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

als@fl.ul.pt, michel.genereux@eurac.edu, aidacard@gmail.com, cfm.agostinho@gmail.com, silvanaabalada@gmail.com

Introduction

Goals

.to present a new corpus of child and child-directed speech: the SANTOS database (new version of the corpus of Santos (2006), which now includes part-of-speech tagging (POS)).

.to show how tools developed for more widely available data can be used to automatically annotate child spoken material.

Constitution of the corpus

First version of the corpus (Santos, 2006): 52 files (45-50 min. each) of child-adult interaction (>40 hours of speech), containing the spontaneous production of 3 monolingual children acquiring European Portuguese (EP).

Data Collection

The data were collected (every other week) using videotape and correspond to child-adult interaction in a naturalistic setting: INI (data collection – Freitas, 1997), TOM and INM (data collection – Santos, 2006).

Original Database

The files (one videotape per month) were orthographically transcribed by the author, according to the CHILDES system and using the CLAN software (MacWhinney, 2000).

The original corpus includes 18,492 child utterances.

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

New Version of the Database

This version of the corpus includes: (i) 15 new files with orthographic transcriptions (more 12 hours of speech), performed by one researcher and independently assessed by another one; (ii) sound-transcription alignment (TOM and INM); and (iii) POS tagging.

The enlarged version includes 27,595 child utterances and 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

POS-tagging and lemmatizing the corpus

There is no MOR grammar currently developed for EP within CLAN. Our proposal: a partial solution by tagging the corpus with lemmas and POS.

Corpus Processing

Each utterance was tagged and lemmatized individually (i.e. the tagger did not use context outside the utterance being currently analysed). Annotations and metadata removed or by-passed during the tagging process.

Tagger (developed by 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.

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%**.

Centro de Linguística
da Universidade de Lisboa

Rules

10 hand-crafted rules applied directly on the results produced by the statistical model to provide specificities pertaining to child speech or in some cases to correct outright systematic errors.

Example:
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.

Evaluation

3 tagged files picked randomly from our corpus (one file from each of the three different children): 21,972 tokens, 1,572 types, and 4,736 utterances. Revised manually by a human expert.

POS-tagging errors: 1,128, for a **precision of 94.9%**
Lemmatizing errors: 442 , for a **precision of 98%**

Results are in the same precision bracket as the evaluation made on written material.

Ten most frequent POS-tagging errors

#Ocurrences	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'	Conjunction	Adverb
26	onde 'where'	Relative	Interrogative
25	olha 'look'	Discourse Marker	Verb
25	outra 'other'	Adjective	Indefinite
24	quem 'who'	Relative	Interrogative

Source of the errors

Some of the POS-tagging errors are clearly related to the distinction between spoken and written data. (e.g. "olha" 'look', "pois" 'indeed')

Ten most frequent lemmatizing errors

#Occurrences	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	abracinho 'little hug'	abracinho 'little hug'	abraço 'hug'

Source of the errors
Lemmatization errors are often caused by ambiguity of word forms and inherent to the POS-tagging model. In some rare cases ("outra" 'other.FEM') the conflicting lemmas were normalized to be consistent with the general behaviour of the lemmatizer. The error rate for lemmatization therefore includes errors not specific to child spoken material.

Conclusion

The database we have presented is a relevant resource for language acquisition research, namely on the acquisition of syntax and the development of the syntax-discourse interface.

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 and we 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.

Acknowledgements

The present work was funded by 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 grant FCT/SFRH/BD/80331/2011. The INI data were collected by Maria João Freitas within the project PCSH/C/LIN/524/93, developed at Lab. de Psicolinguística (FLUL). We are grateful to the people who helped us with sound and video files: Ana Isabel Mata and João Miguel Santos.