Adding Syntactic Annotations to Transcripts of Parent-Child Dialogs

Kenji Sagae*, Brian MacWhinney† and Alon Lavie*

*Language Technologies Institute and †Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
sagae@cs.cmu.edu, macw@cmu.edu, alavie@cs.cmu.edu

Abstract

We describe an annotation scheme for syntactic information in the CHILDES database (MacWhinney, 2000), which contains several megabytes of transcribed dialogs between parents and children. The annotation scheme is based on grammatical relations (GRs) that are composed of bilexical dependencies (between a head and a dependent) labeled with the name of the relation involving the two words (such as subject, object and adjunct). We also discuss automatic annotation using our syntactic annotation scheme.

1 Introduction

Transcripts of dialogs between parents and children are the basic empirical data that support the bulk of work in child language acquisition and the study of developmental language disorders. The standard database of such transcripts is the CHILDES database (MacWhinney, 2000), which has been used in over 1500 studies. While many of these studies have relied on the transcribed words alone, some have benefited from automatic analyses of morphology and parts-of-speech (POS). The English corpora in CHILDES are now fully tagged for part of speech (POS). This newly completed tagging has opened the door for the automatic generation of syntactic analyses for these corpora. It is our goal to provide such syntactic information in the form of syntactic dependency structures labeled with grammatical relations (GRs) for the English portion of the CHILDES database. Although only a limited amount of transcript data will be annotated manually (10,000 words), tools will be available for automatic analysis of the entire English database of over 100 megabytes of transcripts.

The manually annotated corpus will include GR information for parent and child utterances. We will use it to develop a system that performs automatic syntactic analyses of English CHILDES data. In addition to its uses in language acquisition research, the information provided by this system would allow for the automation of clinical measures of child language development, such as DSS (Lee, 1974) and IPSyn (Scarborough, 1990), which are based on lexical and syntactic analyses of utterances.

In this paper, we discuss a GR syntactic annotation scheme developed specifically to address the needs of child language researchers and clinicians, the design considerations behind it, and how it relates to similar schemes proposed in recent years. We also describe the current state of development of tools that can parse transcribed spoken language into the GR representation.

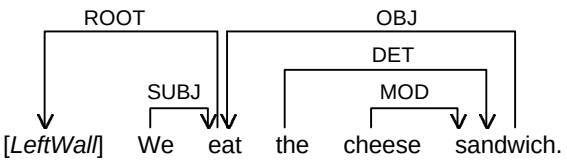
2 Syntactic Annotation in CHILDES

We represent syntactic information in CHILDES data in terms of labeled dependencies that correspond to grammatical relations, such as subjects, objects, and adjuncts. As in many flavors of dependency-based syntax, each GR in our scheme represents a relationship between two words in a sentence: a head and a dependent. Each word in a sentence must be a dependent of exactly one head word (but heads may have several dependents). The single exception to this rule is that every sentence has one "root" word that is not a dependent of any other word in the sentence. To restore consistency across the entire sentence, we make the root word a dependent of a special empty word, appended to the beginning of every sentence. We call this empty word the "LeftWall" (Sleator and Temperley, 1991). In addition to a head word and a dependent word, each GR is labeled with one out of 30 different GR types. Figure 1 shows the syntactic annotation of a sentence from the Eve corpus (Brown, 1973) of the CHILDES database.

The specific set of GR types included in our annotation scheme was designed based on a survey of the child language literature (Fletcher & MacWhinney, 1995) and a review of existing measures of syntactic development (MacWhinney, 2000). Using the detailed GR annotation scheme developed by Carroll et al. (2003) for parser evaluation as a starting point, we identified 30 grammatical relations of specific interest for the study of child language. A complete list of GR types used in our scheme is shown in figure 2.

Sentence:
We eat the cheese sandwich.

Grammatical Relations:



Annotation:

1	We	SUBJ	2
2	eat	ROOT	0
3	the	DET	5
4	cheese	MOD	5
5	sandwich	OBJ	2

Figure 1: Sentence and syntactic annotation

SUBJ, ESUBJ, CSUBJ, XSUBJ Subject, expletive subject, clausal subject (finite and non-finite)
OBJ, OBJ2, IOBJ Object, second object, indirect object
COMP, XCOMP Clausal complements, finite and non-finite
PRED, CPRED, XPRED Predicative, clausal predicative (finite and non-finite)
JCT, CJCT, XJCT Adjunct, clausal adjunct (finite and non-finite)
MOD, CMOD, XMOD Nominal modifier, clausal nominal modifier (finite and non-finite)
AUX, NEG Auxiliary and negation
DET, QUANT Determiner and quantifier
POBJ Prepositional object
PTL Verbal particle (of phrasal verbs)
CPZR Complementizer
COM Communicator
INF Infinitival particle
VOC Vocative
COORD Coordination
ROOT Special relation for the top node

Figure 2: GR types for CHILDES annotation

2.1 Recent Related Annotation Schemes

Our approach combines characteristics of other recent annotation schemes. Like Carroll et al (2003) we use a rich set of GRs that provide detailed information about syntactic relationships. Like Rambow et al (2002) we use dependency structures instead of constituent structures. Dependency structures provide increased ease and reliability of manual annotation. Moreover, it is far simpler to determine grammatical relations between words from their dependencies than from their constituent structures. In addition, much of the syntax-based child language work that uses CHILDES data is predominantly GR-driven.

In spite of the similarities mentioned above, our annotation scheme differs from those of Carroll et al. and Rambow et al. in important ways. The scheme of Carroll et al. does not annotate sentences with a full dependency structure, but rather just lists a set of GRs that occur in the sentence. By doing away with the requirement of a complete and consistent dependency structure, their scheme allows for n-ary relations (while our GRs are strictly binary) and greater flexibility in working with GRs that may not fit together globally. In our framework, n-ary relations (with $n > 2$) must be represented indirectly, using combinations of binary GRs. Their set of GRs, although detailed, is meant for general-purpose annotation of text, and does not include specific pieces of information we have identified as important to the child language community. In addition, they distinguish between "initial" (or deep) GRs, and actual (or surface) GRs, while we report surface GRs only.

The scheme of Rambow et al., on the other hand, is dependency-based. However, their dependency labels are limited to seven syntactic roles (SRole and DRole features, which can have the values of subj, obj, obj2, pobj, pobj2, adj and root). These seven roles suffice for some applications, but do not offer the granularity needed for CHILDES annotation. Both Rambow et al. And Carroll et al., annotate surface and deep relations, while we currently annotate only surface relations. Extending our scheme to represent deep syntactic relations is under consideration for future work.

2.2 Specific Representation Choices

GRs typically assign content words as heads and function words as dependents. For example, nouns are the heads of determiners (forming a GR of type DET with the noun as the head and the determiner as the dependent), and verbs are chosen as the heads of auxiliaries (forming a GR of type AUX with the verb as the head and the auxiliary as the dependent). When both words in a GR are members of lexical categories, the direction of the relation follows common practice with adjectives dependent on nouns, adverbs dependent on verbs, and nouns dependent on verbs. In the case of prepositional phrases, the preposition is chosen as the head of the prepositional object (in a POBJ relation). By convention, vocatives and communicators are dependents of the main verb of their sentences.

In cases where a clause has a relation to another clause, the verb of the lower clause is used as a dependent. Thus, in the relative clause of figure 3, the verb *saw* of the relative clause is treated as dependent in a CMOD (clausal modifier of a nominal) relation with the noun *boy*. The other relations in this sentence are as expected: *The* and *a* are dependents in DET relations with *boy* and *picture*, *boy* is the dependent in a SUBJ relation with *saw*, and *picture* is in a OBJ relation with *drew*. Finally, *drew* is the root of the sentence, and by convention is the dependent in a ROOT relation with the special word *LeftWall*, which we append to the beginning of the sentence.

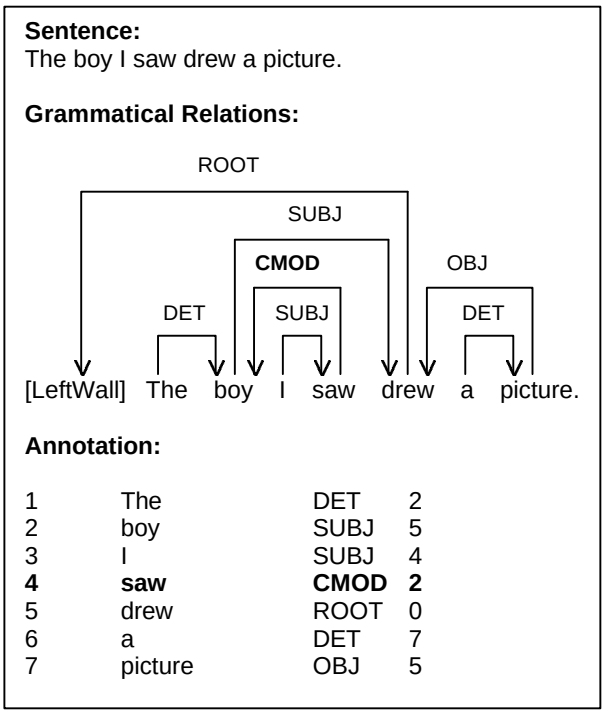


Figure 3: Annotation of a clausal modifier of a noun

A point worth noting is that, in general, only words that appear in a sentence can be participants in a GR as either a head or a dependent. This is in contrast to the scheme of Rambow et al., where an empty word *e* can be added to the sentence in control structures. For example, in the sentence *I wanted to run*, Rambow et al. annotate the empty word *e* as the subject of *run*, while we do not annotate a subject for *run*. The exception for this general rule is in the case of ellipsis, where we do insert elided material back into the sentence (as do Rambow et al. in the case of VP-ellipsis).

2.3 Inter-Annotator Agreement

Inter-annotator agreement was measured by having a corpus of 285 words (48 sentences from a CHILDES corpus) annotated manually by two annotators, independently. The annotators had at least a basic background in Linguistics, and one was familiar with the annotation scheme. The other was trained for about one hour before annotating a separate 10-sentence trial corpus under close guidance. Annotation of the 285-word corpus took about 90 minutes. Annotator agreement was 96.5%, which is about the same as inter-annotator agreement figures for related annotation schemes involving dependencies and grammatical relations. Carroll et al. report 95% agreement, while Rambow et al., report 94% agreement. Because of the size of the corpora used for rating annotator agreement, these differences are not significant. Out of 285 possible labeled dependencies, there were 10 disagreements between the annotators. Of particular interest, four of them were disagreements on the attachments of adjuncts, and three of them were incorrect labeling (with correct dependent-head links) involving COMP, PRED and OBJ.

3 Automatic Syntactic Annotation

While a limited amount of manually annotated data might be of great value, child language researchers may need to examine the occurrence of syntactic patterns over several megabytes of text. In addition, diagnosis of language development in children requires analysis of utterances of the specific child being assessed. Towards these ends, we are developing tools that automatically annotate text with the scheme described above.

3.1 Identifying GRs with a Rule-Based Parser

Our initial approach for extracting grammatical relations from CHILDES transcripts relied on a rule-based robust parser (Rosé and Lavie, 2001), and a small manually written grammar (153 rules) loosely based on Lexical Functional Grammar designed specifically for the task. The parser's robustness comes from features designed for analysis of spoken language. More precisely, the parser has the ability to skip words, or insert parse tree nodes in an analysis, when necessary to make a sentence conform to the given grammar. The parser outputs f-structures, from which GRs are extracted (GRs can be read almost directly from the functions present in the f-structures). In an evaluation limited to subjects (SUBJ, CSUBJ and XSUBJ), objects (OBJ, OBJ2), adjuncts (JCT), and predicate nominals (PRED, CPRED and XPRED), the rule-based system achieved high precision (about 0.85), but low recall. The lack of recall is due to sentences that receive no analysis from the parser because they are not covered by the grammar. Although the parser's robustness features does allow for analysis of sentences that deviate from the grammar in specific ways, recall was still only 0.64. A detailed description of the rule-based system can be found in (Sagae et al, 2001) and (Sagae et al, in press).

3.2 Combining Rule-Based and Data-Driven Approaches

To remedy the low recall of the rule-based system, we designed a simple data-driven GR identifier that always outputs an analysis for each sentence. An interesting feature of the data-driven system is that it uses no new manually annotated data[1], and is trained solely on (unverified) output of the rule-based system.

The data-driven system is comprised of freely available natural language processing tools, and works as follows. Sentences are first parsed with the Charniak parser, which outputs a constituent tree structure. A customized version of the commonly used head-percolation table for Penn Treebank trees (Magerman, 1995; Collins, 1996) is then used to extract head-dependent unlabelled links. A part-of-speech (POS) tagger is trained to assign GR labels to each word. The label assigned is that of the GR in which the word is a dependent (recall that in every sentence each word is a dependent in exactly one GR). Training is done on pairs of POS tags and GR labels, where the POS tags are given

[1] The system described here uses an existing statistical parser (Charniak, 2000) pre-trained on the Penn Treebank (Marcus et al., 1993) to determine unlabelled GR links. However, preliminary results indicate that similar results can be achieved by training a statistical dependency parser on output of the rule-based parser alone.

as input at run-time, and the tagger must produce a GR label. Some lexicalization of this procedure is achieved by using error-driven transformation based learning, with the fn-TBL toolkit (Ngai and Florian, 2001), where the word forms, POS tags, and GR labels produced by the tagger are used as features. This system was intended only as a proof-of-concept for the combination of rule-based and data-driven systems, and a more suitable data-driven dependency parser is currently under development.

The output of the data-driven system is used only when the rule-based system fails to produce an analysis. In this scheme, precision and recall of GRs were both just above 0.80. A detailed explanation and evaluation of results obtained with the combined system can be found in (Sagae and Lavie, 2003).

4 Conclusions and Future Work

Our annotation scheme for representing syntactic grammatical relations in transcripts of spoken dialogues is specifically designed to address the needs of the child language acquisition community. We plan to manually annotate about 10,000 words of English CHILDES data with the described annotation scheme, and then use the manually annotated data to train high accuracy parsers that will automatically annotate over 100 megabytes of English CHILDES data. The manually annotated data we will produce should prove to be of great value for applications beyond child language research, including design, training and evaluation of natural language applications. Although we focus on CHILDES transcripts, much of our approach to syntactic parsing and the general framework of our annotation scheme are applicable to spoken language in general. Developments in syntactic analysis of spoken language, which has received little attention compared to the analysis of written text, are necessary for the improvement of several natural language processing applications, such as speech-to-speech machine translation and spoken interfaces.

The next steps in this project involve the manual annotation of larger representative samples from CHILDES data. These samples will be used in further development and evaluation of syntactic parsers that are capable of producing the grammatical relations in our scheme as output.

Topics under current investigation include the impact of robustness (word skipping, tree-node insertion) on the precision and recall of GRs obtained with rule-based parsing, data-driven models of labeled dependency parsing, and ways of combining multiple sources of information in GR extraction (such rule-based and statistical parsers, and even other end-to-end GR extraction systems).

References

Brown, R. (1973). A First Language. Cambridge, MA: Harvard University Press.

Carroll, J., Minnen, G., and Briscoe, E. (2003). Parser evaluation: using a grammatical relation annotation scheme. In A. Abeillé (ed.), Treebanks: Building and Using Parsed Corpora, Dordrecht: Kluwer.

Charniak, E. (2000). A maximum-entropy-inspired parser. In Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics. Seattle, WA.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. In Proceedings of the 34th Meeting of the Association for Computational Linguistics (pp. 184-191). Santa Cruz: ACL.

Fletcher, P., & MacWhinney, B. (Eds.). (1995). The Handbook of Child Language. Oxford: Blackwell.

Lee, L. (1974). Developmental Sentence Analysis. Evanston, IL: Northwestern University Press.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. Mahwah, NJ: Lawrence Erlbaum Associates.

Magerman, D. (1995). Statistical decision-tree models for parsing. In Proceedings of the Association for Computational Linguistics.

Ngai, G., & Florian, R. (2001). Transformation-based learning in the fast lane. In Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (pp. 40-47). Pittsburgh, PA: ACM.

Rambow, O., Creswell, C., Szekely, R., Taber, H., and Walker, M. (2002). A Dependency Treebank for English. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002), Las Palmas, Spain.

Rosé, C. P., & Lavie, A. (2001). Balancing robustness and efficiency in unification-augmented context-free parsers for large practical applications. In A. van Noord & A. Junqua (Eds.), Robustness in language and speech technology. Amsterdam: Kluwer.

Sagae, K., Lavie, A., & MacWhinney, B. (2001). Parsing the CHILDES database: Methodology and lessons learned. In Proceedings of the Seventh International Workshop in Parsing Technologies. Beijing, China.

Sagae, K., Lavie, A. (2003). Combining Rule-based and Data-driven Techniques for Grammatical Relation Extraction in Spoken Language. In Proceedings of the Eighth International Workshop in Parsing Technologies. Nancy, France.

Sagae, K., MacWhinney, B., & Lavie, A. (in press). Automatic parsing of parent-child interactions. In Behavior Research Methods, Instruments, and Computers.

Sleator, D., & Temperley, D. (1991). Parsing English with a Link Grammar. In Proceedings of the Third International Workshop in Parsing Technologies. Tilburg, Netherlands.

Scarborough, H. (1990). Index of productive syntax. In Applied Psycholinguistics, 11, 1-22.