

***Parsing of Grammatical Relations
in Transcripts of Parent-Child Dialogs***

Thesis Summary

Kenji Sagae

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Alon Lavie, co-chair
Brian MacWhinney, co-chair
Jaime Carbonell
Lori Levin
John Carroll, University of Sussex

Abstract

Automatic analysis of syntax is one of the core problems in natural language processing. Despite significant advances in syntactic parsing of written text, the application of these techniques to spontaneous spoken language has received more limited attention. The recent explosive growth of online, accessible corpora of spoken language interactions opens up new opportunities for the development of high accuracy parsing approaches to the analysis of spoken language. The availability of high accuracy parsers will in turn provide a platform for development of a wide range of new applications, as well as for advanced research on the nature of conversational interactions. One concrete field of investigation that is ripe for the application of such parsing tools is the study of child language acquisition.

In this thesis, we describe an approach for analyzing the syntactic structure of spontaneous conversational language in parent-child interactions. Specific emphasis is placed on the challenge of accurately annotating the English corpora in the CHILDES database with grammatical relations (such as subject, objects and adjuncts) that are of particular interest and utility to researchers in child language acquisition. This work involves rule-based and corpus-based natural language processing techniques, as well as a methodology for combining results from different parsing approaches. We present novel strategies for integrating the results of different parsers into a system with improved accuracy.

One practical application of this research is the automation of language competence measures used by clinicians and researchers of child language development. We present an implementation of an automatic version of one such measurement scheme. This provides not only a useful tool for the child language research community, but also a task-based evaluation framework for grammatical relation identification.

Through experiments using data from the Penn Treebank, we show that several of the techniques and ideas presented in this thesis are applicable not just to analysis of parent-child dialogs, but to parsing in general.

1 Introduction

Automatic analysis of syntax is one of the core problems in natural language processing. The process of determining one or more syntactic structures corresponding to an input sentence is commonly referred to as *syntactic parsing*, or simply *parsing*, and it is generally regarded as a necessary step in natural language understanding, as well as in several natural language applications such as machine translation or question answering. After decades of research in linguistics, computer science and related fields, reasonably accurate natural language parsers are now available for a number of languages. However, most of this research has focused on written text, and the application of these analysis methods to spontaneous spoken language corpora has received more limited attention. At the same time, the recent explosive growth of online, accessible corpora of spoken language interactions opens up new opportunities for the development of high accuracy parsing approaches to the analysis of spoken language. The availability of high accuracy parsers will in turn provide a platform for developing a wide range of new applications, as well as for advanced research on the nature of conversational interactions. One concrete field of investigation that is ripe for the application of such parsing tools is the study of child language acquisition. The CHILDES database (MacWhinney, 2000), containing several megabytes of 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. In this thesis, we will present an approach for syntactic parsing of the child and adult language in CHILDES transcripts, focusing on a syntactic representation that suits the needs of the child language research community.

Despite the enormous impact CHILDES has had on the study of language acquisition, the lack of tools for automatic analysis above the word-level stands as a barrier to the realization of the full potential of this system for research on normal language development and the evaluation of language disorders. Currently, the corpus includes no syntactic information, because parsing technology specifically targeted to the production of such annotations has not been available. Lacking these tools, several major research groups¹ have been forced to engage in resource-intensive syntactic annotation of small segments of the database. Because this work is not based on shared community standards, the resulting analyses are discrepant and non-replicable, and current corpus-based research on syntactic development has been only weakly cumulative.

The work presented in this thesis aims to correct this situation through the development of a new system specifically tailored to the needs of the child language community. Developing technology for reliable analysis of the syntax of spontaneous spoken language is a challenging task, and shaping this technology for the productions of young children is a still greater challenge. We present an annotation scheme designed to represent syntactic information in CHILDES corpora in a way that addresses specific needs of the child language community. We then investigate the use of different parsing technologies to analyze transcripts according to our annotation scheme. Finally, we will show that novel ways of combining different parsing

¹ Such as those in the University of Wisconsin (Madison), the University of California (San Diego and Berkeley), Rutgers, Purdue, Cornell, the Massachusetts Institute of Technology, Nagoya University, the University of Potsdam, among others.

approaches allow for accurate analysis of syntax in both child and adult language in CHILDES transcripts.

1.1 Motivation

Explaining the enigma of child language acquisition is one of the main challenges facing cognitive science. Although all normal children succeed in learning their native tongue, neither psychology nor linguistics has yet succeeded in accounting for the many complexities of language learning. Within this general area, there has been particular attention to the acquisition of grammar, as it is expressed through morphosyntax, stimulated in large measure by Chomsky's theory of Universal Grammar and its attendant claims regarding innate principles and parameters (Chomsky, 1982). Beyond its theoretical importance, the measurement of morphosyntactic competence is crucial to applied work in fields such as developmental language disorders, schooling and literacy, and second language acquisition.

To examine the development of morphosyntax, researchers have come to rely increasingly on large corpora of transcript data of verbal interactions between children and parents. The standard database in this area is the CHILDES database (MacWhinney, 2000; <http://childes.psy.cmu.edu>), which provides 300MB of transcript data for over 25 human languages, as well as a large amount of digitized audio and video linked to the transcripts. There are now several hundred studies that have used the CHILDES database to study the development of morphosyntax. However, these studies have typically been forced to use the database in its raw lexical form, without tags for part-of-speech, syntactic parses, or predicate-argument information. Lacking this information, researchers have devoted long hours of hand analysis to locating and coding the sentences relevant to their hypotheses. If syntactic parses were available, these analyses could be automated, allowing investigators to conduct a wider variety of tests in a more reliable fashion. Automatic syntactic analysis systems would also be of great value in clinical settings, allowing clinicians and clinical researchers to construct profiles of language delays by comparing small speech samples collected in structured interviews with a larger database of normed data.

Producing automated analyses of child language corpora is one instance of the more general challenge of developing a comprehensive NLP approach to the analysis of spontaneous speech. Although many of the ideas presented in this thesis are applicable to other genres of spoken language, we focus on the child language problem for three reasons. First, a large, publicly accessible corpus of parent-child spoken language interactions is available, and its importance in the study of child language makes the task of syntactic analysis immediately meaningful in a practical sense. Second, research in this area works towards bridging the gap between cognitive science and natural language processing, thereby improving interaction between these fields. Third, the child language field has already developed a clear set of criteria for the standardized measurement of morphosyntactic development (Scarborough, 1990). The automation of these pre-existing measures provides clear benchmarks for new NLP systems.

1.2 Research Goals

The research I propose encompasses five main objectives:

- Developing effective parsing approaches for high accuracy identification of grammatical relations in spoken language (including utterances spoken by adults and children at various stages of first language acquisition);
- Developing effective methods for combining the results of multiple sources of information, including rule-based and statistical parsers, in order to boost the accuracy of the combined system by accounting for specific strengths of the different approaches;
- Evaluating the performance of our resulting grammatical relation parsing approaches using the standard performance metrics of precision and recall;
- Validating the utility of the resulting systems to child language acquisition research by automating a standard application for measuring the grammatical complexity of child language corpora;

2 Background: Data and Previous Related Work

Natural language syntactic parsers are systems that output one or more syntactic structures corresponding to an input sentence. Parsing is a complex problem, and several different approaches have demonstrated some amount of success in different aspects of the task. Parsers differ with respect to what underlying algorithm they use to determine syntactic structures, what kind of syntactic structures they provide, whether they output one or more syntactic structures for a single sentence, among other things. One characteristic that varies significantly across several parsing approaches is how much the system relies on linguistic knowledge encoded as rules, such as with a manually-written grammar, and how much the system learns based on manually annotated examples (sentences with their correct syntactic structures). In chapter 2 we start by looking at this source of variation, as one of the issues addressed in this thesis is the combination of rule-based and data-driven parsing strategies: by leveraging on the strengths of approaches that tackle the parsing challenge in different ways, we attempt to achieve overall results that surpass the performance of the state-of-the-art of the different approaches taken in isolation. We then turn to the data used in this work, and discuss what data we target for syntactic analysis, what data is used for evaluation of the techniques we develop, and what data is used for development and training of our systems. Finally, we discuss previous work in parsing and related fields that are of particular relevance to the work presented in this thesis.

2.1 The CHILDES Database

The standard source for corpus-based research in child language is the CHILDES Database (MacWhinney, 2000), which contains hundreds of megabytes of transcripts of dialogs between children and parents in several languages. CHILDES transcripts have been used in over 1,500 studies in child language acquisition and developmental language disorders. We focus on the English portion of the database, and pay particular attention to the Eve corpus (Brown, 1973), which we partition into training, development and testing sections that we use with the different parsing strategies we pursue.

2.2 The Penn Treebank

The Penn Treebank (Marcus, Santorini, & Marcinkiewics, 1993) contains a large amount of text annotated with constituent (phrase structure) trees. The Wall Street Journal (WSJ) corpus of the

Penn Treebank contains about one million words of text from news stories in the financial domain. Although the type of language in the WSJ domain differs significantly from our target domain (CHILDES transcripts with utterances from children and parents), there are two important reasons for our use of the WSJ corpus of the Penn Treebank in this work: (1) it contains a much larger amount of text annotated with syntactic structures than would be feasible us to annotate manually using CHILDES data, and certain data-driven systems require very large training corpora; (2) it allows us to compare our methods directly to a vast and well-known body of research in parsing that has used the WSJ corpus for training, development and testing.

3 The CHILDES GR Annotation Scheme

One crucial aspect of producing useful automatic syntactic analysis for child-parent dialog transcripts is the definition of a suitable annotation scheme that targets the specific type of syntactic information needed by the child language community. To address this need, we have developed the CHILDES Grammatical Relation (GR) annotation scheme, described chapter 3.

Syntactic information in CHILDES transcripts is represented according to our scheme as labeled dependency structures. Such structures are trees where each node is a word in the sentence, and there are labels associated with each edge that reflect the type of syntactic relationship (or grammatical relation) that exists between the parent word (the head), and its children (the dependents). The GRs specified by the dependency labels include relations such as subject, object, adjunct, etc. The specific set of GR types (or labels) included in our annotation scheme was developed by using the GRs in the annotation scheme of Carroll et al. (2003) as a starting point, and adapting the set to suit specific needs of child language research. Sources of refinement included a survey of the child language literature (Fletcher & MacWhinney, 1995), a review of existing measures of syntactic development (MacWhinney, 2000), and input from child language researchers.

In general, content words (such as nouns and verbs) are considered to be the heads of function words (such as determiners and auxiliaries), except in the cases of prepositional phrases, where the preposition is the head of its object (usually a noun), and coordination, or a conjunction is the head of its coordinated phrases. Only surface grammatical relations are considered, and the annotation scheme does not deal with issues such as control and ellipsis (although extensions for handling such phenomena are possible). We have measured inter-annotator agreement for the CHILDES GR annotation scheme at 96.5%.

4 A Rule-Based GR Parsing Approach

Chapter 4 describes the use of rule-based parsing for automatically annotating CHILDES corpora with the grammatical relation scheme described in chapter 3. The form of rule-based parser we consider here is more specifically described as grammar-driven parsing, which uses a set of production rules (a grammar) that specify how each syntactic constituent may be expanded into other constituents or words as a model of natural language.

The use of a grammar effectively shapes the search space of possible syntactic analyses for a sentence, since the rules specify exactly how words may combine to form larger structures. It is then important to use a carefully designed grammar that allows the system to analyze as many syntactic configurations in the text as possible, while constraining the space of analyses so that

ambiguities present in the grammar can be resolved effectively, and a correct analysis for each sentence can be found. In practice, this results in a trade-off between coverage (the portion of input sentences that are covered by the rules in the grammar) and ambiguity (how many possible analysis for each sentence are licensed by the grammar). Grammars that cover too many ways words and phrases may be combined can be highly ambiguous, making the process of choosing a correct analysis for a given sentence (in other words, disambiguation) a difficult task. On the other hand, more restrictive grammars may not cover all syntactic structures in the input text. In spontaneous spoken language this problem is exacerbated, since sentence structures do not always conform to the more rigid standards of written text.

We developed a syntactic analysis system based mainly on grammar-driven robust parsing using the LCFlex parser (Rosé & Lavie, 2001). The system analyzes utterances from CHILDES transcripts in three steps: (1) word-level analysis (morphology and part-of-speech tagging); (2) grammar-driven syntactic analysis; and (3) statistical disambiguation.

4.1 Word-level Analysis: MOR and Part-of-Speech Tagging

We use the morphological analyzer MOR (MacWhinney, 2000), which was designed for use with CHILDES data, and returns a set of possible morphological analyses for each word in a input sentence. For part-of-speech tagging and disambiguation of the output of MOR, we designed a scheme that involves MXPOST (Ratnaparkhi, 1996), a part-of-speech tagger trained on the Penn Treebank WSJ corpus, and a classifier that converts the output of MXPOST to the CHILDES POS tag set. With this approach, we obtain over 96% accuracy of part-of-speech tagging using the CHILDES tag set.

4.2 Grammar-driven Syntactic Analysis: LCFlex

LCFlex is a robust grammar-driven parser with features that are designed specifically for analysis of spontaneous spoken language. Through the tuning of a few parameters, the parser allows the insertion of pre-specified words or constituents into the input sentence, or the skipping of certain portions of the input. The ability to insert and skip allows the parser to analyze sentences that deviate from the language specified by the grammar in use. This results in a way to manage aspects of the coverage/ambiguity trade-off while keeping the grammar constant.

Grammars used by LCFlex consist of a set of rules, each containing a context-free backbone and a set of unification equations. A bottom-up chart-parsing algorithm with top-down left-corner predictions is used to process sentences according to the context-free backbone of the grammar, and unification according to the equations in each rule is performed in interleaved fashion. LCFlex outputs feature structures that can be easily converted to labeled dependency structures. To parse CHILDES data with LCFlex, we designed a custom grammar for our domain, containing unification features that correspond to the GRs in the CHILDES annotation scheme. The grammar is relatively small, and it covers about 65% of the Eve test set. While 35% of the sentences in our test set cannot be parsed with this grammar, ambiguity is kept low. This means that while the recall of specific GRs is low, precision is high. By allowing the parser to insert and skip according to its robustness parameters, we can parse more sentences, while introducing more ambiguity. In other words, we increase recall at the expense of precision.

4.3 Statistical Disambiguation: Two-level PCFG

When the grammar allows for multiple analyses of a single sentence, we choose one of the possible analyses using statistical disambiguation. In our system, disambiguation is done with a two-level probabilistic context-free grammar (two-level PCFG).

PCFGs are generative models that simply have a probability associated with each context-free rule in the grammar (and the probabilities for all rules with the same left-hand side sum to 1.0). According to a PCFG, the probability of an analysis is the product of the probabilities of each rule used in the derivation of the analysis. A two-level PCFG considers rule bigrams (while a regular PCFG considers rule unigrams). In other words, instead of each rule having a probability (as in the case of PCFGs), each pair of rules has a probability. Similar to the parent annotation described by Johnson (1998), this model creates bounded sensitivity to context. The rule-bigram probabilities in our two-level PCFG are determined by maximum likelihood estimation using the (manually annotated) Eve training corpus.

4.4 Tailoring an Analysis System for Adult Utterances

While the adult language in the CHILDES corpora generally conforms to standard spoken language, the child language in the corpora varies from the language of a child in the very early stages of language learning to fairly complex syntactic constructions. Child and adult utterances differ significantly enough that we can analyze them more accurately by doing so separately, often with different strategies. We first explore the “easier” (in the sense that it is better defined) problem of analyzing adult utterances, which are theoretically of equal importance as child utterances, since theories of learning depend heavily on consideration of the range of constructions provided to children in the input (MacWhinney, 1999; Moerk, 1983).

Although the grammar covers about 65% of the sentences in the Eve test set, this does not mean that the system returns the correct analysis for each of these sentences. What it does mean, however, is that the correct analysis is included in the set of analyses the parser finds for a sentence, and it is up to the disambiguation procedure to find it. Allowing the parser to use its robustness features (skip portions of the input and insert lexical and non-lexical items in the parsing process, as needed) increases coverage, while also increasing ambiguity (and making the search for the correct analysis more difficult). In other words, by controlling the amount of flexibility that the parser is allowed to use, we can manipulate the trade-off between precision and recall of GRs. For example, consider the GR SUBJ (subject). If we look at the output of system when it is not allowed to insert or skip at all, we obtain high precision (94.9%) and low recall (51.6%). This is because ambiguity and coverage were limited. If we increase coverage and ambiguity by allowing the system to insert a noun phrase, and to skip one word, precision drops due to the increased ambiguity (84.3%), but recall improves due to the increase in coverage (76.2%).

Although the performance of our rule-based system is poor in certain aspects, it is not meant to be used by itself for analysis of CHILDES data. Instead, it is used in combination with other approaches described in this thesis. As we show in chapter 6, when combined with data-driven approaches, the rule-based system can be valuable.

4.5 Grammar-driven Parsing of Child Utterances

While every adult utterance in CHILDES transcripts is expected to have valid grammatical relations, utterances spoken by children younger than five years old often deviates enough from standard adult speech that the recognition of GRs is not possible. Grammar-driven parsing is particularly appropriate in this situation, since it can be used to detect sentences where the system should not attempt to find GRs. Because the well-formed sentences spoken by young children are syntactically simpler than adult speech, our grammar covers those sentences quite well. Sentences that are rejected by the rule-based system are considered to be in the category of sentences where GR analysis is not possible. The rule-based system finds such sentences with accuracy above 85%. In chapter 6, we will see how this discrimination capability of our rule-based system can be combined with other parsing approaches for accurate GR analysis of child utterances.

5 Data-Driven GR Parsing

In chapter 5 we consider two data-driven approaches to syntactic analysis of CHILDES data. The first is based on existing statistical parsing using a generative model trained on the WSJ corpus of the Penn Treebank. The second is based on deterministic classifier-based parsing, where a classifier decides the actions of a shift-reduce parser. To compare this last approach to other work on parsing English, we first train and test it using the standard division of the WSJ corpus. We then apply it to our task of parsing CHILDES data, training and evaluating it on the appropriate sections of the Eve corpus.

5.1 GR Parsing with a Generative Statistical Parser

This first data-driven GR parsing approach we describe illustrates how existing natural language processing tools and resources can be utilized to produce a high-performance system for GR analysis in CHILDES data with relatively little cost. This approach is particularly useful in languages for which accurate parsers already exist. The main idea is obtain syntactic parses for the target data using an existing parser, the well-known Charniak (2000) parser, and convert those analyses into GRs in the CHILDES annotation scheme. The GR system we developed according to this general idea analyzes CHILDES transcripts in three steps: text preprocessing, unlabeled dependency identification, and dependency labeling.

In the text processing step we simply use existing tools for processing transcripts in the CHILDES database to remove dysfluencies such as false-starts, retracings and repetitions. This step also tokenizes the input text, providing text that is as clean as possible to the Charniak parser.

The second step of unlabeled dependency identification is where we actually find a bare-bones syntactic analysis for the input text, using the Charniak parser. Since the parser is trained on Penn Treebank constituent trees, and outputs the same style of syntactic representation, it is then necessary to convert the output of the parser into the dependency format used for representing syntactic structures in the CHILDES annotation scheme. We use the standard procedure of lexicalizing the output constituent trees using a head percolation table (Collins, 1996; Magerman, 1995) to extract an unlabeled dependency structure from the output of the Charniak parser. The rules in the head table we use are similar to the ones used in the Collins (1996) parser, but modified to reflect our dependency annotation scheme.

Once we have obtained unlabeled dependencies, we proceed to the third step: dependency labeling. This is accomplished by using a classifier that determines a GR label for each of the dependencies in a dependency structure. The input we pass to the classifier is a set of features extracted from the unlabeled dependency structure, including the head and dependent words in the dependency in question, their part-of-speech tags, the distance between the two words, and the label (phrase type) of the lowest constituent that includes both words in the original constituent tree output of the Charniak parser. The output of the classifier is one of the possible GR labels in the CHILDES GR annotation scheme. The classifier is trained by extracting pairs of correct labels and feature sets for every dependency in the Eve training corpus. The classifier used in our system is the k-nearest neighbors implementation in the TiMBL package (Daelemans, Zavrel, van der Sloot, & van den Bosch, 2004).

This approach is quite effective in identifying GRs in CHILDES data, as we verify by testing it on the Eve test corpus. When tested on WSJ data, the Charniak parser has an unlabeled dependency accuracy of over 92%. On the Eve test corpus, the unlabeled dependency accuracy of the parser is 90.1%. Despite the degradation in performance due to the differences in the training and testing domains, the accuracy of the parser on CHILDES data is still high. The accuracy of the dependency labeling step using the k-nn classifier (on gold standard unlabeled dependencies) is 91.4%. When we combine the unlabeled dependency parsing step and the dependency labeling step, the overall labeled dependency accuracy of the system is 86.9%. Although accuracy is high, the precision and recall of certain GRs is, in fact, quite low. The GRs corresponding to clausal complements COMP and XCOMP, for example, have precision and recall below 60%. Other frequent GRs, such as subject, objects and adjuncts, are recognized with high levels of precision and recall (above 80%).

5.2 GR Analysis with Classifier-Based Parsing

Although the statistical parsing approach to GR identification described in the previous section performs well for most GRs in our annotation scheme, there are important reasons for also using additional data-driven approaches to GR analysis. One of the main reasons, which is discussed in detail in chapter 6 and is directly related to one of the central themes in this thesis, is that we can improve the precision and recall of identification of individual GRs as well as overall system accuracy by utilizing several different GR identification approaches. Henderson and Brill (1999) have shown that, in the context of constituent parsing of the Penn Treebank, combining the outputs of different parsers can result in improved accuracy, even if all each parser uses the same training data. As we show in chapter 6, this aspect of parser diversity is also applicable to dependency parsing, and combination schemes using different approaches to GR parsing result in improved precision and recall of GRs. An additional reason for pursuing different analysis approaches is, of course, that our statistical parsing approach discussed in the previous section does not identify certain GRs (such as COMP and XCOMP) reliably. Finally, this second data-driven approach we consider allows us to develop a parser that works natively with the syntactic representation used in our CHILDES GR scheme.

We first present a general approach for classifier-based parsing with constituent structures, and evaluate two parsers developed in this framework using the standard split for training, development and testing of the Penn Treebank. The main idea our classifier-based parsing approach is to have a classifier that decides on the parsing action of a shift-reduce parser. The

parsing algorithm is very simple, and parsing is done in linear time. Each time the parser must decide on a shift or a reduce action, a set of features that reflect the current configuration of the parser is given as input to a classifier, which then outputs a parser action. A parser using k -nearest neighbors for classification achieves slightly above 80% precision and recall of constituents, and 86.3% dependency accuracy on Penn Treebank WSJ data. Using support vector machines, the classifier-based parser achieves over 87% precision and recall of constituents, and 90.3% dependency accuracy. The SVM-based parser, although not as accurate as state-of-the-art statistical parsers that achieve about 89% precision and recall on the same data, is more accurate than several more complex parsers, while parsing considerably faster than popular statistical parsing approaches.

We then turn to the task of parsing CHILDES data using a classifier-based parser. First, we show that our general classifier-based parser for constituents can be easily adapted into a labeled dependency parser. The resulting parser is similar to the MALT parser (Nivre & Scholz, 2004), but it uses a slightly different set of features. Using support vector machines for classification, and training on the Eve training set, the performance of the parser on the Eve test set is surprisingly high. Overall labeled dependency accuracy is 87.3%, putting this approach on the same level as the one using the more complex Charniak parser (trained on the larger WSJ corpus). More interestingly, the precision and recall of GRs that were problematic for our previous approach is much improved. For example, the precision and recall of XCOMP are above 80%, and the precision and recall of COMP are above 70%.

6 Combining Different Approaches for GR Parsing

In chapter 6 we address the issues related to the combination of the several parsing approaches discussed in chapters 4 and 5 to achieve improved identification of GRs. We start with a simple voting scheme, examine progressively more sophisticated combination schemes, and arrive at a novel way of performing parser combination using a parsing algorithm. We then turn to the issue of analyzing utterances from young children using rule-based parser to discriminate between utterances that contain GRs and those where GRs cannot be reliably identified.

6.1 Dependency Voting

First we consider the simplest case of combining unlabeled dependency structures created by different parsers. The simplest voting scheme assigns equal weight to the dependencies generated by every parser. Voting is done on a word-by-word basis. For each word, we check the head chosen by each of the parsers. Each of these heads receives one vote. The head with most votes is chosen as the head for the current word.

We tested this combination scheme using different parsers generated according to the classifier-based framework described in chapter 5. Accuracy of the combination on the standard WSJ test set is 91.9%, a 14% error reduction over the most accurate parser in the combination. Similar but more complex schemes that assign weights to the votes of each parser result in further accuracy improvements, with the best scheme achieving 92.3% accuracy. For comparison purposes, the statistical parsers of Collins (1997) and Charniak (2000) have unlabeled dependency accuracy of 91.2% and 92.3%, respectively. When these two statistical parsers are included in the best-performing voting scheme in addition to the deterministic parsers, we achieve 93.9% accuracy, surpassing the highest published results in the WSJ test set. When applied to the

Eve test set, a weighted voting combination gives us 94.0% unlabeled dependency accuracy (over 30% relative error reduction from the single best parser in the combination scheme), and 91.6% labeled dependency accuracy.

6.2 Obtaining Well-Formed Dependency Trees

Although the voting schemes perform well in producing more accurate dependencies from multiple parsers, there is no guarantee that the resulting dependencies form a dependency tree. In fact, the resulting set of dependencies may form a structure with cycles, or even disconnected graphs. We present two novel ways to combine dependency structures that perform as well as the voting schemes in terms of accuracy, while building well-formed dependency structures.

In both cases, we start by creating a graph where each word in the sentence is a node. We then create directed edges between nodes corresponding to words for which dependencies are obtained from any of the parsers. In cases where more than one parser indicates that the same edge should be created, the weights are added, just as in the voting scheme. As long as any of the parsers creates a valid dependency tree for the sentence, the directed weighted graph created this way will be fully connected.

Once the graph is created, we can simply find its maximum spanning tree, using, for example, the Chu-Liu/Edmonds directed MST algorithm (Chu & Liu, 1965; Edmonds, 1967). The maximum spanning tree maximizes the votes for dependencies given the constraint that the resulting structure must be a tree. However, there is no guarantee against crossing branches. While this may seem undesirable, the resulting tree generated from the combination of parsers should rarely contain crossing branches (for English). In addition, this would be a suitable scheme for combining structures in free-word-order languages, where branches are expected to cross.

A second option is to use dynamic programming to “reparse” the sentence. We proceed just as if we were parsing the sentence using the CKY parsing algorithm for PCFGs, but we restrict the creation of new items in the CKY chart to pairs of words connected by the appropriate edges in the directed weighted graph, and assign these items the weight of their respective edges. Instead of the usual multiplication of probabilities, we simply add the values associated with each item used in the creation of a new item, and the value of the graph edge that allowed that item to be created. The resulting syntactic structure is guaranteed to be a tree with no crossing branches. Unlabeled dependency accuracy on the WSJ test set is only slightly lower than with the best-performing voting method, and labeled accuracy on the Eve test set is higher than with any voting method.

6.3 Handling Child Utterances by Combining Rule-Based and Data-Driven Approaches

There are two challenges in parsing child language that are not addressed by the data-driven GR analysis methods presented in chapter 5:

1. Certain utterances by young children cannot be annotated according to our GR annotation scheme, simply because they do not contain the syntactic structure associated with GRs. Young children may produce word strings that do not conform to a grammar that we can

interpret. Rather than to *guess* what the child is trying to say (but not saying), we should simply not annotate any GRs in such utterances. The data-driven systems we have developed, as most of the recent work in data-driven parsing, are inherently robust, assigning a syntactic structure to (almost) any string of words.

2. Young children may produce utterances that are mostly grammatical, but missing certain words (auxiliaries, possessive case markers, pronouns, prepositions, etc). Although the intent of the utterance is clear (and no guessing is required), words that are missing in the utterance make its analysis problematic for a data-driven parser trained on fully grammatical language.

We have seen that the data-driven parsing approaches outperform overall performance of our rule-based approach. However, the rule-based approach is better equipped to handle the two challenges mentioned above. First we address challenge (1). By using a small grammar with as little over-generation as possible, we can attempt to identify utterances where GRs should be found, and those where they should not. This is done simply by verifying if an utterance can be parsed using the grammar or not. While it may seem that the brittleness usually associated with rule-based systems may be a problem, there is a characteristic of the task that works to our advantage: the sentences in question tend to be very simple, so the grammatical coverage problem is greatly diminished. To address challenge (2), we use the rule-based parser's ability to perform limited insertions (one of the robustness features of LCFlex). By using the same grammar, we can also attempt to identify utterances where a word may be missing, and even determine what the word should be. If a missing word is correctly identified, it can be inserted in the sentence, which can then be passed as input to a data-driven system or a combination of systems as described in the previous section.

In the task of determining whether or not a sentence should be analyzed for GRs, we achieve better than 85% accuracy. In the task of identifying missing words, the system performs well in the most frequent cases (missing copula, missing possessive case marker). Certain less frequent (and more challenging) word insertions cannot be reliably, such as with missing prepositions. These rule-based techniques allow the overall performance of GR identification in young children's utterances to go from 69% to 87% accuracy.

7 Automated Measurement of Syntactic Development

In chapter 7 we present a practical end-to-end application of the methods for syntactic annotation and automatic analysis described so far. The task we explore is the automated measurement of syntactic development in child language, which has both clinical and theoretical value in the field of child language acquisition. Specifically, we present a fully automated way of computing the Index of Productive Syntax, or IPSyn (Scarborough, 1990), a popular measure for syntactic development that has traditionally required significant manual effort by trained researchers or clinicians.

In addition to its inherent value to the child language community, automatic computation of IPSyn scores serves as a task-based evaluation for our GR analysis approach. Although researchers in natural language parsing have become accustomed to evaluating systems in terms of precision and recall of certain pieces of information (such as constituent bracketing, or grammatical relations, as we have done in previous chapters), a syntactic analysis system often operates as a piece in a larger system designed to perform a task that goes beyond determining

parse trees or grammatical relations. Because task-based evaluations focusing on the effects of the performance of specific NLP components are relatively rare, the relationship between the standard precision/recall measures and the performance of larger systems that include these NLP components is still somewhat unclear. Through a task-based evaluation where we examine the results of an end-to-end system that computes IPSyn scores, we can determine the impact of the accuracy of our GR system in a practical setting. Accurate computation of IPSyn scores validates the usefulness of our annotation scheme and our approach to automatic GR analysis in its current levels of precision and recall of grammatical relations.

7.1 The Index of Productive Syntax (IPSyn)

The Index of Productive Syntax (Scarborough, 1990) is a measure of development of child language that provides a numerical score for grammatical complexity. IPSyn was designed for investigating individual and group differences in child language acquisition, and has been used in numerous studies. It addresses weaknesses in the widely popular Mean Length of Utterance measure, or MLU, with respect to the assessment of development of syntax in children. Because it addresses syntactic structures directly, it has gained popularity in the study of grammatical aspects of child language learning in both research and clinical settings.

Calculation of IPSyn scores requires a corpus of 100 transcribed child utterances, and the identification of 56 specific language structures in each utterance. These structures are counted and used to compute numeric scores for the corpus in four categories (noun phrases, verb phrases, questions and negations, and sentence structures), according to a fixed score sheet. IPSyn scores vary from zero to 112, with higher scores reflecting more syntactic complexity in the corpus. Language structures that must be identified for computation of IPSyn vary from simple patterns such as determiner-noun sequences to more complex structures such as embedded clauses, relative clauses and bitransitive predicates.

7.2 Automating IPSyn

Calculating IPSyn scores manually is a laborious process that involves identifying 56 syntactic structures (or their absence) in a transcript of 100 child utterances. Currently, researchers work with a partially automated process by using transcripts in electronic format and spreadsheets. However, the actual identification of syntactic structures, which accounts for most of the time spent on calculating IPSyn scores, still has to be done manually. Long, Fey and Channell (2004) have attempted, with limited success, to automate the computation of IPSyn using patterns of part-of-speech tags to search for the syntactic structures that are used in IPSyn scoring. Their Computerized Profiling (CP) program can be used for identification of simpler structures within IPSyn, but reliability of overall scores is well below manual level. Syntactic analysis of transcripts as described in chapters 4, 5 and 6 allows us to go a step further, fully automating IPSyn computations and obtaining a level of reliability comparable to that of human scoring. The ability to search for syntactic patterns using both grammatical relations and parts-of-speech makes searching both easier and more reliable. Automating the process of computing IPSyn scores consists of two main steps: (1) parse each sentence in the input transcript to obtain GRs according to the CHILDES annotation scheme; and (2) use patterns of GRs to search each sentence in the transcript for each of the syntactic structures named in IPSyn.

7.3 Evaluation

We evaluate our implementation of IPSyn in two ways. The first is *Point Difference*, which is calculated by taking the (unsigned) difference between scores obtained manually and automatically. The point difference is of great practical value, since it shows exactly how close automatically produced scores are to manually produced scores. The second is *Point-to-Point Accuracy*, which reflects the overall reliability over each individual scoring decision in the computation of IPSyn scores. It is calculated by counting how many decisions (identification of presence/absence of language structures in the transcript being scored) were made correctly, and dividing that number by the total number of decisions. The point-to-point measure is commonly used for assessing the inter-rater reliability of metrics such as the IPSyn. In our case, it allows us to establish the reliability of automatically computed scores against human scoring.

Using two sets of transcripts (41 transcripts in total) with corresponding IPSyn scoring that were provided to us by child language research groups, we measured the average point difference of our GR-based IPSyn system to be 3.3, and the point-to-point reliability to be 92.8%. For comparison purposes, CP's average point difference on the same transcripts is significantly higher at 8.3, and its point-to-point reliability is significantly lower at 85.4%. Inter-rater reliability among human coders is about 94%.

Our experiments show that the results obtained from automatic IPSyn scoring using our GR analysis for CHILDES data are only slightly less reliable than human scoring, and much more reliable than scoring based on part-of-speech analysis alone. This validates not only our analysis approach, but also our CHILDES GR annotation scheme.

8 Conclusions

This thesis presents a multi-strategy approach for syntactic analysis of transcripts of parent-child dialogs. Experiments using data from the CHILDES database show that our approach achieves high accuracy in the identification of grammatical relations. Through experiments using data from the Penn Treebank, we show that several of the ideas developed in this thesis are applicable not just to analysis of parent-child dialogs, but to parsing in general.

8.1 Summary of Contributions

The major contributions of this thesis are:

- A scheme for annotating syntactic information as grammatical relations in transcripts of child-parent dialogs focusing on information relevant to the study of child language. The annotation scheme is based on labeled dependency structures that represent several grammatical relations (GRs), such as subjects, objects and adjuncts.
- A linear-time classifier-based deterministic parsing approach for constituent structures. By using classifiers to determine the actions of a shift-reduce parser, we obtain high levels of precision and recall of constituent structures at a greater speed than methods with comparable accuracy.
- The application of the rule-based robust parsing techniques of Lavie (1996) and Rosé and Lavie (2001) to a high-precision grammatical relation identification system for parent-child dialogs. We use a multi-pass approach, allowing a gradual increase of coverage and

ambiguity by setting robust parsing parameters. Although the recall of grammatical relations obtained with the system is low, precision is very high, allowing the system to contribute as one of the components in a high-accuracy combination system.

- The development of data-driven parsers for grammatical relations, based on statistical and classifier-based parsing approaches. First, using the Penn Treebank (Marcus et al., 1993) and the Charniak (2000) parser, we show how existing resources and parsing technologies can be adapted to a different domain using a different syntactic representation. We then develop a classifier-based approach for constituents and labeled dependencies. We show that the performance of a classifier-based parser trained on a small corpus is comparable to that of a more complex system trained on a much larger corpus of text in a different domain.
- The use of different weighted voting schemes for combining different dependency structures, and a novel methodology for using parsing algorithms to combine grammatical relation analysis from multiple systems. We extend the work of Henderson and Brill (1999) on parser combination to the case where several dependency parsers are combined. We also develop new ways of determining different voting weights, and show that more specific weights can produce improved accuracy. Finally, we present a novel way of using maximum spanning trees or the CKY algorithm to combine the results of different parsers producing well- formed dependency structures.
- The demonstration of the effectiveness of the GR annotation scheme and GR identification approach through a task-based evaluation. We implement an automated version of the Index of Productive Syntax, or IPSyn (Scarborough, 1990), a measure of syntactic development in child language used by clinicians and researchers. We show that by using our GR parsing approach, we can produce IPSyn scores fully automatically with accuracy comparable to that of manual scoring. This serves not only as a useful tool for the child language community, but as a way of showing the quality and value of our syntactic analysis approach in a practical setting.

8.2 Future Research Directions

There are several possible directions for research that extends the work presented in this thesis. One area of future work is strongly tied to an important characteristic of the CHILDES database, which served as the source of much of the data used in our experiments. The CHILDES database contains transcripts in several languages, and the state of NLP resources for these languages varies greatly. No other language, however, offers us a wealth of resources comparable to what is available for English. This situation points to the research question of what level of success can be expected from the application of the work in this thesis to other languages. Work towards parsing different languages includes the investigation of the applicability of several of the analysis approaches developed for English data.

Another direction of future research is further exploration of the classifier-based framework for parsing. The classifier-based parsers in this thesis are deterministic, but it is possible to extend them to perform a beam search that considers multiple analyses and outputs a k-best list of parses while keeping a linear run-time, in a similar way as done by Ratnaparkhi (1997) in his maximum-entropy parser. Another area of improvement for classifier-based parsing is the set of features used to determine the parser's actions. One interesting possibility is the use of tree features, since the data structures operated on by the parser are in fact trees. This can be

accomplished with the use of support vector machines and tree kernels, or other structured classification techniques, such as the tree boosting algorithm by Kudo and Matsumoto (2004).

One issue that deserves further research in our parser combination approach is the selection of parsers to be included in the combination. Intuitively, we know that it is desirable to have parsers that rely on different methods for determining syntactic structures, and in practice we have relied on checking the performance of different parser sets on development data. While this simple strategy has produced good results, finding a scheme for parser selection that is robust to the addition of weak parsers or several very similar parsers remains a challenge. One possible direction for addressing this issue is the use of classification techniques, such as Henderson and Brill's (1999) use of a naïve Bayes classifier for selecting constituents from different parsers. Using such techniques, however, would require that scores can be assigned to different dependencies from different parsers depending on their parts-of-speech or GR types, as we found that such scoring produces superior results than parser-wide scoring, and simple yes/no decisions on the inclusion of dependencies would be insufficient for our maximum-spanning-tree or CKY combination approaches.

Finally, accurate parsing of child language transcripts allows for several future research directions in child language. One area that is directly related to an issue explored in this thesis is the measurement of grammatical complexity in child language. A measure such as IPSyn (Scarborough, 1990) was designed with the knowledge that transcripts would have to be scored manually. There is no doubt that such a constraint shaped the design of the scoring task. Although the design of entirely new metrics that make better use of the currently available technology is a worthwhile area of research, it is also one that is likely to require a great deal of expertise in child language research. A more tractable goal for natural language processing research is to find better ways to utilize parsing technology to determine scores for existing metrics, such as IPSyn. In our current implementation, we simply use an automated system to mimic the manual scoring process. Knowing the precision and recall levels of different GRs produced automatically allows us to sets of transcripts and corresponding scores to increase the reliability of automatic scoring even further. For example, instead of using whole counts of IPSyn items, fractional counts could be used depending on the expected accuracy of identification of GRs involved in particular items. A more ambitious (but entirely plausible) alternative is the use of machine learning techniques to go from a GR-annotated transcript to an IPSyn score without the use of manually encoded rules, allowing a system to learn the mapping from GRs to scores.

References

- Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
- Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 132-139). Seattle, WA.
- Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding*. Cambridge, MA: MIT Press.
- Chu, Y. J., & Liu, T. H. (1965). On the shortest arborescence of a directed graph. *Science Sinica*(14), 1396-1400.
- Collins, M. (1996). A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, CA.
- Collins, M. (1997). Three generative, lexicalized models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* (pp. 16-23).
- Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2004). TiMBL: Tilburg Memory Based Learner, version 5.1, reference guide. *ILK Research Group Technical Report Series*(04-02, 2004).
- Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards*(71B), 233-240.
- Henderson, J., & Brill, E. (1999). Exploiting diversity in natural language processing: combining parsers. In *Proceedings of the Fourth Conference on Empirical Methods in Natural Language Processing*. College Park, MD.
- Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics*, 24, 613-632.
- Kudo, T., & Matsumoto, Y. (2004). A boosting algorithm for classification of semi-structured text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain.
- Long, S. H., Fey, M. E., & Channell, R. W. (2004). *Computerized Profiling (Version 9.6.0)*. Cleveland, OH: Case Western Reserve University.
- MacWhinney, B. (1999). *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum Associates.
- MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Magerman, D. (1995). Statistical decision-tree models for parsing. In *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*.
- Marcus, M. P., Santorini, B., & Marcinkiewics, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19.
- Moerk, E. (1983). *The mother of Eve as a first language teacher*. Norwood, N.J.: ALEX.
- Nivre, J., & Scholz, M. (2004). Deterministic dependency parsing of English text. In *Proceedings of the 20th International Conference on Computational Linguistics* (pp. 64-70). Geneva, Switzerland.

- Ratnaparkhi, A. (1996). A maximum-entropy part-of-speech tagger. In *Proceedings of the the First Conference on Empirical Methods in Natural Language Processing*. Philadelphia, PA.
- Ratnaparkhi, A. (1997). A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*. Providence, RI.
- 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.
- Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*(11), 1-22.