An annotated English child language database

Aline Villavicencio[♣♠], Beracah Yankama[♠], Rodrigo Wilkens[♣], Marco A. P. Idiart[♣], Robert Berwick[♠]

[♣]Federal University of Rio Grande do Sul (Brazil)
[♠]MIT (USA)

alinev@gmail.com, beracah@mit.edu, rswilkens@gmail.com, marco.idiart@gmail.com, berwick@csail.mit.edu

1 Introduction

The use of large-scale naturalistic data has been opening up new investigative possibilities for language acquisition studies, providing a basis for empirical predictions and for evaluations of alternative acquisition hypotheses. One widely used resource is CHILDES (MacWhinney, 1995) with transcriptions for over 25 languages of interactions involving children, with the English corpora available in raw, part-of-speech tagged, lemmatized and parsed formats (Sagae et al., 2010; Buttery and Korhonen, 2005). With a recent increase in the availability of lexical and psycholinguistic resources and robust natural language processing tools, it is now possible to further enrich child-language corpora with additional sources of information.

In this paper we describe the English CHILDES Verb Database (ECVD), which extends the original lexical and syntactic annotation of verbs in CHILDES with information about frequency, grammatical relations, semantic classes, and other psycholinguistic and statistical information. In addition, these corpora are organized in a searchable database that allows the retrieval of data according to complex queries that combine different sources of information. This database is also modular and can be straightforwardly extended with additional annotation levels. In what follows, we discuss the tools and resources used for the annotation (§2), and conclude with a discussion of the implications of this initial work along with directions for future research (§3).

2 Linguistic and Statistical Properties

The English CHILDES Verb Database contains information about the English corpora in CHILDES parsed using three different pipelines: (1) MEGRASP; (2) RASP; and (3) the CHILDES Treebank. In the first, made available as part of the CHILDES distribution[1], the corpora are POS tagged (in %mor), and parsed using MEGRASP (Sagae et al., 2010) which provides information about dependency parses and grammatical relations (in %gra):[2]

```
*MOT:  I said (.) Adam you could have a banana
       and offer Robin and Ursula one (.)would you
       ?
%mor:  pro|I v|say&PAST n:prop|Adam pro|you
       aux|could v|have det|a n|banana ...
%gra:  1|2|SUBJ 2|6|CJCT 3|2|OBJ 4|6|SUBJ
       5|6|AUX 6|9|COORD 7|8|DET 8|6|OBJ ...
```

In the second pipeline, the RASP system (Briscoe et al., 2006) is used for tokenisation, tagging, lemmatization and parsing of the input sentences, outputting syntactic trees (in %ST) and grammatical relations (%GR).[3] In both examples each GR denotes a relation, along with its head and dependent:

```
*MOT:  oh no # he didn't say anything about win-
       dow .
%ST:   (T Oh:1 no:2 ,:3 (S he:4 (VP do+ed:5
       not+:6 say:7 anything:8 (PP about:9 (N1
       window:10)))) .:11)
%GR:   (|ncsubj| |say:7_VV0| |he:4_PPHS1| _)
       (|aux| |say:7_VV0| |do+ed:5_VDD|)
       (|ncmod| _ |say:7_VV0| |not+:6_XX|)
       (|iobj| |say:7_VV0| |about:9_II|) (|dobj|
       |say:7_VV0| |anything:8_PN1|) (|dobj|
       |about:9_II| |window:10_NN1|)
```

The third focuses on the Adam corpus from the Brown data set (Brown, 1973) and uses the Charniak parser with Penn Treebank style part of speech tags and output, followed by hand-curation, as described by Pearl and Sprouse (2012):

```
(S1 (SBARQ (WHNP (WP who)) (SQ (VP (COP is)
(NP (NN that)))) (. ?)))
```

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

[2] In an evaluation MEGRASP produced correct dependency relations for 96% of the relations in the gold standard, with the dependency relations being labelled with the correct GR 94% of the time.

[3] The data was kindly provided by P. Buttery and A. Korhonen and generated as described in (Buttery and Korhonen, 2005).

Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss, pages 23–25,
Avignon, France, April 24 2012. ©2012 Association for Computational Linguistics

The use of annotations from multiple parsers enables the combination of the complementary strengths of each in terms of coverage and accuracy, similar to inter-annotator agreement approaches. These differences are also useful for optimizing search patterns in terms of the source which produces the best accuracy for a particular case. Information about corpora sizes and the annotated portions for each of the parsers is displayed in table 1.

Information	Sentences
Total Raw	4.84 million
MEGRASP & RASP Raw	2.5 million
MEGRASP Parsed	109,629
RASP Parsed	2.21 million
CHILDES Treebank	26,280
MEGRASP & RASP Parsed	98,456

Table 1: Parsed Sentences

The verbs in each sentence are also annotated with information about shared patterns of meaning and syntactic behavior from 190 fine-grained subclasses that cover 3,100 verb types (Levin, 1993). This annotation allows searches defined in terms of verb classes, and include all sentences that contain verbs that belong to a given class. For instance, searching for verbs of running would return sentences containing not only *run* but also related verbs like *slide, roll* and *stroll*.

Additional annotation of properties linked to language use and recognition include extrinsic factors such as word frequency and intrinsic factors such as the length of a word in terms of syllables; age of acquisition; imageability; and familiarity. Some of this annotation is obtained from the MRC Psycholinguistic Database (Coltheart, 1981) which contains 150,837 entries with information about 26 properties, although not all properties are available for every word (e.g. IMAG is only available for 9,240 words).

For enabling complex search functionalities that potentially combine information from several sources, the annotated sentences were organized in a database, and Tables 2 and 3 list some of the available annotations. Given the focus on verbs, for search efficiency each sentence is indexed according to the verbs it contains. In addition, verbs and nouns are further annotated with information shown in table 3 whenever it is available in the existing resources.

These levels of annotation allow for complex searches involving for example, a combination of information about a verb's lemma, target grammatical relations, and occurrence of Levin's classes in the corpora.

Not all sentences have been successfully analyzed, and the comments field contains information about the missing annotations and cases of near perfect matches that arise from the parsers using different heuristics for e.g. non-words, meta-characters and punctuation. These required more complex matching procedures for identifying the corresponding cases in the annotations of the parsers.

Fields
Sentence ID
Corpus
Speaker
File
Raw sentence
MOR and POST tags
MEGRASP dep. and GRs
RASP syntactic tree
RASP dep. and GRs
Comments

Table 2: Information about Sentences

Fields
Word ID
Sentence ID
Levin's classes
Age of acquisition
Familiarity
Concreteness
Frequency
Imageability
Number of syllables

Table 3: Information about Words

3 Conclusions and future work

This paper describes the construction of the English CHILDES Verb Database. It combines information from different parsing systems to capitalize on their complementary recall and precision strengths and ensure the accuracy of the searches. It also includes information about Levin's classes for verbs, and some psycholinguistic information for some of the words, like age of acquisition, familiarity and imageability. The result is a large-scale integrated resource that allows complex searches involving different annotation levels. This database can be used to inform analysis, for instance, about the complexity of the language employed with and by a child as her age increases, that can shed some light on discussions about the poverty of the stimulus. This is an ongoing project to make the annotated data available to the research community in a user-friendly interface that allows complex patterns to be specified in a simple way.

Acknowledgements

This research was partly supported by CNPq Projects 551964/2011-1, 202007/2010-3, 305256/2008-4 and 309569/2009-5.

References

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

R. Brown. 1973. *A first language: The early stages.* Harvard University Press, Cambridge, Massachusetts.

P. Buttery and A. Korhonen. 2005. Large-scale analysis of verb subcategorization differences between child directed speech and adult speech. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes.*

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

B. Levin. 1993. *English verb classes and alternations - a preliminary investigation.* The University of Chicago Press.

B. MacWhinney. 1995. *The CHILDES project: tools for analyzing talk.* Hillsdale, NJ: Lawrence Erlbaum Associates, second edition.

L. Pearl and J. Sprouse, 2012. *Experimental Syntax and Islands Effects*, chapter Computational Models of Acquisition for Islands. Cambridge University Press.

K. Sagae, E. Davis, A. Lavie, B. MacWhinney, and S. Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(03):705–729.