On Parsing CHILDES

Aarre Laakso
Department of Psychology
Indiana University, Bloomington
1101 East 10th Street
Bloomington, IN 47405
alaakso@indiana.edu

Abstract

Research on child language acquisition would benefit from the availability of a large body of syntactically parsed utterances between parents and children. We consider the problem of generating such a "treebank" from the CHILDES corpus, which currently contains primarily orthographically transcribed speech tagged for lexical category.

1 Introduction

The CHILDES database (MacWhinney, 2000) is a 300-megabyte collection of electronic corpora of transcribed speech between parents and the children. It contains corpora in several genres (informal speech, narrative), from a wide range of situations (structured experiments, lunch room conversations, bathtime), from 26 languages, across a range of ages, and for a variety of special populations (bilinguals, second-language learners, clinical populations).

Although there are a few phonetically transcribed corpora, most of the corpora in the CHILDES database are transcribed orthographically. The English CHILDES corpora were recently released with a complete set of morphosyntactic tags identifying the lexical category (part of speech) and lemma (stem) of every uttered word. Tags were assigned using the MOR program that is part of the CHILDES distribution, then disambiguated using the POST program (Parisse and Le Normand, 2000). For example:

```
*MOT:   what's that ?
%mor:   pro:wh|what~v|be&3S pro:dem|that ?
*CHI:   yyy .
%mor:   unk|yyy .
*MOT:   it's a chicken .
%mor:   pro|it~v|be&3S det|a n|chicken .
*CHI:   yeah .
%mor:   co|yeah .
*MOT:   yeah .
%mor:   co|yeah .
```

As a consequence of the nature of the data that is available, most of the analyses conducted with CHILDES have focused on lexical and morphological phenomena. However, questions about the acquisition of grammar are at least as important, and many of the most important arguments for an innate Universal Grammar hinge on claims about the learnability of grammatical rules that depend on highly abstract structural principles and operations, including *c-command* (Crain and Pietroski, 2002) and *merge* and *move* (Baker, 2005). Having an accurate set of constituent structure parses for CHILDES would make it possible to determine the extent to which such phenomena are or are not modeled by parents, as well as the the ages (and order) in which they emerge in children's spontaneous production.

To our knowledge, however, there has been only one attempt to parse CHILDES (Sagae et al., 2004). That work focuses on the Eve corpus only and reports 65% accuracy using a rule-based grammar tuned to that corpus, combined with statistical disambiguation. Somewhat surprisingly, they report that using the part-of-speech tags from MOR/POST *decreased* accuracy to 57.5%. They explain this by noting that the accuracy of the part-of-speech tagger is about 95%, which means that there will be an

Submitted to Midwest Computational Linguistics Colloquium (MCLC) 4/10/2005.

incorrect tag in every 20 words on average.

2 Challenges with CHILDES

The relatively poor performance of the existing work on parsing CHILDES when compared with parsing written English — state-of-the-art parsers of written English now routinely exceed 90% accuracy — is understandble in light of the unique challenges of parsing CHILDES. The most difficult obstacle that must be overcome in any attempt to parse CHILDES is the fact that it is transcribed *conversation*, which means that it not only includes large numbers of indexicals and ambiguities but also many disfluencies, including fillers ("um", "ah"), parentheticals ("you know"), repetitions ("I hafta. . . , I hafta go. . . "), stuttering ("It. . . , it. . . , it. . . , it. . . , it's a dog"), repairs ("I went. . . I mean, you went. . . "), and other material (and omissions) characteristic of conversational speech. CHAT, the transcription system used in the CHILDES database, provides mechanisms for indicating such material, but they are not used consistently across the various corpora that make up the entire CHILDES collection because the corpora were collected by different researchers over many years. Thus, unlike the materials from which most existing treebanks have been developed (generally written materials such as newspaper articles), which consist mostly of grammatical sentences, CHILDES consists mostly of sentences that would be considered ungrammatical in writing.

Parsing CHILDES is further complicated by the fact that the transcribed speech consists of conversations between parents, typically mothers, and children of various ages, roughly from 12 months to 6 years. Even the adult utterances in the transcripts therefore contain a variety of features that are unusual not only in written text but also in adult conversation, including onomatopoeia ("woof-woof"), mothers' uses of child-invented forms ("baba" instead of "bottle"), singing ("la-la-la"), references to letters of the alphabet as objects ("That's an 'A"') and word play ("goobarumba") (MacWhinney, 2000).

Finally, there is no "gold standard" for parsing CHILDES because the goal is precisely to produce the *first* constituent structure parse of the database. In particular, there is no set of parse trees generated by human experts on which a statistical parser may be trained or against which the results of any parser may be scored.

3 Our contribution

In this paper, we report on attempts to parse CHILDES using a variety of parsers, including a simple context-free grammar, off-the-shelf broad-coverage parsers (including both rule-based parsers and statistical parsers) and a customized parser we are in the process of developing. On the assumption that parental speech would be more like written English than children's speech — and therefore that the parsing task would be easier for parental speech — we attempted to parse only parental speech. This assumption has also been made in previous attempts to parse CHILDES (Sagae et al., 2004), where it has been justified on the grounds that many studies of child language acquisition rely heavily, or even exclusively, on analysis of parental input.

The specific parsers we have tried include: a bottom-up chart parser using a simple hand-written Context Free Grammar; the XTAG Parser, based on a lexicalized Tree Adjoining Grammar (XTAG Research Group, 2001); the Link Grammar Parser, a robust parser for link grammers (Grinberg et al., 1995); the Collins statistical parser (Collins, 1999); Dan Bikel's Multilingual Statistical Parsing Engine (Bikel, 2002); and RASP, the Robust Accurate Statistical Parser (Briscoe and Carroll, 2002).

In general, the rule-based parsers performed poorly. As a baseline, we used a simple Context-Free Grammar (CFG) to parse directly from the lexical category / subcategory tags produced by MOR. This baseline CFG achieved approximately 15% coverage of the corpus and revealed a number of deficiencies in the tags assigned by MOR. Improving coverage would involve a laborious process of handcrafting rules and carefully balancing accuracy versus completeness. The XTAG parser got stuck in infinite loops on many utterances and had to be restarted repeatedly.

The statistical parsing approach was slightly more successful, achieving greater than 30% coverage in every case with the default models, which are based on parsed Wall Street Journal text from the Penn Treebank. We attribute this to the fact that such

parsers do not depend on strict grammatical rules but instead define probability distributions over them, so are better able to cope with "ungrammaticalities" in the input. It would be possible to train some of the statistical parsers on corpora that contain the same or similar kinds of "ungrammaticalities" — i.e., other transcribed speech corpora, such as the SWITCHBOARD corpus (Godfrey et al., 1992) — making the grammatical differences between transcribed speech and standard written language less significant. This work is in progress, but training on conversational data will not be a panacea, because of the differences between the kinds of utterances found in CHILDES and those found in adult conversation, as discussed above.

It would be theoretically possible to train the statistical parsers on the parsed Eve corpus. However, the available data on the Eve corpus is in the form of dependency structures over grammatical relations, not the full constituent structure trees expected by the training modules of the statistical parsers. We are in the process of converting the available dependency structures into constituent structure trees, but it is unlikely that it will prove valuable — the sparsity of training data (the Eve corpus represents less than 1% of the CHILDES database) will likely be a more serious problem.

In light of the difficulties with off-the-shelf parsers, we are in the process of developing our own statistical parser based on a neural network (Lane and Henderson, 2001) and complemented by lexical information and heuristics. For example, like Charniak and Johnson (2001), we treat the identification and removal of disfluencies as a pre-parsing operation as much as possible, on the grounds that the information that they contain is more likely to be a hindrance than a help in parsing. We are also using a "bootstrapping" approach whereby we iteratively train the parser on its own parses until improvement levels off. We are mitigating the risk that this will lead to innaccuracies by having linguistic experts verify randomly selected parses, which then become part of our growing "gold standard". This work is still underway, but we will report on progress made and remaining challenges. We hope that the attendees at MCLC will be able to suggest other ideas for meeting this difficult parsing challenge.

References

Mark C. Baker. 2005. Mapping the terrain of language learning. *Language Learning and Development*, 1(1):93–129.

Daniel M. Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of the Human Language Technology Conference*, San Diego.

Edward Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1499–1504, Las Palmas, Canary Islands.

Eugene Charniak and Mark Johnson. 2001. Edit detection and parsing for transcribed speech. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 118–126.

Michael John Collins. 1999. *Head-driven statistical models for natural language parsing*. Ph.d. dissertation, University of Pennsylvania.

Stephen Crain and Paul Pietroski. 2002. Why language acquisition is a snap. *Linguistic Review*, 19(1–2):163–183.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*, volume 1, pages 517–520, San Francisco.

Dennis Grinberg, John Lafferty, and Daniel Sleator. 1995. A robust parsing algorithm for link grammars. Technical Report CMU-CS-95–125, School of Computer Science, Carnegie-Mellon University.

Peter Lane and James Henderson. 2001. Incremental syntactic parsing of natural language corpora with simple synchrony networks. *IEEE Transactions on Knowledge and Data Engineering*, 13(2):219–231.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, volume 2: The Database. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition.

C. Parisse and M.-T. Le Normand. 2000. Automatic disambiguation of morphosyntax in spoken language corpora. *Behavior Research, Methods, Instruments & Computers*, 32:468–481.

Kenji Sagae, Brian MacWhinney, and Alon Lavie. 2004. Automatic parsing of parental verbal input. *Behavior Research Methods, Instruments and Computers*, 36(1):113–126.

XTAG Research Group. 2001. A lexicalized tree adjoining grammar for english. Technical report, IRCS, University of Pennsylvania.

