <2 years; ≥ 2 and < 3 years; ≥ 3 years.
5) Age of the first occurrence of each lemma for each child (year, month and day).
6) Observations.

The 2 201 lemmas which were retrieved include 1043 common nouns (and 375 proper nouns), 302 verbs (and 74 past participles and 1 auxiliary verb, in distinct categories), 130 adjectives and 57 adverbs.

The way in which this lexicon is presented allows researchers to obtain quick measures of lexical growth. For instance, if an open class category such as common noun is taken into account, and the lexicon of a particular child, e.g. TOM, followed from 1;06 to 3;10, is under study, the results will show that only 197 (28% of the total

707 common noun lemmas documented in this child's lexicon) are produced before 2;00 and that this number reaches 325 (46% of the total common nouns) in the period between 2;00 and 2;11. As for verbs (past participle forms and auxiliaries excluded), a total of 226 lemmas are documented in this child's lexicon; nevertheless, only 44 (20%) lemmas are documented in the period before 2;00, but this number reaches 128 (57%) in the period between 2;00 and 2;11.

On the other hand, a researcher may be interested in determining the common lexicon of all the children included in the corpora, in a particular age range. If, for instance, we are interested in verbs produced before 2;00, the search will show that only 12 lemmas were attested in the speech of all the five children whose speech before 2;00 was included in the lexicon. Of course, in this case, general frequency effects and the diversity of situations of the data collection (which always corresponded to naturalistic settings) constrain the results. However, we believe that the information provided by this lexicon is a useful tool for researchers in language acquisition, as well as for researchers in the area of speech therapy and clinical linguistics in general.

4. Conclusion

CEPLEXicon is a resource available to the community that provides information on lexicon (including lemmas and morphosyntactic categories). This lexicon focuses on the L1 acquisition by monolingual children between 1 and 4 years of age, thus providing information on lexical development. This resource can be relevant in different areas, including:

(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).

In fact, CEPLEXicon was already used as a baseline reference in the project *Tracking Studies and Validation of the MacArthur-Bates Communicative Development Inventories for European Portuguese* (PTDC/MHC-PED/4725/2012, FCT, COMPETE e FEDER), which is currently developing the adaptation of the MacArthur-Bates Communicative Development Inventories for European Portuguese. The CEPLEXicon was also used in the validation process of the phonological assessment tool developed by Ramalho, Almeida & Freitas (2014) at CLUL (*Cross-linguistic Child Phonology Project – EP*, registration IGAC 67/2014), under the Cross-linguistic Child Phonology Project, coordinated by M. Bernardht and J. Stemberger at the University of British Columbia (funding Conseil de Recherches en Sciences Humaines du Canada (#410-2009-0348); in the case of the Portuguese tool, SFRH/BD/88966/2012, Pest-OE/LIN/UI0214/2013 and UID/LIN/00214/2013). Moreover, CEPLEXicon was used in Afonso (2015) to validate the lexical stimuli included in the phonological awareness assessment tools proposed by the author.

This is a free resource distributed by ELRA. The full reference to CEPLEXicon should be included in all types of work using it as a source of information, according to the manual and the contract established with ELRA.

5. Acknowledgements

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.

6. Bibliographical 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., Hendrickx, I. and Mendes, A. (2012). Introducing the Reference Corpus of Contemporary Portuguese On-Line. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation - LREC 2012*. European Language Resources Association (ELRA), pp. 2237--2244.

Freitas, M.J. (1997). *Aquisição da estrutura silábica do Português Europeu*. Ph.D. Dissertation. Universidade de Lisboa.

Freitas, M.J., Tanganho, A., Rocha, M. and Oliveira, P. (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 and 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*. PhD. Dissertation. Universidade de Lisboa. Publication: 2009, Amsterdam / Philadelphia: John Benjamins.

Santos, A.L., Génereux, M., Cardoso, A., Agostinho, C. and Abalada, S. (2014). A corpus of European Portuguese child and child-directed speech. In *Proceedings of the 9th Conference on Language Resources and Evaluation – LREC 2014*. European Language Resources Association (ELRA).

7. Language Resource References

Santos, A.L., Freitas, M.J. and Cardoso, A. (2014). CEPLEXicon – A Lexicon of Child European Portuguese. Lisboa: Anagrama (CLUL, FLUL). ISLRN 408-817-203-152-3, ELRA ID: ELRA-L0094.

8. Appendix

Tag	Morphosyntactic Category	Examples
ADJ	Adjectives	bom, brilhante, eficaz…
ADV	Adverbs	hoje, sim, felizmente…
CARD	Cardinals	zero, dez, cem, mil…
CJ	Conjunctions	e, ou, mas, porque…
CL	Clitics	o, lhe, se…
CN	Common Nouns	computador, cidade, ideia…
DA	Definite Articles	o, os, a, as.
DEM	Demonstratives	este, esses, aquele…
DFR	Denominators of Fractions	meio, terço…
DM	Discourse Marker	pronto, enfim…
EXC	Exclamatives	que, quanto...
IA	Indefinite Articles	uns, umas…
IND	Indefinites	tudo, alguém, ninguém…
INT	Interrogatives	quem, como, quando…
ITJ	Interjection	olá, fogo…
LTR	Letters	a, b, c…
MGT	Magnitude Classes	unidade, dúzia, resma…
MTH	Months	Janeiro, Dezembro…
ORD	Ordinals	primeiro, centésimo…
PADR	Part of Address	rua, avenida…
PNM	Part of Name (proper nouns)	Lisboa, António, João…
POSS	Possessives	meu, teu, seu…
PPA	Past Participles not in compound tenses	pintado, afirmados, vivida…
PREP	Prepositions	de, para, desde , em…
PRS	Personals	eu, tu, ele…
QNT	Quantifiers	todos, muitos, nenhum…
REL	Relatives	que, cujo, quem…
STT	Social Titles	Presidente, dr., prof….
UM	"um" or "uma"	um, uma
UNIT	Measurement units in abbreviated form	Kg, h, seg, Hz, Mbytes...
VAUX	Finite "ter" or "haver" in compound tenses	temos, havia…
V	Verbs (other than PPA, PPT, INF or GER)	falou, falaria…
WD	Week Days	segunda, terça-feira, sábado…

Table 2: List of POS-tags from CRPC corpus included in CEPLEXicon