High-accuracy Annotation and Parsing of CHILDES Transcripts

Kenji Sagae
Department of Computer Science
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
sagae@is.s.u-tokyo.ac.jp

Eric Davis
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
dhdavis@cs.cmu.edu

Alon Lavie
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
alavie@cs.cmu.edu

Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
macw@cmu.edu

Shuly Wintner
Department of Computer Science
University of Haifa
31905 Haifa, Israel
shuly@cs.haifa.ac.il

Abstract

Corpora of child language are essential for psycholinguistic research. Linguistic annotation of the corpora provides researchers with better means for exploring the development of grammatical constructions and their usage. We describe an ongoing project that aims to annotate the English section of the CHILDES database with grammatical relations in the form of labeled dependency structures. To date, we have produced a corpus of over 65,000 words with manually curated gold-standard grammatical relation annotations. Using this corpus, we have developed a highly accurate data-driven parser for English CHILDES data. The parser and the manually annotated data are freely available for research purposes.

1 Introduction

In order to investigate the development of child language, corpora which document linguistic interactions involving children are needed. The CHILDES database (MacWhinney, 2000), containing transcripts of spoken interactions between children at various stages of language development with their parents, provides vast amounts of useful data for linguistic, psychological, and sociological studies of child language development. The raw information in CHILDES corpora was gradually enriched by providing a layer of morphological information. In particular, the English section of the database is augmented by part of speech (POS) tags for each word. However, this information is usually insufficient for investigations dealing with the syntactic, semantic or pragmatic aspects of the data.

In this paper we describe an ongoing effort aiming to annotate the English portion of the CHILDES database with syntactic information based on grammatical relations represented as labeled dependency structures. Although an annotation scheme for syntactic information in CHILDES data has been proposed (Sagae et al., 2004), until now no significant amount of annotated data had been made publicly available. In the process of manually annotating several thousands of words, we updated the annotation scheme, mostly by extending it to cover syntactic phenomena that occur in real data but were unaccounted for in the original annotation scheme.

The contributions of this work fall into three main categories: revision and extension of the annotation scheme for representing syntactic information in CHILDES data; creation of a manually annotated 65,000 word corpus with gold-standard syntactic analyses; and implementation of a complete parser that can automatically annotate additional data with high accuracy. Both the gold-standard annotated data and the parser are freely available. In addition to introducing the parser and the data, we report on many of the specific annotation issues that we encountered during the manual annotation pro-

Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, pages 25–32,
Prague, Czech Republic, June 2007

cess, which should be helpful for those who may use the annotated data or the parser. The annotated corpora and the parser are freely available from `http://childes.psy.cmu.edu/`.

We describe the annotation scheme in the next section, along with issues we faced during the process of manual annotation. Section 3 describes the parser, and an evaluation of the parser is presented in section 4. We analyze the remaining parsing errors in section 5 and conclude with some applications of the parser and directions for future research in section 6.

2 Syntactic annotation

The English section of the CHILDES database is augmented with automatically produced ambiguous part-of-speech and morphological tags (MacWhinney, 2000). Some of these data have been manually disambiguated, but we found that some annotation decisions had to be revised to facilitate syntactic annotation. We discuss below some of the revisions we introduced, as well as some details of the syntactic constructions that we account for.

2.1 The morphological annotation scheme

The English morphological analyzer incorporated in CHILDES produces various part-of-speech tags (there are 31 distinct POS tags in the CHILDES tagset), including ADJective, ADVerb, COmmunicator, CONJunction, DETerminer, FILler, Noun, NUMeral, ONomatopoeia, PREPosition, PROnoun, ParTicLe, QuaNtifier, RELativizer and Verb[1]. In most cases, the correct annotation of a word is obvious from the context in which the word occurs, but sometimes a more subtle distinction must be made. We discuss some common problematic issues below.

Adverb vs. preposition vs. particle The words *about, across, after, away, back, down, in, off, on, out, over, up* belong to three categories: ADVerb, PREPosition and ParTicLe. To correctly annotate them in context, we apply the following criteria.

First, a preposition must have a prepositional object, which is typically realized as a noun phrase (which may be topicalized, or even elided). Second, a preposition forms a constituent with its noun phrase object. Third, a prepositional object can be fronted (for example, *he sat on the chair* becomes *the chair on which he sat*), whereas a particle-NP sequence cannot (**the phone number up which he looked* cannot be obtained from *he looked up the phone number*). Finally, a manner adverb can be placed between the verb and a preposition, but not between a verb and a particle.

To distinguish between an adverb and a particle, the meaning of the head verb is considered. If the meaning of the verb and the target word, taken together, cannot be predicted from the meanings of the verb and the target word separately, then the target word is a particle. In all other cases it is an adverb.

Verbs vs. auxiliaries Distinguishing between Verb and AUXiliary is often straightforward, but special attention is given when tagging the verbs *be*, *do* and *have*. If the target word is accompanied by an non-finite verb in the same clause, as in *I have had enough* or *I do not like eggs*, it is an auxiliary. Additionally, in interrogative sentences, the auxiliary is moved to the beginning of the clause, as in <u>*have*</u> *I had enough?* and <u>*do*</u> *I like eggs?*, whereas the main verb is not. However, this test does not always work for the verb *be*, which may head a non-verbal predicate, as in *John* <u>*is*</u> *a teacher*, vs. *John* <u>*is*</u> *smiling*. In verb-participle constructions headed by the verb *be*, if the participle is in the progressive tense, then the head verb is labeled as auxiliary.

Communicators vs. locative adverbs COmmunicators can be hard to distinguish from locative adverbs, especially at the beginning of a sentence. Our convention is that CO must modify an entire sentence, so if a word appears by itself, it cannot be a CO. For example, utterances like *here* or *there* are labeled as ADVerb. However, if these words appear at the beginning of a sentence, are followed by a break or pause, and do not clearly express a location, then they are labeled CO. Additionally, in *here/there you are/go*, *here* and *there* are labeled CO.

2.2 The syntactic annotation scheme

Our annotation scheme for representing grammatical relations, or GRs (such as subjects, objects and adjuncts), in CHILDES transcripts is a slightly extended version of the scheme proposed by Sagae et al. (2004), which was inspired by a general annota-

[1] We use capital letters to denote the actual tag names in the CHILDES tagset.

tion scheme for grammatical relations (Carroll et al., 1998), but adapted specifically for CHILDES data. Our scheme contains 37 distinct GR types. Sagae et al. reported 96.5% interannotator agreement, and we do not believe our minor updates to the annotation scheme should affect interannotator agreement significantly.

The scheme distinguishes among SUBJects, (finite) Clausal SUBJects[2] (e.g., *that he cried moved her*) and XSUBJects (*eating vegetables is important*). Similarly, we distinguish among OBJects, OBJect2, which is the second object of a ditransitive verb, and IOBjects, which are required verb complements introduced by prepositions. Verb complements that are realized as clauses are labeled COMP if they are finite (*I think that was Fraser*) and XCOMP otherwise (*you stop throwing the blocks*). Additionally, we mark required locative adjectival or prepositional phrase arguments of verbs as LOCatives, as in *put the toys in the box/back*.

PREDicates are nominal, adjectival or prepositional complements of verbs such as *get*, *be* and *become*, as in *I'm not sure*. Again, we specifically mark Clausal PREDicates (*This is how I drink my coffee*) and XPREDicates (*My goal is to win the competition*).

Adjuncts (denoted by JCT) are optional modifiers of verbs, adjectives or adverbs, and we distinguish among non-clausal ones (*That's much better*; *sit on the stool*), finite clausal ones (CJCT, *Mary left after she saw John*) and non-finite clausal ones (XJCT, *Mary left after seeing John*).

MODifiers, which modify or complement nouns, again come in three flavors: MOD (*That's a nice box*); CMOD (*the movie that I saw was good*); and XMOD (*the student reading a book is tall*).

We then identify AUXiliary verbs, as in *did you do it?*; NEGation (*Fraser is not drinking his coffee*); DETerminers (*a fly*); QUANTifiers (*some juice*); the objects of prepositions (POBJ, *on the stool*); verb ParTicLes (*can you get the blocks out?*); ComPlementiZeRs (*wait until the noodles are cool*); COMmunicators (*oh, I took it*); the INfinitival *to*; VOCatives (*Thank you, Eve*); and TAG questions (*you know how to count, don't you?*).

[2]As with the POS tags, we use capital letters to represent the actual GR tags used in the annotation scheme.

Finally, we added some specific relations for handling problematic issues. For example, we use ENUMeration for constructions such as *one, two, three, go* or *a, b, c*. In COORDination constructions, each conjunct is marked as a dependent of the conjunction (e.g., *go and get your telephone*). We use TOPicalization to indicate an argument that is topicalized, as in *tapioca, there is no tapioca*. We use SeRiaL to indicate serial verbs as in *come see if we can find it* or *go play with your toys*. Finally, we mark sequences of proper names which form the same entity (e.g., *New York*) as NAME.

The format of the grammatical relation (GR) annotation, which we use in the examples that follow, associates with each word in a sentence a triple $i|j|g$, where i is the index of the word in the sentence, j the index of the word's syntactic head, and g is the name of the grammatical relation represented by the syntactic dependency between the i-th and j-th words. If the topmost head of the utterance is the i-th word, it is labeled `i|0|ROOT`. For example, in:

```
a          cookie     .
1|2|DET    2|0|ROOT   3|2|PUNCT
```

the first word *a* is a DETerminer of word 2 (*cookie*), which is itself the ROOT of the utterance.

2.3 Manual annotation of the corpus

We focused our manual annotation on a set of CHILDES transcripts for a particular child, Eve (Brown, 1973), and we refer to these transcripts, distributed in a set of 20 files, as the Eve corpus. We hand-annotated (including correcting POS tags) the first 15 files of the Eve corpus following the GR scheme outlined above. The annotation process started with purely manual annotation of 5,000 words. This initial annotated corpus was used to train a data-driven parser, as described later. This parser was then used to label an additional 20,000 words automatically, followed by a thorough manual checking stage, where each syntactic annotation was manually verified and corrected if necessary. We retrained the parser with the newly annotated data, and proceeded in this fashion until 15 files had been annotated and thoroughly manually checked.

Annotating child language proved to be challenging, and as we progressed through the data, we noticed grammatical constructions that the GRs could not adequately handle. For example, the original GR scheme did not differentiate between locative arguments and locative adjuncts, so we created a new GR label, LOC, to handle required verbal locative arguments such as *on* in *put it on the table*. *Put* licenses a prepositional argument, and the existing JCT relation could not capture this requirement.

In addition to adding new GRs, we also faced challenges with telegraphic child utterances lacking verbs or other content words. For instance, *Mommy telephone* could have one of several meanings: *Mommy this is a telephone*, *Mommy I want the telephone*, *that is Mommy's telephone*, etc. We tried to be as consistent as possible in annotating such utterances and determined their GRs from context. It was often possible to determine the VOC reading vs.the MOD (*Mommy's telephone*) reading by looking at context. If it was not possible to determine the correct annotation from context, we annotated such utterances as VOC relations.

After annotating the 15 Eve files, we had 18,863 fully hand-annotated utterances, 10,280 adult and 8,563 child. The utterances consist of 84,226 GRs (including punctuation) and 65,363 words. The average utterance length is 5.3 words (including punctuation) for adult utterances, 3.6 for child, 4.5 overall. The annotated Eve corpus is available at `http://childes.psy.cmu.edu/data/Eng-USA/brown.zip`. It was used for the *Domain adaptation task* at the CoNLL-2007 dependency parsing shared task (Nivre, 2007).

3 Parsing

Although the CHILDES annotation scheme proposed by Sagae et al. (2004) has been used in practice for automatic parsing of child language transcripts (Sagae et al., 2004; Sagae et al., 2005), such work relied mainly on a statistical parser (Charniak, 2000) trained on the Wall Street Journal portion of the Penn Treebank, since a large enough corpus of annotated CHILDES data was not available to train a domain-specific parser. Having a corpus of 65,000 words of CHILDES data annotated with grammatical relations represented as labeled dependencies allows us to develop a parser tailored for the CHILDES domain.

Our overall parsing approach uses a best-first probabilistic shift-reduce algorithm, working left-to-right to find labeled dependencies one at a time. The algorithm is essentially a dependency version of the data-driven constituent parsing algorithm for probabilistic GLR-like parsing described by Sagae and Lavie (2006). Because CHILDES syntactic annotations are represented as labeled dependencies, using a dependency parsing approach allows us to work with that representation directly.

This dependency parser has been shown to have state-of-the-art accuracy in the CoNLL shared tasks on dependency parsing (Buchholz and Marsi, 2006; Nivre, 2007)[3]. Sagae and Tsujii (2007) present a detailed description of the parsing approach used in our work, including the parsing algorithm. In summary, the parser uses an algorithm similar to the LR parsing algorithm (Knuth, 1965), keeping a stack of partially built syntactic structures, and a queue of remaining input tokens. At each step in the parsing process, the parser can apply a *shift* action (remove a token from the front of the queue and place it on top of the stack), or a *reduce* action (pop the two topmost stack items, and push a new item composed of the two popped items combined in a single structure). This parsing approach is very similar to the one used successfully by Nivre et al. (2006), but we use a maximum entropy classifier (Berger et al., 1996) to determine parser actions, which makes parsing extremely fast. In addition, our parsing approach performs a search over the space of possible parser actions, while Nivre et al.'s approach is deterministic. See Sagae and Tsujii (2007) for more information on the parser.

Features used in classification to determine whether the parser takes a shift or a reduce action at any point during parsing are derived from the parser's current configuration (contents of the stack and queue) at that point. The specific features used are:[4]

- Word and its POS tag: $s(1)$, $q(2)$, and $q(1)$.
- POS: $s(3)$ and $q(2)$.

[3]The parser used in this work is the same as the probabilistic shift-reduce parser referred to as "Sagae" in the cited shared task descriptions. In the 2007 shared task, an ensemble of shift-reduce parsers was used, but only a single parser is used here.

[4]$s(n)$ denotes the n-th item from the top of the stack (where $s(1)$ is the item on the top of the stack), and $q(n)$ denotes the n-th item from the front of the queue.

- The dependency label of the most recently attached dependent of: $s(1)$ and $s(2)$.

- The previous parser action.

4 Evaluation

4.1 Methodology

We first evaluate the parser by 15-fold cross-validation on the 15 manually curated gold-standard Eve files (to evaluate the parser on each file, the remaining 14 files are used to train the parser). Single-word utterances (excluding punctuation) were ignored, since their analysis is trivial and their inclusion would artificially inflate parser accuracy measurements. The size of the Eve evaluation corpus (with single-word utterances removed) was 64,558 words (or 59,873 words excluding punctuation). Of these, 41,369 words come from utterances spoken by adults, and 18,504 come from utterances spoken by the child. To evaluate the parser's portability to other CHILDES corpora, we also tested the parser (trained only on the entire Eve set) on two additional sets, one taken from the MacWhinney corpus (MacWhinney, 2000) (5,658 total words, 3,896 words in adult utterances and 1,762 words in child utterances), and one taken from the Seth corpus (Peters, 1987; Wilson and Peters, 1988) (1,749 words, 1,059 adult and 690 child).

The parser is highly efficient: training on the entire Eve corpus takes less that 20 minutes on standard hardware, and once trained, parsing the Eve corpus takes 18 seconds, or over 3,500 words per second.

Following recent work on dependency parsing (Nivre, 2007), we report two evaluation measures: labeled accuracy score (LAS) and unlabeled accuracy score (UAS). LAS is the percentage of tokens for which the parser predicts the correct head-word and dependency label. UAS ignores the dependency labels, and therefore corresponds to the percentage of words for which the correct head was found. In addition to LAS and UAS, we also report precision and recall of certain grammatical relations.

For example, compare the parser output of *go buy an apple* to the gold standard (Figure 1). This sequence of GRs has two labeled dependency errors and one unlabeled dependency error. `1|2|COORD` for the parser versus `1|2|SRL` is a labeled error because the dependency label produced by the parser (COORD) does not match the gold-standard annotation (SRL), although the unlabeled dependency is correct, since the headword assignment, `1|2`, is the same for both. On the other hand, `5|1|PUNCT` versus `5|2|PUNCT` is both a labeled dependency error and an unlabeled dependency error, since the headword assignment produced by the parser does not match the gold-standard.

4.2 Results

Trained on domain-specific data, the parser performed well on held-out data, even though the training corpus is relatively small (about 60,000 words). The results are listed in Table 1.

	LAS	UAS
Eve cross-validation	92.0	93.8

Table 1: Average cross-validation results, Eve

The labeled dependency error rate is about 8% and the unlabeled error rate is slightly over 6%. Performance in individual files ranged between the best labeled error rate of 6.2% and labeled error rate of 4.4% for the fifth file, and the worst error rates of 8.9% and 7.8% for labeled and unlabeled respectively in the fifteenth file. For comparison, Sagae et al. (2005) report 86.9% LAS on about 2,000 words of Eve data, using the Charniak (2000) parser with a separate dependency-labeling step. Part of the reason we obtain levels of accuracy higher than usually reported for dependency parsers is that the average sentence length in CHILDES transcripts is much lower than in, for example, newspaper text. The average sentence length for adult utterances in the Eve corpus is 6.1 tokens, and 4.3 tokens for child utterances[5].

Certain GRs are easily identifiable, such as DET, AUX, and INF. The parser has precision and recall of nearly 1.00 for those. For all GRs that occur more than 1,000 times in the Eve corpus (which contrains more than 60,000 tokens), precision and recall are above 0.90, with the exception of COORD, which

[5]This differs from the figures in section 2.3 because for the purpose of parser evaluation we ignore sentences composed only of a single word plus punctuation.

```
          go          buy         an        apple       .
parser:   1|2|COORD   2|0|ROOT    3|4|DET   4|2|OBJ     5|1|PUNCT
gold:     1|2|SRL     2|0|ROOT    3|4|DET   4|2|OBJ     5|2|PUNCT
```

Figure 1: Example output: parser vs. gold annotation

occurs 1,163 times in the gold-standard data. The parser's precision for COORD is 0.73, and recall is 0.84. Other interesting GRs include SUBJ, OBJ, JCT (adjunct), COM, LOC, COMP, XCOMP, CJCT (subordinate clause acting as an adjunct), and PTL (verb particle, easily confusable with prepositions and adverbs). Their precision and recall is shown in table 2.

GR	Precision	Recall	F-score
SUBJ	0.96	0.96	0.96
OBJ	0.93	0.94	0.93
JCT	0.91	0.90	0.90
COM	0.96	0.95	0.95
LOC	0.95	0.90	0.92
COMP	0.83	0.86	0.84
XCOMP	0.86	0.87	0.87
CJCT	0.61	0.59	0.60
PTL	0.97	0.96	0.96
COORD	0.73	0.84	0.78

Table 2: Precision, recall and f-score of selected GRs in the Eve corpus

We also tested the accuracy of the parser on child utterances and adult utterances separately. To do this, we split the gold standard files into child and adult utterances, producing gold standard files for both child and adult utterances. We then trained the parser on 14 of the 15 Eve files with both child and adult utterances, and parsed the individual child and adult files. Not surprisingly, the parser performed slightly better on the adult utterances due to their grammaticality and the fact that there was more adult training data than child training data. The results are listed in Table 3.

	LAS	UAS
Eve - Child	90.0	91.7
Eve - Adult	93.1	94.8

Table 3: Average child vs. adult results, Eve

Our final evaluation of the parser involved testing the parser on data taken from a different parts of the CHILDES database. First, the parser was trained on all gold-standard Eve files, and tested on manually annotated data taken from the MacWhinney transcripts. Although accuracy was lower for adult utterances (85.8% LAS) than on Eve data, the accuracy for child utterances was slightly higher (92.3% LAS), even though child utterances were longer on average (4.7 tokens) than in the Eve corpus.

Finally, because a few aspects of the many transcript sets in the CHILDES database may vary in ways not accounted for in the design of the parser or the annotation of the training data, we also report results on evaluation of the Eve-trained parser on a particularly challenging test set, the Seth corpus. Because the Seth corpus contains transcriptions of language phenomena not seen in the Eve corpus (see section 5), parser performance is expected to suffer. Although accuracy on adult utterances is high (92.2% LAS), accuracy on child utterances is very low (72.7% LAS). This is due to heavy use of a GR label that does not appear at all in the Eve corpus that was used to train the parser. This GR is used to represent relations involving *filler syllables*, which appear in nearly 45% of the child utterances in the Seth corpus. Accuracy on the sentences that do not contain filler syllables is at the same level as in the other corpora (91.1% LAS). Although we do not expect to encounter many sets of transcripts that are as problematic as this one in the CHILDES database, it is interesting to see what can be expected from the parser under unfavorable conditions.

The results of the parser on the MacWhinney and Seth test sets are summarized in table 4, where *Seth (clean)* refers to the Seth corpus without utterances that contain filler sylables.

5 Error Analysis

A major source for parser errors on the Eve corpus (112 out of 5181 errors) was telegraphic speech,

	LAS	UAS
MacWhinney - Child	92.3	94.8
MacWhinney - Adult	85.8	89.4
MacWhinney - Total	88.0	91.2
Seth - Child	72.7	82.0
Seth - Adult	92.2	94.4
Seth - Total	84.6	89.5
Seth (clean) - Child	91.1	92.7
Seth (clean) - Total	92.0	93.9

Table 4: Training on Eve, testing on MacWhinney and Seth

as in *Mommy telephone* or *Fraser tape+recorder floor*. Telegraphic speech may be the most challenging, since even for a human annotator, determining a GR is difficult. The parser usually labeled such utterances with the noun as the ROOT and the proper noun as the MOD, while the gold annotation is context-dependent as described above.

Another category of errors, with about 150 instances, is XCOMP errors. The majority of the errors in this category revolve around dropped words in the main clause, for example *want eat cookie*. Often, the parser labels such utterances with COMP GRs, because of the lack of *to*. Exclusive training on utterances of this type may resolve the issue. Many of the errors of this type occur with *want*: the parser could be conditioned to assign an XCOMP GR with *want* as the ROOT of an utterance.

COORD and PRED errors would both benefit from more data as well. The parser performs admirably on simple coordination and predicate constructions, but has troubles with less common constructions such as PRED GRs with *get*, e.g., *don't let your hands get dirty* (69 errors), and coordination of prepositional objects, as in *a birthday cake with Cathy and Becky* (154 errors).

The performance drop on the Seth corpus can be explained by a number of factors. First and foremost, Seth is widely considered in the literature to be the child who is most likely to invalidate any theory (Wilson and Peters, 1988). He exhibits false starts and filler syllables extensively, and his syntax violates many "universal" principles. This is reflected in the annotation scheme: the Seth corpus, following the annotation of Peters (1983), is abundant with *filler syllables*. Because there was no appropriate GR label for representing the syntactic relationships involving the filler syllables, we annotated those with a special GR (not used during parser training), which the parser is understandably not able to produce. Filler syllables usually occur near the start of the sentence, and once the parser failed to label them, it could not accurately label the remaining GRs. Other difficulties in the Seth corpus include the usage of *dates*, of which there were no instances in the Eve corpus. The parser had not been trained on the new DATE GR and subsequently failed to parse it.

6 Conclusion

We described an annotation scheme for representing syntactic information as grammatical relations in CHILDES data, a manually curated gold-standard corpus of 65,000 words annotated according to this GR scheme, and a parser that was trained on the annotated corpus and produces highly accurate grammatical relations for both child and adult utterances. These resources are now freely available to the research community, and we expect them to be instrumental in psycholinguistic investigations of language acquisition and child language.

Syntactic analysis of child language transcripts using a GR scheme of this kind has already been shown to be effective in a practical setting, namely in automatic measurement of syntactic development in children (Sagae et al., 2005). That work relied on a phrase-structure statistical parser (Charniak, 2000) trained on the Penn Treebank, and the output of that parser had to be converted into CHILDES grammatical relations. Despite the obvious disadvantage of using a parser trained on a completely different language genre, Sagae et al. (2005) demonstrated how current natural language processing techniques can be used effectively in child language work, achieving results that are close to those obtained by manual computation of syntactic development scores for child transcripts. Still, the use of tools not tailored for child language and extra effort necessary to make them work with community standards for child language transcription present a disincentive for child language researchers to incorporate automatic syntactic analysis into their work. We hope that the GR representation scheme and the parser presented here will make it possible and convenient for the child language community to take advantage of some of the recent developments in natural language parsing, as was the case with part-of-speech tagging when CHILDES specific tools were first made available.

Our immediate plans include continued improvement of the parser, which can be achieved at least in part by the creation of additional training data from other English CHILDES corpora. We also plan to release automatic syntactic analyses for the entire English portion of CHILDES.

Although we have so far focused exclusively on English CHILDES data, dependency schemes based on functional relationships exist for a number of languages (Buchholz and Marsi, 2006), and the general parsing techniques used in the present work have been shown to be effective in several of them (Nivre et al., 2006). As future work, we plan to adapt existing dependency-based annotation schemes and apply our current syntactic annotation and parsing framework to other languages in the CHILDES database.

Acknowledgments

We thank Marina Fedner for her help with annotation of the Eve corpus. This work was supported in part by the National Science Foundation under grant IIS-0414630.

References

A. Berger, S. A. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Roger Brown. 1973. *A first language: the early stages*. George Allen & Unwin Ltd., London.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, June. Association for Computational Linguistics.

John Carroll, Edward Briscoe, and Antonio Sanfilippo. 1998. Parser evaluation: a survey and a new proposal. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, pages 447–454, Granada, Spain.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 132–139, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

D. Knuth. 1965. On the translation of languages from left to right. *Information and Control*, 8(6):607–639.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, third edition.

Joakim Nivre, Johan Hall, Jens Nilsson, Gulsen Eryigit, and Svetoslav Marinov. 2006. Labeled pseudo-projective dependency parsing with support vector machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

Joakim Nivre, editor. 2007. *CoNLL-XI Shared Task on Multilingual Dependency Parsing*, Prague, June. Association for Computational Linguistics.

Ann M. Peters. 1983. *The Units of Language Acquisition*. Monographs in Applied Psycholinguistics. Cambridge University Press, New York.

Ann M. Peters. 1987. The role of immitation in the developing syntax of a blind child. *Text*, 7:289–311.

Kenji Sagae and Alon Lavie. 2006. A best-first probabilistic shift-reduce parser. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 691–698, Sydney, Australia, July. Association for Computational Linguistics.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with lr models and parser ensembles. In *Proceedings of the Eleventh Conference on Computational Natural Language Learning*.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2004. Adding syntactic annotations to transcripts of parent-child dialogs. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 197–204, Ann Arbor, Michigan, June. Association for Computational Linguistics.

B. Wilson and Ann M. Peters. 1988. What are you cookin' on a hot?: A three-year-old blind child's 'violation' of universal constraints on constituent movement. *Language*, 64:249–273.