Searching the Annotated Portuguese Childes Corpora

Rodrigo Wilkens
Institute of Informatics
Federal University of Rio Grande do Sul
Brazil
rswilkens@inf.ufrgs.br

1 Introduction

Annotated corpora of child language data are valuable resources for language acquisition studies, for instance, providing the basis for developmental comparisons and evaluation of different hypotheses. For computational investigations annotated corpora can serve as an approximation to the linguistic environment to which a child is exposed, as discussed by Wintner (2010).

Recently there has been a growing number of initiatives for annotating children's data for a number of languages, with for instance, part-of-speech (PoS) and syntactic information (Sagae et al., 2010; Buttery and Korhonen, 2007; Yang, 2010) and some of these are available as part of CHILDES (MacWhinney, 2000). For resource rich languages like English these annotations can be further extended with detailed information, for instance, from WordNet (Fellbaum, 1998) about synonymy, from the MRC Psycholinguistic Database (Coltheart, 1981) about age of acquisition, imagery, concreteness and familiarity among others. However, for many other languages one of the challenges is in annotating corpora in a context where resources and tools are less abundant and many are still under development.

In this paper we describe one such initiative, for annotating the raw Portuguese corpora available in the CHILDES database with (psycho)linguistic and distributional information (§2) . It also describes a modular searching environment for these corpora that allows complex and flexible searches that combine different levels of annotation, and that can be easily extended (§3). We finish with some conclusions and future work.

2 Resource Description

The Portuguese, CHILDES contains 3 corpora:

- Batoréo (Batoreo, 2000) with 60 narratives, 30 from adults and 30 from children, about two stories
- Porto Alegre (Guimarães, 1994; Guimarães, 1995) with data from 5 to 9 year old children, collected both cross-sectionally and longitudinally and
- Florianópolis with the longitudinal data for one Brazilian child: 5530 utterances in broad phonetic transcription.

The total number of sentences and words per age in these corpora is shown in Table 1

Table 1: Frequency of words and sentences per age in the Portuguese corpora

Age	words	sentences
0	0	0
1	7k	2k
2	8k	1k
3	0	0
4	1k	61
5	38k	1k
6	47k	1k
7	56k	1k
8	56k	1k

In order to annotate the transcribed sentences in the CHILDES Portuguese corpora we used the PALAVRAS parser[1] (Bick, 2000). It is a statistical robust Portuguese parser, which always return

[1]Tagset avaliable at http://beta.visl.sdu.dk/visl/pt/info/symbolset-manual.html.

Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss, pages 26–28,
Avignon, France, April 24 2012. ©2012 Association for Computational Linguistics

at least one analysis even for incomplete or ungrammatical sentences. This parser has a high accuracy: 99% for part-of-speech and 96-97%. The parser also has a named entity recognizer (Bick, 2003) and provides some semantic information for nouns, verbs and adjectives (e.g. organization, date, place, etc). The annotations process consisted of the following steps:

1. automatic pre-processing for dealing with incomplete words and removing transcription notes;

2. tagging and parsing with PALAVRAS parser;

3. annotation of verbs and nouns with psycholinguistic information like age of acquisition and concreteness from (Cameirao and Vicente, 2010).

For enabling age related analysis, the sentences were subsequently divided according to the child's age reported in each corpus, and annotated with frequency information collected considering separately each type of annotation per age.

3 System Description

In order to allow complex searches that combine information from different levels of annotation for each age, the sentences were organized in a database, structured as in Tables 2 and 3, respectively presenting the structure of words and sentences).

Table 2: Information about Words

Word
age of acquisition
part-of-speech
corpus frequency
frequency by age
adult frequency

Table 3: Information about Sentences

Sentence
children gender
PoS tags
dependency tree
semantic tags

Using a web environment, a user can choose any combination of fields in the database to perform a query. It is possible to examine, for instance, the usage of a particular word and its evolution according to grammatical class per age.

The environment provides two modes for queries: an expert mode, where database queries can be dynamically specified selecting the relevant fields, and a guided mode which contains predefined query components and a set of filters that users can combine in the queries. The results are available both as a file containing the relevant annotated sentences for further processing, or in a graphical form. The latter shows a chart of frequency per age, which can be displayed in terms of absolute or relative values.

The guided mode provides an iterative way for query construction where the user selects a relevant field (e.g. age of acquisition) at a time and adds it to the query until all desired fields have been added, when the resulting query is saved. The user can repeat this process to create combined queries and at the end of the process can chose the form for outputting the result (graphic or file).

4 Conclusion

This paper describes the (psycho)linguistic and distributional annotation of the Portuguese corpora in CHILDES, and presents an environment for searching them. This environment allows complex searches combining multiple levels of annotation to be created even by non-expert users. Therefore this initiative not only produced an integrated and rich annotation schema so far lacking for these corpora, but also provided a modular environment for structuring and searching them through a more user friendly interface. As next steps we foresee the extension of the annotation using other resources. We also plan to add corpora for other languages to the environment, such as English and Spanish.

Acknowledgements

This research was partly supported by CNPq Projects 551964/2011-1 and 478222/2011-4.

References

Batoréo, H. 2000. *Expressão do Espaço no Português Europeu. Contributo Psicolinguístico para*

o Estudo da Linguagem e Cognição. PhD Dissertation, Fundação Calouste Gulbenkian e Fundação para a Ciência e a Tecnologia, Ministério da Ciência e da Tecnologia, Lisboa

Bick, E. 2000. *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* [S.l.]: University of Arhus.

Bick, E. 2003. *Multi-level NER for Portuguese in a CG framework.* Proceedings of the Computational Processing of the Portuguese Language.

Briscoe, E., Carroll, J., and Watson, R. 2006. *The second release of the rasp system.* COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia.

Buttery, P., Korhonen, A. 2007. *I will shoot your shopping down and you can shoot all my tins– Automatic Lexical Acquisition from the CHILDES Database.* Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition. Association for Computational Linguistics.

Cameirao, M.L. and Vicente, S.G. 2010. *Age-of-acquisition norms for a set of 1,749 portuguese words.* Behavior research methods 42, Springer.

Coltheart, M. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33A:497–505.

Fellbaum, C. 1998. *WordNet An Electronic Lexical Database.*. The MIT Press, Cambridge, MA ; London.

Guimarães, A. M. 1994. *Desenvolvimento da linguagem da criança na fase deletramento.* Cadernos de Estudos Linguísticos, 26, 103-110

Guimarães, A. M. 1994. *The use of the CHILDES database for Brazilian Portuguese.* I. H. Faria & M. J. Freitas (Eds.), Studies on the acquisition of Portuguese. Lisbon: Colibri

MacWhinney, B. 2000. *The CHILDES project: tools for analyzing talk.* Lawrence Erlbaum Associates, second edition.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B. and Wintner, S. 2010. *Morphosyntactic annotation of CHILDES transcripts.* Journal of Child Language.

Wintner, S. 2010. *Computational Models of Language Acquisition.* CICLing'10.

Charles, Yang 2010. *Three factors in language variation.* Lingua.