A corpus of European Portuguese child and child-directed speech

Ana Lúcia Santos*, Michel Généreux, Aida Cardoso*, Celina Agostinho*, Silvana Abalada***
*Universidade de Lisboa (FLUL / CLUL) / ** Universidade de Lisboa (CLUL) and EURAC Research
Faculdade de Letras da Universidade de Lisboa — Institute for Specialised Communication and Multilingualism
Alameda da Universidade, 1600-214 Lisboa — Viale Druso1, 39100 Bolzano/Italy
E-mail: als@fl.ul.pt, michel.genereux@eurac.edu, aidacard@gmail.com, cfm.agostinho@gmail.com, silvanaabalada@gmail.com

Abstract

We present a corpus of child and child-directed speech of European Portuguese. This corpus results from the expansion of an already existing database (Santos, 2006). It includes around 52 hours of child-adult interaction and now contains 27,595 child utterances and 70,736 adult utterances. The corpus was transcribed according to the CHILDES system (Child Language Data Exchange System) and using the CLAN software (MacWhinney, 2000). The corpus itself represents a valuable resource for the study of lexical, syntax and discourse acquisition. In this paper, we also show how we used an existing part-of-speech tagger trained on written material (Généreux, Hendrickx & Mendes, 2012) to automatically lemmatize and tag child and child-directed speech and generate a line with part-of-speech information compatible with the CLAN interface. We show that a POS-tagger trained on the analysis of written language can be exploited for the treatment of spoken material with minimal effort, with only a small number of written rules assisting the statistical model.

Keywords: acquisition, child corpus, part-of-speech-tagging

1. Introduction

The main purpose of this paper is to present a new database of child and child-directed speech, known as the SANTOS database, which was transcribed according to the CHILDES[1] (Child Language Data Exchange System) system and using the CLAN software (MacWhinney, 2000), and which is an enlarged version of the database of Santos (2006). This database has already been used by the author and collaborators as the basis of research on syntactic and discourse development; here, we present an enlarged and enriched version of the same database, which now includes part-of-speech tagging.

As tools to annotate automatically child spoken material are still in their infancy, so to speak, especially for Portuguese (Branco et al., 2012), a second purpose of this work is to show how we can use existing tools developed for more widely available data. Porting existing tools to annotate data substantially different from the training set is not a trivial matter, given that training and target sets of data differ in two aspects: written versus spoken and adult versus child. Our approach is based on the post-processing of the output provided by a statistical model using a set of rules designed after a careful examination of the corpus. Portability of NLP (Natural Language Processing) tools between closely related languages is also an important area of research.

First, we present the corpus in detail, including options of data collection and transcription. In the second part, we describe how this corpus was lemmatized and tagged using a general-purpose tagger trained on written text and adapted to work on speech data.

[1] http://childes.psy.cmu.edu/.

2. Constitution of the corpus

The first version of this corpus resulted from a Ph.D. project (Santos, 2006). It included 52 files, each corresponding to 45-50 minutes of child-adult interaction (more than 40 hours of speech), and containing the spontaneous production of three different monolingual children acquiring European Portuguese (INI – age ranging from 1;6.6 to 3;11.12; TOM – age ranging from 1;6.18 to 2;9.7; INM – age ranging from 1;5.9 to 2;7.24). The data were collected using videotape and correspond to child-adult interaction in a naturalistic setting: children were taped at their homes interacting with their family (most often their mother) and the researcher. The data from one of the children (INI) were collected by Maria João Freitas (Freitas, 1997). The data from the other two children were collected between 1999 and 2002 by the author of the original database. These children were videotaped every other week, even though only one videotape per month was selected for transcription.

The original CLAN files in the corpus developed for Santos (2006) contained only orthographic transcriptions, carried out by the author. Since the data was meant to serve research on syntax and the syntax-discourse interface, all adults and children utterances were transcribed. The initial transcription of INI and TOM was based on audio copies of the video files. The transcription of data from INM was based on DVD copies of the video files. Given the better quality of the DVD copies, all the transcripts from INI and TOM were compared with the video DVD files at the end of 2004. The data from INM were also subjected to a revision using additional information provided by video. Finally, the transcription always progressed from the earlier to the later stages, since this strategy facilitated the process and ultimately

improved accuracy.
The original database of Santos (2006) contained 18,492 child utterances. The mean length of utterances in words (MLUw) in this database is presented in Table 1, according to the counts made available by the author.

Child	Age	MLUw	Number of files	Number of child's utterances
INI	1;6.6 - 3;11.12	1.527 - 3.815	21	6,591
TOM	1;6.18 - 2;9.7	1.286 - 2.954	16	6,800
INM	1;5.9 - 2;7.24	1.315 - 2.370	15	5,101

Table 1 – Spontaneous production in the original database.

Within the same Ph.D. project, data corresponding to more advanced stages of acquisition were collected for the same children, along with data from other children, but these data could not be transcribed at the time.
The corpus that we now present is an extension of this initial corpus. This enlarged version of the corpus includes not only more data but also new facilities, namely sound-transcription alignment and tagging. Occasionally, revisions were also made to the original corpus, resulting in small changes, e.g. in values of MLUw.

First, this enlarged version includes 15 new files with orthographic transcriptions, which were added to the initial data (corresponding to an increase of 12 hours of child-adult interaction). Transcription was based on the video files and was performed by one researcher and independently assessed by another researcher. All the cases in which both researchers did not agree were signalled and subjected to discussion, after which a final decision was taken or the case was marked as doubtful. In order to align the transcription with sound, the 15 new files and the files of TOM and INM from the original corpus (a total of 46 files) were converted from the original videos (Hi8 format) to digital video and audio. The digital videos are QuickTime files (mpeg 4 format) with a H.264 codec and an AAC audio codec, with dimensions of 480x360 pixels at 25 frames per second. The digital audios are in wave format, with 16 bit mono at 44 KHz sampling rate. Sound-transcription alignment was carried out with sound-text linking facilities within the CLAN software. For the time being, only the files for TOM and INM are linked to sound, but we intend to extend this facility to the entire database.

In Table 2, we present general information on the present corpus, namely age and MLUw (calculated with the *dates* and the *mlu* commands in CLAN). This corpus now includes 27,595 child utterances and also a total of 70,736 adult utterances.

Child	Age	MLUw	Number of files	Number of child's utterances
INI	1;6.6 - 3;11.12	1.530 - 3.827	21	6,591
TOM	1;6.18 - 3;10.16	1.286 - 3.089	30	15,548
INM	1;5.9 - 2;9.3	1.345 - 2.834	16	5,456

Table 2 – Child speech in SANTOS corpus.

3. POS-tagging and lemmatizing the corpus

In this section, we describe our work on automatically lemmatizing and tagging the corpus with part-of-speech (POS). The CLAN software includes the MOR program, a morphological analyser, with the possibility of building MOR grammars for each particular language. As there is no MOR grammar currently developed for European Portuguese, we propose a partial solution to this state-of-affairs by tagging (annotating) the corpus with lemmas and part-of-speech. Failing to have a MOR grammar for Portuguese, users of this corpus are now able to read and search the annotations we provide with the usual CLAN interface. Note that the tagger we used was trained statistically on large domain written material, so we have adapted the tagger by specializing it, at least partly, for child spoken material.

During the transcription process, various annotations and metadata were introduced. These annotations were either removed or by-passed. For example, the utterance:

(1) CHILD: xxx que(r) bo(n)eca
want doll
'(He?) wants the doll.'

is tagged as follows:

(2) %mor: V|querer CN|boneco
V|want.INF CN|doll

Since "xxx" denotes unintelligible speech and letters between parenthesis mean that the child did not pronounce the corresponding sound, these annotations were disregarded. Unintelligible speech is reintroduced in the transcription, albeit untagged. The fully-tagged utterance indicates that the word "quer" was assigned the POS-tag "Verb" and the lemma "querer", while the word "boneca" was assigned the POS-tag "Common Noun" and the lemma "boneco". Each utterance was tagged and lemmatized individually, which means that the tagger did not use context outside the utterance being currently analysed. Note that as in any automated tagging process, there are inevitably errors, which we will report later in

this paper.

The tagger we used was developed in our research group and is described in length in Généreux, Hendrickx & Mendes (2012). Here we will only highlight its main features. The POS-tagger was statistically trained on 644K tokens from a written corpus using a set of 80 POS-tag labels. The tagger has been evaluated and obtained an F-score of 0.954. The lemmatizer combines a machine learning algorithm with a lookup into a dictionary of 120,768 wordform-lemma combinations produced in-house. The lemmatizer has been evaluated and achieved an accuracy of 96.7%. As child-spoken data represents a serious challenge for any system statistically trained on written material, we decided to include a number of rules to assist the statistical model. Hand-crafted rules were applied directly on the results produced by the statistical model, mostly to provide specificities pertaining to child speech or in some cases to correct outright systematic errors. The rules are as follows, in no particular order:

1. a list of 80 words typically used as interjections were always tagged as such;
2. if the first word of an utterance is tagged as relative, change the tag for interrogative;
3. if the first word of an utterance is either "quando" 'when', "porque" 'why / because', "como" 'how / like' or "quanto" 'how much / as' and is tagged as a conjunction, change the tag for interrogative;
4. the lemmas "pronto" 'ready', "vá" 'go' or "olha" 'look' opening a sentence are always tagged as a discourse marker;
5. if a word is tagged as past participle not in compound tense and the lemma is "segurar,seguro" 'hold/secure', change the POS-tag for verb and the lemma for "segurar" 'hold';
6. the word "segura" 'hold / secure.FEM' should always be POS-tagged with verb and lemmatized to "segurar" 'hold';
7. if a word is POS-tagged as para-linguistic material and lemmatized as "queque" 'cake', change the POS-tag for common noun;
8. if a word is POS-tagged as a prepositional phrase and lemmatized as "porquê" 'why', change the POS-tag for interrogative;
9. if a word is lemmatized as "mamã" 'mommy', change its POS-tag for common noun;
10. if the word "se" 'if / CLITIC' is POS-tagged as a conjunction and follows a word POS-tagged as a verb, change the POS-tag for clitic.

4. Evaluation

In this section we provide an evaluation of the lemmatizer-tagger that we adapted for child spoken material. We tagged three files[2] picked randomly from our corpus, one file from each of the three different children. The files had a total of 21,972 tokens, 1,572 types and 4,736 utterances. These three files were revised manually by a human expert for tagging errors. We found a total of 1,128 POS-tagging errors, for a precision of 94.9%. We also found 442 lemmatizing errors, for a precision of 98%. These two results are in the same precision bracket as the evaluation we mentioned earlier made on written material, which is a very encouraging result. Table 3 below summarizes the ten most frequent POS-tagging errors and Table 4 the ten most frequent lemmatizing errors.

#Occur-rences	Word	Assigned tag	Corrected tag
148	que 'that'	Relative	Interrogative
52	olha 'look'	Verb	Discourse Marker
51	se 'CL' /'if'	Clitic	Conjunction
45	a 'PREP'/ 'the'	Preposition	Definite Article
36	a 'PREP'/ 'the'	Definite Article	Preposition
36	pois 'because'/ 'indeed'	Conjunc-tion	Adverb
26	onde 'where'	Relative	Interrogative
25	olha 'look'	Discourse Marker	Verb
25	outra 'other'	Adjective	Indefinite
24	quem 'who'	Relative	Interrogative

Table 3 – Ten most frequent POS-tagging errors.

Some of the POS-tagging errors are clearly related to the distinction between spoken and written data. For instance, "olha" 'look' and "pois" 'indeed' are frequently used in dialogues: in child-adult spoken interaction, 'olha' is frequently used to catch the child's attention and "pois" as an answer to a yes-no question or generally as an expression of agreement.

[2] Each file represents a full child speech production during one session.

#Occur-rences	Word	Lemma assigned	Lemma corrected
52	olha 'look'	olhar 'look.INF'	olha 'look'
25	olha 'look'	olha 'look'	olhar 'look.INF'
25	outra 'other.FEM'	outro 'other.MASC'	outra 'other.FEM'
21	conta 'tell'/ 'account'	conta 'account'	contar 'tell.INF'
12	foi 'was'/ 'went'	ser 'be'	ir 'go'
9	bolas 'balls' / 'to hell'	bolas 'to hell'	bola 'ball'
9	espera 'wait'/ 'delay'	espera 'delay'	esperar 'wait.INF'
9	gira 'turn' / 'cute.FEM'	girar 'turn.INF'	giro 'cute.MASC'
7	carrinho 'little car'	carrinho 'little car'	carro 'car'
6	abraci-nho 'little hug'	abracinho 'little hug'	abraço 'hug'

Table 4 – Ten most frequent lemmatizing errors.

Lemmatization errors are often caused by ambiguity of word forms and inherent to the POS-tagging model. This is exemplified with a case like "olha" 'look', which can be a verb or a discourse marker or the case of "foi", which can either be a form of the verb "ser" 'be' or "ir" 'go'. In some rare cases ("outra" 'other.FEM') the conflicting lemmas were normalized to be consistent with the general behaviour of the lemmatizer, which assumes different lemmas for the masculine and the feminine in the case of closed class categories.[3] The error rate for lemmatization therefore includes errors not specific to child spoken material.

[3] (http://alfclul.clul.ul.pt/CQPweb/doc/CRPCmanual.v1_en.pdf)

5. Conclusion

The database we have presented is a relevant resource for language acquisition research. Given the fact that it presents child and child-directed speech, reproducing the complete child-adult interaction in the original recordings, it may be a source of information on both the acquisition of syntax and the development of the syntax-discourse interface.

As far as tagging and lemmatizing go, our experiments showed that, given a set of well-crafted rules, a statistical model trained and developed for written material can be ported to POS-tag and lemmatize spoken data from children with almost the same performance. Only ten simple rules have been developed to assist the statistical model, which cannot be considered prohibitively high in man-hour cost. We would think that similar minor adjustments could be made to successfully bring other statistically trained systems for other languages to a par with their performance on the same type of material on which they were trained.

6. Acknowledgements

The present work was funded by FCT-Fundação para a Ciência e a Tecnologia, within the project *Complement Clauses in the Acquisition of Portuguese* (PTDC/CLE-LIN/120897/2010). Silvana Abalada is supported by the FCT grant FCT/SFRH/BD/80331/2011. The data corresponding to INI were collected by Maria João Freitas within the project PCSH/C/LIN/524/93, developed at Laboratório de Psicolinguística da Faculdade de Letras da Universidade de Lisboa. We are also grateful to the people who helped us with sound and video files, namely Ana Isabel Mata and João Miguel Santos.

7. References

Branco, A.; Mendes, A.; Pereira, S.; Henriques, P.; Pellegrini, T.; Meinedo, H.; Trancoso, I.; Quaresma, P. and Strube de Lima, V. L. (2012). A Língua Portuguesa na Era Digital / The Portuguese Language in the Digital Age, White Paper Series, Berlim, Springer, ISBN9783642295928.

Freitas, M.J. (1997). Aquisição da estrutura silábica do Português Europeu. 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.

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

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).