

Automatic Parsing of Parent-child Interactions

Kenji Sagae
Brian MacWhinney
Alon Lavie

Carnegie Mellon University

Send correspondence to:

Kenji Sagae
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
Email: sagae@cs.cmu.edu

ABSTRACT

To evaluate the major theoretical proposals regarding the course of child language acquisition, researchers need to rely on the processing of large numbers of syntactically parsed utterances, both from children and their parents. However, hand parsing is an exceedingly difficult and unreliable process. Using the MOR tagger, a rule-based parser, and statistical disambiguation techniques, we developed a set of automatic procedures that can obtain nearly 80% correct parses for the sentences spoken to children. To achieve this level, we had to construct a particular processing sequence that minimizes problems caused by the coverage/ambiguity trade-off in parser design. These procedures are particularly appropriate for use with the CHILDES database, an international corpus of transcripts. The data and programs are made freely available over the Internet.

1 Introduction

Explaining the enigma of child language acquisition is one of the core 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, 1984).

To investigate these proposals, 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 a large amount of transcript data for over 25 human languages. There are now several hundred studies that have used the CHILDES database to study the development of morphosyntax. However, most of these studies have been forced to use the database in its raw lexical form, without tags for part-of-speech and without syntactic parses. Lacking this information, researchers have devoted long hours of hand analysis to locate and code the sentences relevant to their hypotheses. If tags and parses were available, these analyses could be automated, allowing investigators to conduct a wider variety of tests in a more reliable fashion.

As an initial move in this direction, a morphological tagger called MOR (MacWhinney, 2000) has been developed for English, French, German, Italian, Spanish, Japanese, and Cantonese. The results of the MOR tagger can be disambiguated using the POST program (Parisse, 2001). The level of accuracy of this combination of the MOR and POST programs has now reached levels as high as 95%, which is close to the current state of the art for automatic tagging (Garside & Smith, 1997). The current work seeks to build on these advances to add a deeper layer of syntactic information to the utterances in the CHILDES corpora. This additional structure contains information on the constituent structure found in these utterances, and the grammatical functions of these constituents.

The idea of annotating natural language corpora with syntactic structures is not a new one. Over the past decade a number of annotation efforts have resulted in large amounts of text annotated with syntactic parse trees. These collections are known as "treebanks" (Marcus, Santorini, & Marcinkiewics, 1993). However, none of the treebanks currently in existence specifically addresses the needs of child language acquisition research. In fact, the language found in these treebanks is often taken from written material, such as newspaper texts¹, and the annotation style is designed to facilitate the training of statistical language analysis tools, rather than the study of language acquisition.

¹ One exception is the annotated portion of the Switchboard corpus, which contains spontaneous telephone conversation between adults.

Creating even a relatively small annotated corpus (15,000 sentences) would require months of work by a trained linguist. Recent advances in natural language processing have created the possibility of performing automatic (or semi-automatic) syntactic analysis of natural language with a high degree of accuracy. This analysis is commonly referred to as “automatic syntactic parsing”, or simply “parsing”. In this article, we describe our efforts in using state-of-the-art natural language analysis technologies to parse the parent language used in one of the corpora from the CHILDES database, creating annotations suitable for use in the research of child language acquisition.

2 Parsing

The main idea in syntactic parsing is to use a computational model of natural language to analyze a sentence and produce a syntactic structure of the sentence as output. The syntactic structure may take several forms, such as a constituent tree (c-structure or parse tree), a syntactic feature structure (or f-structure), or a dependency structure (figure 1). The different syntactic representations of a sentence may differ in the level of information they contain (parts-of-speech, case, syntactic function labels, etc.), but each of them describe in some way how words combine to form a sentence. The choice of syntactic representation, and therefore the choice of a syntactic parser, depends mainly on the purpose the syntactic analyses will serve.

Natural language processing researchers have made use of a number of existing models and theories of language to develop systems that perform syntactic parsing with increasingly high levels of accuracy. Rule-based parsers rely on a small set of grammatical rules that implement specific linguistic theories and principles of syntax (Hauser, 1999). Statistical parsers rely on regularities in the data and the training corpus to extract parts of speech and cooccurrence regularities (Charniak, 1997). Whichever design is chosen, no parser is able to achieve completely accurate results. There are at least four reasons for these problems. First, natural language is highly ambiguous. As native speakers with good intuitions, we often fail to sense the scope of this ambiguity. However, when we come to articulating a system for automatic parsing, the fundamental ambiguity of attachment, roles, interpretations, and relations in human language stands as a formidable challenge. Second, much of the information needed to determine the correct parse lies outside the scope of the sentence, in the domain of the discourse or the situational context. Third, most parsers are trained using material from a specific genre or area of language use. When the parser is then extended to cover material from a new domain, often there is a serious decrement in performance. Finally, there are some problems that are unique to the task of parsing spoken language. Specifically, spoken language differs from written language by including large numbers of ungrammaticalities, dysfluencies, filled pauses, retracings, and other conversational features.

In addition to the inherent ambiguity commonly expected in natural languages, parsers usually face a larger number of choices than one would expect for the analysis of all but the simplest sentences. Allowing for the analysis of a large number of syntactic constructions leads to the development of a model

that suffers from more ambiguity than a more restrictive model. The balance of coverage and ambiguity is a crucial issue in parsing. As an example, consider the simple context-free grammar and sentences:

- | | |
|-----------------|---------------------|
| (1) S -> NP VP | (7) DET -> the |
| (2) NP -> DET N | (8) N -> boy |
| (3) NP -> PRO | (9) N -> dog |
| (4) VP -> V NP | (10) N -> telescope |
| (5) VP -> VP PP | (11) P -> with |
| (6) PP -> P NP | (12) PRO -> I |
| | (13) V -> saw |

(S1) I saw the dog with the telescope.

(S2) I saw the boy with the dog.

In spite of the similar part-of-speech sequences in sentences S1 and S2, their syntactic structures differ in the place where the prepositional phrases (PP) “with the telescope” and “with the dog” should be attached. Sentence S1 can be paraphrased as “I used the telescope to see the dog,” while sentence S2 could not be paraphrased as “I used the dog to see the boy.” A more plausible paraphrase for S2 would be “I saw the boy who had a dog with him.” In S1, the prepositional phrase “with the telescope” is an adjunct to the verb phrase, since it modifies the way the action of seeing was done. In S2, however, “with the dog” does not modify the action, but rather the noun phrase (NP) “the boy,” and that is where it should be attached (figure 2a).

While sentence S1 has the acceptable analysis shown in figure 2c according to our simple grammar, the analysis obtained for sentence S2 is incorrect (figure 2b). To get the acceptable analysis in figure 2a, the grammar would need an additional rule:

- (2') NP -> NP PP

However, the addition of a new rule to handle a previously uncovered syntactic structure may have adverse side effects in the overall performance of the grammar. Even though the addition of rule 2' to the grammar allows for the correct analysis of sentence S2 (figure 2a), the incorrect analysis (figure 2b) is still possible. What is even worse is that this modification allows for the incorrect analysis of sentence S1 (figure 2d), which could only be analyzed correctly and unambiguously before the addition of rule 2'. While there are ways to resolve the ambiguities in sentences S1 and S2 and produce the correct analysis in each case (for example, by using additional knowledge sources, more complex grammar constructions, feature unification, or statistical disambiguation models), this example illustrates how increasing the coverage of a grammar may result in unwanted ambiguity.

3 The CHILDES Database and the Eve Corpus

Among the corpora in the CHILDES database, we chose to focus on the Eve corpus (Brown, 1973). Our choice was motivated by the fact that we² had already created a clean transcription with manually verified part-of-speech tags for the child utterances, as well as its central role in child language acquisition research (Moerk, 1993). The corpus includes utterances from the child (Eve), as well as her parents. An example of child utterances in the corpus can be seen in figure 3. In this example, the first line is a transcription of one of Eve’s utterances (indicated by *CHI:), and the following line contains part-of-speech and morphological annotations for that utterance (indicated by %mor:). Adult sentences in the corpus include a line with part-of-speech information, but that information is produced fully automatically and is often ambiguous. An example can be seen in figure 4.

Our work used rule-based parsing techniques to analyze each adult utterance in the corpus to produce syntactic annotations in the form of syntactic feature structures (marked by %fst:), as illustrated in figure 5. Figure 6 shows a graphical representation of the annotations in figure 5. The syntactic feature structures we produced resemble the feature-value pairs of Lexical Functional Grammar (Bresnan, 2001), although we made no attempt to follow LFG theory closely. In our f-structures, the features are typically syntactic functions, and the values are syntactic constituents or syntactic characteristics of the sentence. The `index` feature provides a cross-reference between a feature structure and its corresponding constituent structure.

Once a corpus has been annotated with automatically generated syntactic feature structures, a corpus browser allows a user to view a graphical representation of the syntactic analysis for a sentence, and label it as correct or incorrect (figure 7). While the annotation process still relies on human expertise at this step, the time required to judge the correctness of a feature structure is only a small fraction of the time it would take a person to generate the analysis. When the analysis of a sentence is found to be incorrect, the user may enter a comment to accompany the analysis.

While the adult language in the CHILDES corpora is usually grammatically acceptable, 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. We believe that the child and adult utterances differ significantly enough that we may be able to analyze them more accurately by doing so separately, possibly with different strategies. In this paper, we explore the “easier” (in the sense that it is better defined) problem of analyzing the adult utterances in the Eve corpus, whose role in child language acquisition has been the subject of extensive research (Moerk, 1993). Although parsing of the adult input is easier than parsing of the child’s forms, it is theoretically of equal importance, since theories of learning depend heavily on consideration of the range of constructions provided to children in the input (MacWhinney, 1999).

² Thanks to Yuriko Oshima-Takane and her students at McGill for producing these disambiguated tags for training of POST.

4 The Syntactic Analysis System

To produce the syntactic analyses necessary for annotation of the corpus, we developed a syntactic analysis system based on grammar-based robust parsing and statistical disambiguation. Grammar-based (or rule-based) parsers use a set of production rules that specify how each syntactic constituent may be expanded into other constituents or words as a model of natural language. The type of grammar used by our system follows a formalism based on a context-free backbone augmented with feature-unification constraints. As an example, figure 8 shows a simple grammar that can be used with our system.

The grammar in figure 8 can be used to parse the sentence “He sees the ball” as follows: a lexical analysis of the words in the input determines the part-of-speech and agreement features (in the cases of “he” and “sees”) of each word. Rule 3 allows the formation of a noun phrase (NP) from the pronoun (PRO) “he”. The single unification equation associated with that rule states that every feature in the first element of the right-hand side of the context-free portion of the rule (represented in the equation by x_1 and corresponding to the pronoun) will be passed to the newly formed constituent, or the left-hand side of the context-free rule (the noun phrase, represented in the equation by x_0). Rule 2 forms another noun phrase from the words “the” and “ball”. The first equation in the rule specifies that the first element in the right-hand side of the context-free rule (DET, represented in the equation by x_1) is the value of a feature named DETERMINER of the second element of the right-hand side. The second equation specifies that every feature of the second element of the right-hand side (including the newly specified DETERMINER feature) be passed to the newly created constituent (NP). Rule 4 can then be applied to form a verb phrase (VP) from the verb “sees” and the noun phrase “the ball”. According to the first equation of rule 4, the noun phrase becomes the OBJECT of the verb. Finally, the noun phrase “he” and the verb phrase “sees the ball” can be combined by rule 1 to produce a sentence (S) constituent that spans the entire input string, completing the parse. The first equation of rule 1 requires that the value of the AGREEMENT features of the verb phrase and noun phrase match. These values are provided by the lexical analysis performed before the parsing process. The second equation makes the noun phrase the subject of the sentence. The final analysis can be seen in figure 9.

Robust parsing technologies seek to augment the coverage of a parser by allowing it to analyze language phenomena that fall outside of the coverage of the parser’s model (in our case, a syntactic grammar). Our use of robustness is targeted towards the analysis unforeseen spoken language phenomena. This is achieved by allowing the parser to: (1) insert lexical or non-terminal items (constituents) that are not present in the input string, and (2) skip certain words in the input string. The specific uses of these techniques are discussed in section 5.4. While the expansion of coverage provided by robust parsing increases the ambiguity problem faced by the analysis system, we employ statistical techniques to allow the parser to cope with such ambiguity. By providing a training corpus of manually disambiguated sentences (of much smaller size than the total amount of text we ultimately analyze), we can build a statistical model

of grammar usage to make certain decisions in the parsing process that result in fairly accurate disambiguation.

The input to our system is a sequence of transcribed utterances, and the output is a syntactic analysis for each of those utterances, as seen in the example in section 3. At a lower level, the system can be divided into three main components (figure 10):

1. POST, a part-of-speech tagger developed especially for the CHILDES database and its custom set of part-of-speech tags for child-parent communication (Parsisse, 2001). POST operates on the tags inserted by the MOR program to provide an unambiguous morphosyntactic analysis at the lexical (word) level with high accuracy.
2. LCFlex (Rosé and Lavie, 2001), a robust parser that provides special features for parsing spoken language. Because the corpora in the CHILDES database consist only of transcribed spontaneous speech (with its dysfluencies and ungrammaticalities), having a parser designed to handle such language is of great importance. Through a set of parameters, LCFlex can be tuned to allow the insertion of specific missing syntactic constituents into a sentence, and to skip extra-grammatical material that would prevent an analysis from being found with the grammar in use. These features allow great flexibility in parsing spoken language, but their parameters must be tuned carefully to balance benefits and the increased ambiguity caused by allowing insertions and skipping.
3. A statistical disambiguation module to pick the correct analysis from the many produced by the parser. The idea behind syntactic disambiguation in LCFlex is that each analysis of a particular utterance is obtained through an ordered succession of grammar rule applications, and the correct analysis should be the one resulting from the most probable succession of rules. The probability of each competing analysis is determined based on a statistical model of bigrams of rule applications (Rosé and Lavie, 2001) obtained from training examples consisting of sentences and their correct analyses.

One of the goals of our work is to reduce the overall need for manual work to the point where reliable annotations could be generated for 80% of a 15,000 sentence corpus in just a few days, as opposed to the months of work that a trained linguist would require to produce the annotations from scratch. However, it will still be necessary for a linguist to check over the results of the automatic parsing. This check only involves a binary decision. In 80% of the cases, the linguist simply has to accept the results of the parser. In the remaining 20%, further processing will be needed.

5 Tailoring a High Performance Analysis System

To obtain high quality syntactic analyses from a system composed of the pieces described in the previous section, each component must be tuned carefully, keeping in mind the behavior of the overall

system. This section focuses on the specific issues that relate to the components of the parser, as well as their integration into a high performance analysis system.

5.1 Grammar

The grammar needed by LCFlex consists of context-free rules augmented with feature unification constraints. Although general-purpose English grammars are available, we found that they were not suitable for our analysis task. The main problems associated with “off-the-shelf” grammars are related to the large amount of ambiguity allowed such grammars (an unavoidable consequence of a grammar designed to analyze unconstrained English sentences), and the lack of support for certain phenomena we find in corpora in the CHILDES database (such as the extensive use of communicators and vocatives commonly used in casual spoken language, onomatopoeia, etc.). It has been reported that practical grammars used to analyze newspaper articles written in English produce an astronomical number of parses per sentence (Moore, 2000), with the vast majority of these parses being completely uninterpretable from a human point-of-view. As a simple example of this phenomenon, Charniak (1997) uses the sentence “Salespeople sold the dog biscuits” and a grammar not unlike the one considered in section 2, but including the rule

NP -> NP NP

Although this rule may seem unusual, it is used to analyze phrases such as “10 dollars a share,” where both “10 dollars” and “a share” are noun phrases that combine to form a larger noun phrase. Charniak gives three analyses for his example sentence (figure 11). While the first two analyses can be easily interpreted as “dog biscuits were sold by the salespeople,” and “biscuits were sold to the dog by the salespeople,” respectively, the third analysis does not seem to have a meaningful interpretation. In fact, it was the result of the application of a rule designed to cover a syntactic construction not present in any plausible interpretation of this sentence. The interested reader is encouraged to see Charniak (1997) for a more detailed account of the ambiguity problem in practical natural language grammars.

Because of the nature and the domain of our target corpus, it does not contain many of the complex syntactic constructions found in newspaper-style text, or even adult conversations. We can take advantage of that fact and attempt to reduce ambiguity by using a grammar that fits our target language more tightly. It is a fairly accepted notion in natural language processing that parsing within a specific domain can be made more accurately with the use of domain-specific resources.

Starting with a general-purpose English grammar with about 600 rules, we pruned or simplified a large number of rules that would never (or rarely) be used in correct analyses for the target corpus. For example, the noun phrase rule mentioned above can be safely discarded, since we are not interested in covering constructions such as “10 dollars a share.” The result was a completely rewritten compact grammar with 152 rules. This final grammar included rules to handle the specific language phenomena likely to appear in the CHILDES database, and represents a much cleaner and tighter model of the language

in the domain we are attempting to analyze. As a result, the potential for ambiguity in parsing was significantly reduced.

5.2 Lexical Ambiguity

Even though a more suitable grammar is a first step towards managing ambiguity, it is not a complete solution to the problem, and further techniques to resolve syntactic ambiguity are needed. One such way is to eliminate lexical ambiguity by selecting a single part-of-speech tag for each word, using the part-of-speech tagger. The first step is to have a corpus of correctly tagged text to train the tagger. Unfortunately, the CHILDES database contains no unambiguous part-of-speech tagged data for adult utterances. While tagged data for child utterances are available, the child and adult languages are significantly different so that a tagger trained on child utterances would perform poorly in tagging adult ones. To create a part-of-speech tagging training corpus for the adult language in the corpus, we used the following bootstrapping process:

1. Use tagged child utterances to train a part-of-speech tagger for adult utterances.
2. Tag adult utterances (4,000 words) and hand correct them.
3. Retrain the tagger with the newly corrected data, and iterate from step 2.

By performing four iterations of the procedure above, we improved the accuracy of the part-of-speech tagger from an initial 87.2% to 94.3%. The improvement in accuracy for each iteration decreased at a rapid pace, and it is unlikely that further iterations would yield significant benefits (at least not cost-effectively).

5.3 Syntactic Ambiguity

Once syntactic ambiguity has been reduced through the elimination of lexical ambiguity, we can attempt to find the single correct analysis produced by the parser (when one exists) using statistical disambiguation. For that, we need a training corpus of correct sentence-analysis pairs. We create this data in a way similar to the bootstrapping process used to generate part-of-speech training data, but this time we start with the results of parsing lexically unambiguous input.

1. Parse all of the part-of-speech tagged utterances.
2. Examine the analyses that are unambiguous (or nearly unambiguous).
3. Add correct analyses to the training corpus.
4. Train the statistical disambiguation module.
5. Use an iterative process similar to the one used for obtaining the part-of-speech training corpus.

We started with an initial training corpus for statistical disambiguation of less than 500 sentences, and increased its size to 3,000 sentences in four iterations of the process described above. As with the process for building a part-of-speech training corpus, the benefits of successive iterations decreased at a fast rate. Lexical disambiguation less to unambiguous parses for only a few sentences. However, these sentences

can be very useful in obtaining an initial training corpus for the statistical disambiguation module, since resolving large amounts of syntactic ambiguity manually may be a practically intractable task.

As a way to enhance the performance of the procedures described above for building training corpora for part-of-speech tagging and syntactic disambiguation, we can exploit the interactions between part-of-speech tagging and parsing. Once we obtain an initial training corpus for statistical disambiguation, we can increase its size while also increasing the size of our part-of-speech tagging training corpus by using a feedback loop between part-of-speech tagging and parsing. We assume that the input sentences for which the parser produces correct analyses have correct part-of-speech tag assignments, and we add those sentences to our part-of-speech training corpus. Improvements in part-of-speech tagging, in turn, result in more correct analyses being produced by the parser.

5.4 Parser Flexibility

Even after grammar development, a large number of sentences in the Eve corpus still could not be parsed with our compact grammar due to specific characteristics of the casual conversational language in the corpus (not due to general syntactic structures). The majority of such sentences were not covered successfully because of omitted words or filled pauses in otherwise fully grammatical utterances, for example:

- Missing auxiliary verbs in questions (“[Do] You want to go outside?”);
- Missing noun phrases, as elided subjects (“[I] Don’t think she wants to play now.”), and even as elided objects (“Give [it] to me.”);
- Missing auxiliary verbs and noun phrases (“[Do] [you] Want to go outside?”);
- Filled pause (“I’d like to tell you, *uh*, something.”).

Adding explicit ad-hoc grammar rules to handle such sentences would cause the grammar to deviate from the clean model of language we were hoping to achieve, and add much harmful ambiguity to the analysis process. Instead, we turned to the robustness features of the parser to handle these sentences. LCFlex allows the addition of specific syntactic nodes to an analysis, or skipping of words in a sentence, making it conform to the grammar and leading to a successful analysis.

5.5 Balancing Coverage and Ambiguity

We now examine the effects of each of the strategies above on the coverage/ambiguity trade-off.

5.5.1 Decreasing Ambiguity

Parsing with the initial general English grammar and without proper training of the disambiguation module yielded very few correct analyses due to ambiguity. We only considered an analysis “correct” if it contains

no errors or ambiguity. Using the initial general grammar coupled with the statistical disambiguation provided by LCFlex, we obtained less than 50% accuracy in analyzing the Eve corpus (measured with a 200 utterance test corpus). We define accuracy as the ratio between correct analyses and the total number of sentences analyzed. Although incorrect analyses often contained correctly analyzed portions that can be considered useful information, our evaluation methodology only considers “correct” an analysis that contains no errors. Using our final rewritten grammar and statistical disambiguation (trained on 3,000 correctly parsed utterances), we reached close to 65% accuracy. This reflects both the improvement in ambiguity resolution and some gain in coverage.

Using the final grammar with part-of-speech tagged input sentences to eliminate lexical ambiguity, the number of correct analyses decreased to 57.5%. However, the set of correct analyses obtained with this setup is not a subset of the set of correct analyses obtained with lexically ambiguous input. Although the ratio of correct analyses over non-failed analyses increases, the system failed to analyze (and produces no output for) a large number of utterances due to errors in part-of-speech assignments. In terms of the trade-off, we decreased ambiguity significantly, but at the cost of a severe reduction in coverage. We achieved a 1.1% improvement in part-of-speech tagging using transformation-based learning of Brill-style rules (Brill, 1995), which resulted in a slight improvement in coverage. However, overall parser accuracy obtained with part-of-speech tagged input was still under 60%.

5.5.2 Increasing Coverage

Setting the parser to allow limited insertions (a single noun-phrase and/or a single auxiliary may be inserted during parsing) led to an improvement in recognition for about 5% of the sentences in the Eve corpus. However, the percentage improvement in accuracy is less than 3%, due to the increased ambiguity and over-generation that results from increasing the search space of possible analyses with insertions. Allowing limited skipping (a single word in the input utterance may be skipped during parsing) actually decreases the overall accuracy. In other words, the number of sentences that are parsed incorrectly due to the increased search space is greater than the number of correct analyses that result from limited skipping.

5.5.3 Putting it all together

Each of the strategies to reduce ambiguity or increase coverage described above has a different impact on the coverage/ambiguity trade-off, and the effect of applying them together by naively combining them all at once is far from optimal. In summary, our efforts to reduce ambiguity come at the cost of reducing coverage, and our efforts to increase coverage result in much increased ambiguity. By applying lexical disambiguation, limited insertions and skipping, and relying only on the statistical model of bigrams of rule applications for parse selection, we achieve less than 70% parsing accuracy with the Eve corpus.

We must then attempt to balance the coverage/ambiguity trade-off to benefit from both decreased ambiguity and increased coverage. We do so by controlling the amount of ambiguity and coverage in

several passes of parsing. We start with the most restrictive settings and the least ambiguity, and upon failures in parsing, gradually increase coverage (and ambiguity). The idea is that we only pay the cost of an increased search space as it becomes necessary, taking advantage of both more limited ambiguity when possible, and increased coverage when needed. Through empirical observation, we arrived at the settings shown in table 1 for each of the passes. In this way, we reached 78.5% correct parses – the highest level we were able to obtain. In the first pass, we parse lexically unambiguous input, and use no coverage-enhancing techniques. From passes two through six, we allow limited lexical ambiguity, and gradually increase coverage through the robust parsing features of LCFlex. Limited lexical ambiguity means that not every possible part-of-speech tag (according to a lexicon available with the CHILDES database) is allowed for each lexical item, which would cause a greater increase in syntactic ambiguity. Instead, we only allow lexical ambiguity for certain lexical categories where the automatic part-of-speech tagger was observed to make frequent mistakes, causing parser failures. We determined those highly confusable parts-of-speech simply by analyzing the cause of failed analyses, and keeping track of the parts-of-speech most frequently associated with those failures. The following sets of tags accounted for more than 95% of failures caused by a part-of-speech tagging error: {verb, auxiliary}, {verb particle, preposition}, {adverb, adjective}, {noun, verb}.

The reason for not combining multiple coverage-increasing techniques in further passes of parsing is that we prefer having no analysis for an utterance to having an analysis that is very likely to be incorrect. This multi-pass approach not only increases ambiguity gradually only as needed, but also allows us to have some sense of how confident we are that an analysis is correct. Figures 12 and 13 illustrate how our final multi-pass analysis system works.

6 Results

6.1 Evaluation

To assess the effectiveness of our methods, we evaluated our current system on 200 randomly chosen previously unused utterances from the Eve corpus, and checked their generated syntactic analyses for correctness. The contribution of each of the six passes to the total number of correct analyses can be seen in the table 2. The overall level of correct parses obtained here is 78.5%. The causes of the remaining errors in incorrect analyses are shown in table 3. The row labeled “insertion” refers to the utterances that were not covered by the grammar but were assigned an incorrect analysis due to limited insertions. The row labeled “over-generation” refers to utterances for which the parser did not produce an appropriate analysis due to lack of grammar coverage, but where the utterance was still covered in an incorrect way due to grammar over-generation. Finally, the causes of parsing failures where no analysis was produced for an utterance is shown in table 4.

6.2 Availability

One of the main goals of this project is to provide data to the language acquisition research community. The current results of the research described in this paper (a version of the Eve corpus with syntactic annotations for adult utterances), as well as related tools and other resources, are available for research purposes at the CHILDES web site (<http://childes.psy.cmu.edu>), or by request from the authors. It is our hope that the data we have produced will be useful in current research efforts in language acquisition³, as well as inspire and fuel new research on natural language learning and various aspects of grammar acquisition.

7 Conclusions and Future Work

Our system is quite effective in producing accurate syntactic annotations for the adult utterances in the Eve corpus. The number of incorrect analyses is acceptably small, making the task of manually checking and possibly correcting the resulting annotations fairly manageable, or even unnecessary if an error rate of about 10% can be tolerated. Practically all of the utterances that failed to be analyzed by the system were not handled due to the occurrence of rare syntactic constructions. Although increasing grammar coverage through the creation of new rules to properly handle these constructions is possible, the increased ambiguity resulting from a larger grammar may hurt the overall accuracy of the system. The net effect of such an increase in both coverage and ambiguity remains to be investigated.

Although our efforts to produce syntactic annotations for the child utterances in the corpus (as opposed to the mother's utterances) is still in very early stages, a preliminary evaluation of the current system on a set of such utterances revealed that more than 60% of them could probably be analyzed correctly with the system as-is. However, significant changes to the overall system would be necessary for analyzing a high percentage of the child utterances accurately and reliably. We are currently working on a different analysis strategy for child utterances, which acknowledges both the global (utterance level) differences and local (fragment or constituent level) similarities between the child and adult languages in the corpus. The analyses produced with this strategy report constituents found in child utterances, without trying to combine them into single global structures when the utterances are ill formed (according to our adult grammar). Our initial heuristic in searching for these constituents is to try to cover as much of the utterance as possible, with as few constituents as possible. Although we recognize the simplistic nature of this approach, our preliminary experiments have yielded very promising levels of accuracy in the analysis of child utterances in the Eve corpus. Further research on analyzing child language is planned as the immediate next step in our work. We also plan to investigate of the effectiveness of the current system on

³ See the word order acquisition investigation in (Villavicencio 2000), which has used similar data in lesser amounts, for an example.

other corpora in the CHILDES database, and possibly the automatic adaptation of the system to other corpora.

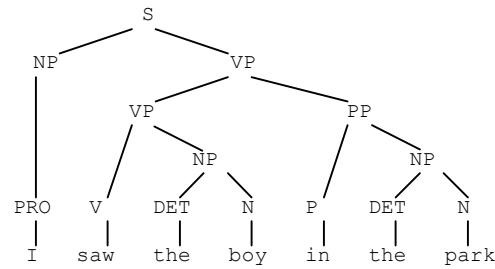
8 References

- Bresnan, J. (2001). *Lexical-functional syntax*. Blackwell publishers, Oxford, UK.
- Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging, *Computational Linguistics*, vol 21.
- Brown, R. (1973). *A first language: The early stages*. Harvard, Cambridge, MA, USA.
- Carrol, J., Minnen, G. and Briscoe, T. (1999) Corpus annotation for parser evaluation, *Journée(s) ATALA sur les corpus annotés pour la syntaxe*. Paris, France.
- Charniak, E. (1997) Statistical Techniques for Natural Language Parsing, *AI Magazine*, vol 18.
- Day, D., Aberdeen, J., Hirschmann, L., Kozierek, R., Robinson, P., and Vilain, M. (1997). Mixed-initiative development of language processing systems, *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*. Washington, DC, USA.
- Garside, R., Leech, G., and McEnery, A. (eds.) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, England.
- Garside, R., and Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 102-121.
- Hauser, R. (1999) *Foundations of computational linguistics*. Berlin: Springer.
- Lieven, E., Pine, J., and Baldwin, G. (1997) Lexically-based learning and early grammatical development, *Journal of Child Language*, 24, 187-219.
- MacWhinney, B. (Ed.) (1999). *The emergence of language*. Lawrence Erlbaum Associates: Mahwah, NJ.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*, Lawrence Erlbaum Associates, Hillsdale, NJ, USA.
- McEnery, A.M., Tanaka, I. and Botley, S.P. (1997) Corpus annotation and reference resolution, in Mitkov, R. & Boguraev, B. (eds) *Proceedings of the Association for Computational Linguistics Workshop on Anaphora Resolution for Unrestricted Texts*, Madrid.

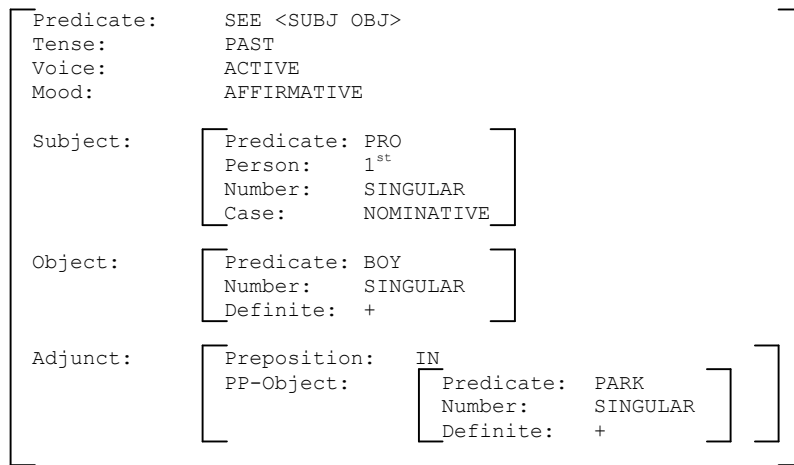
- Marcus, M. P., Santorini, B., and Marcinkiewics, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank, *Computational Linguistics*, vol 19.
- Moerk, E. (1983). *The mother of Eve as a first language teacher*. ABLEX, Norwood, N.J., USA.
- Moore, R. C. (2000). Improved left-corner chart parsing for large context-free grammars. In Proceeding of the Sixth International Workshop on Parsing Technologies.
- Parisse C., and Le Normand, M.T. (2000) Automatic disambiguation of morphosyntax in spoken language corpora, *Behavior Research, Methods, Instruments and Computers*, 32, 3, 468-481.
- Rosé, C. P., and Lavie, A. (2001). Balancing robustness and efficiency in unification-augmented context-free parsers for large practical applications, *Robustness in language and speech technology*, van Noord and Junqua (eds.), ELSNET series, Kluwer Academic Press.
- Skut, W., Krenn, B., Brants, T, and Uszkoreit, H. (1997). An Annotation Scheme for Free Word Order Languages, *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*. Washington, DC, USA.
- Villavicencio, A. (2000). The acquisition of word order by a computational learning system, *Proceedings of the 2nd Learning Language in Logic Workshop*. Lisbon, Portugal.

Sentence: I saw the boy in the park.

Syntactic constituent structure:



Syntactic feature structure:



Syntactic dependency structure:

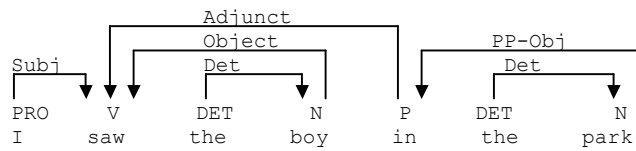


Figure 1: Syntactic constituent structure, feature structure and dependency structure representations of a sentence.

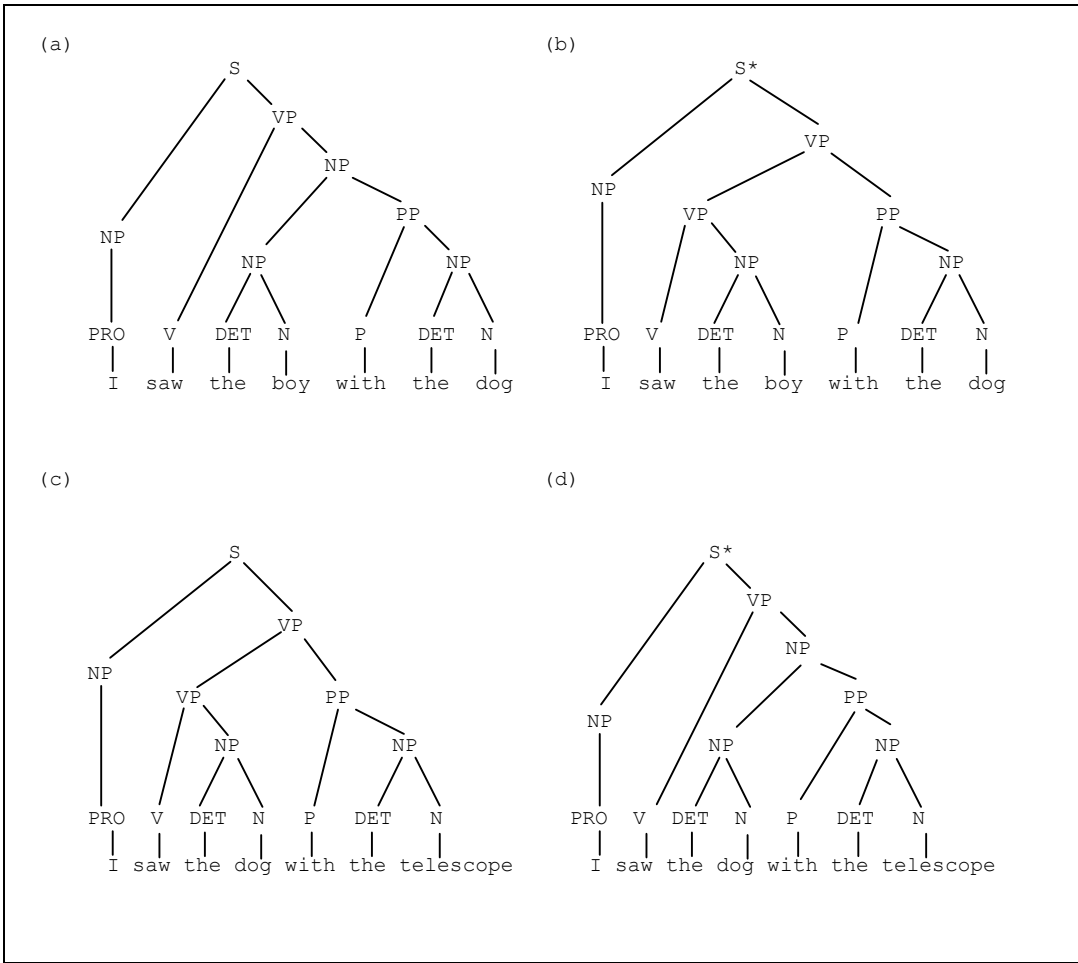


Figure 2: Parse trees for “I saw the boy with the dog” and “I saw the dog with the telescope.” The trees in (a) and (d) can be generated with the given grammar, but the desired trees are (a) and (c). Adding a rule to allow for the analysis in (c) has the undesired side effect of also allowing the tree in (b).

```
*CHI: more cookie.  
%mor: qn|more n|cookie .
```

Figure 3: Sample child utterance from the Eve corpus

```
*MOT: how about another graham  
cracker ?  
%mor: adv:wh|how  
prep|about^adv|about  
det|another n|graham  
n|cracker ?
```

Figure 4: Sample adult utterance from the Eve corpus

```

*MOT:  you kicked it .
%mor:  pro|you v|kick-PAST pro|it .
%fst:  ((mood *declarative) (tense *past)
        (index 2)
        (subject ((cat pro) (num *sg) (pers 2)
                  (case *nom)(index 1) (root *you)))
        (object ((cat pro) (sum sg) (pers 3)
                  (case acc) (index 3) (root *it)))
        (root *kick) (cat v))
%cst:  (sentence (decl (np (pro you))
                    (vp (vbar (v kicked)
                              (np (pro it))))))
        (period .))

```

Figure 5: Sample syntactic annotations in the Eve corpus

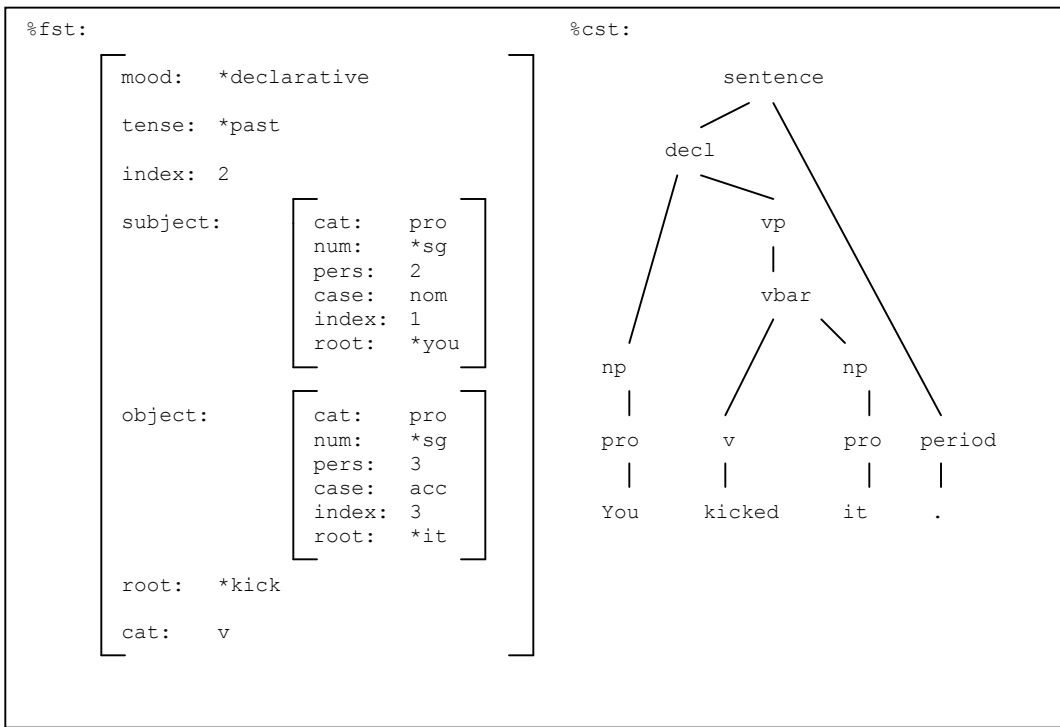


Figure 6: Graphical representations of the `%fst` and `%cst` lines in figure 5.

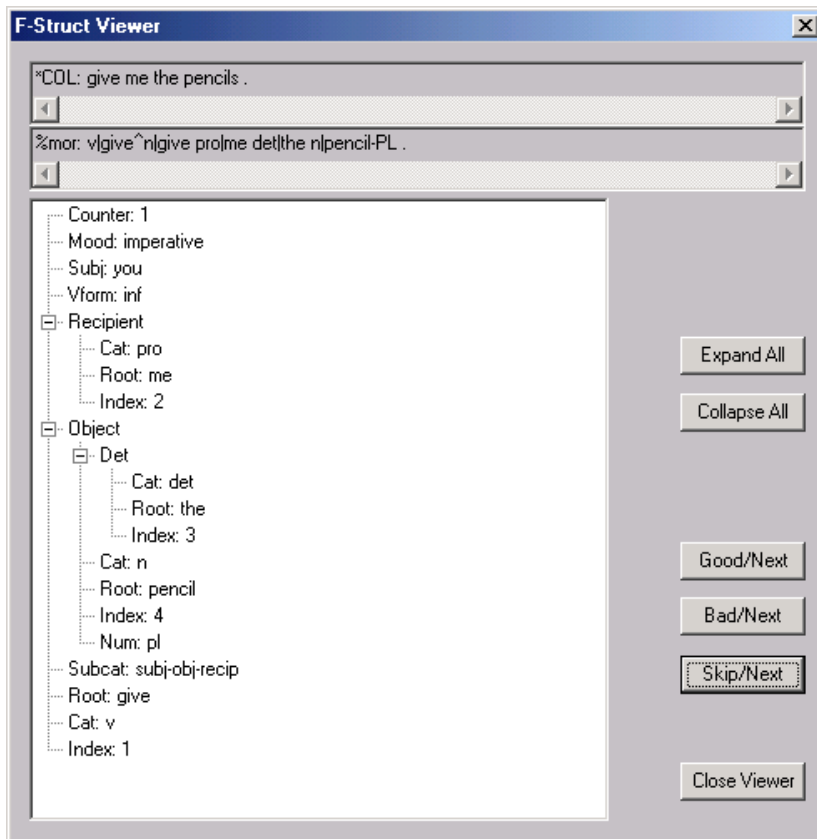


Figure 7: Feature structure viewer. The user may rate the analysis as correct or incorrect simply by using the appropriate buttons.

```
(1) S -> NP VP
    (x1 AGREEMENT) =c (x2 AGREEMENT)
    (x2 SUBJECT) = x1
    x0 = x2

(2) NP -> DET N
    (x2 DETERMINER) = x1
    x0 = x2

(3) NP -> PRO
    x0 = x1

(4) VP -> V NP
    (x1 OBJECT) = x2
    x0 = x1
```

Figure 8: A simple grammar composed of a context-free backbone and unification equations

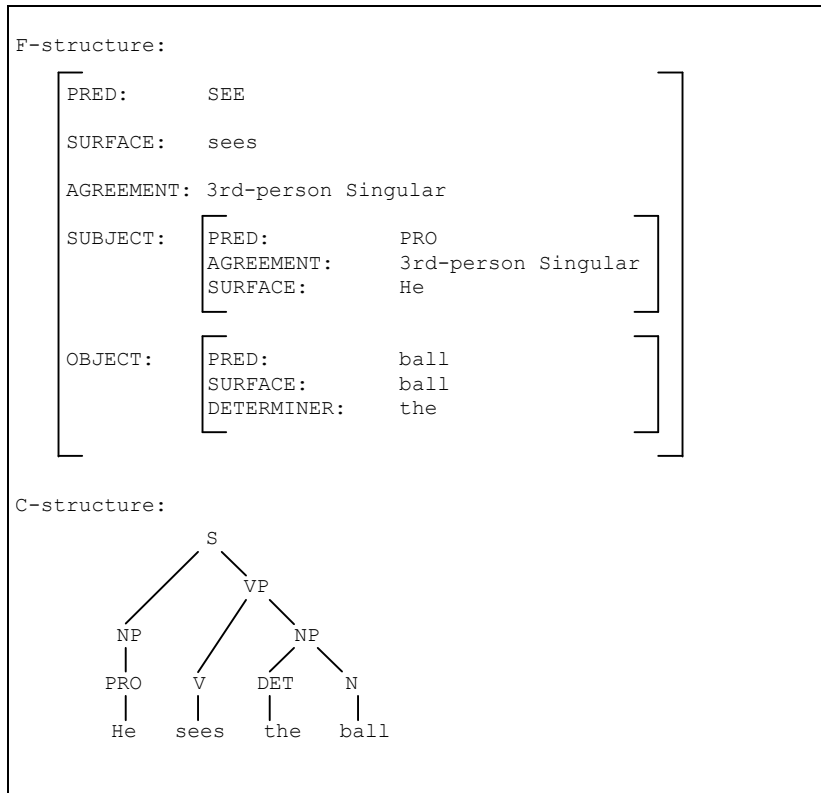


Figure 9: Analysis of the sentence “He sees the ball” according to the grammar in figure 8. The PRED, AGREEMENT and SURFACE features are created during lexical analysis.

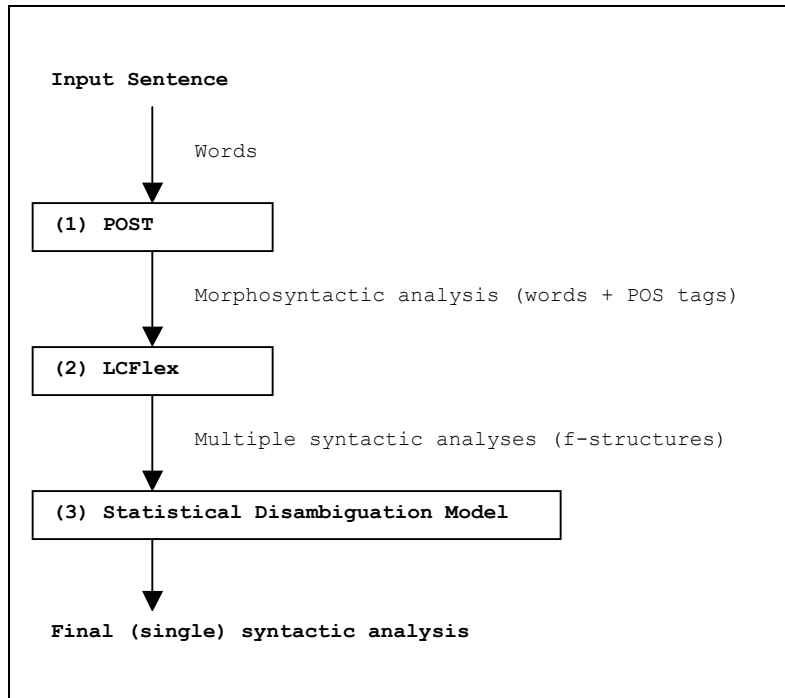


Figure 10: Components of the analysis system

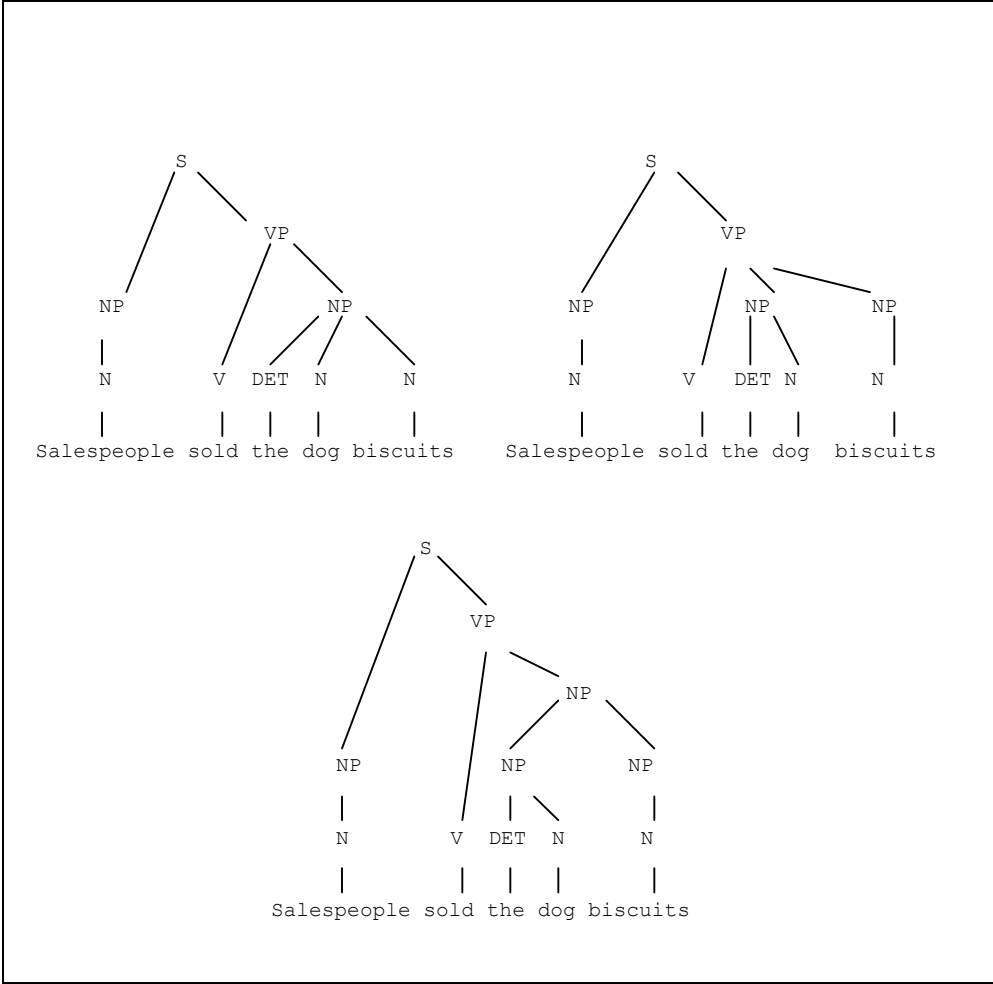


Figure 11: Three syntactic analyses for the sentence "Salespeople sold the dog biscuits."

Input:

we'll play with Sandy later .

Ambiguous morphosyntactic analysis (ambiguous POS):

pro|we~v:aux|will v|play^n|play prep|with n:prop|Sandy
adv|later^adj|late-CP .

Disambiguated morphosyntactic analysis (disambiguated POS):

pro|we~v:aux|will v|play prep|with n:prop|Sandy adv|later .

Pass 1 (using disambiguated POS):

```
((COUNTER 1)
(MOOD *DECLARATIVE)
(AUX ((MODAL +) (ROOT *WILL) (CAT V-AUX) (INDEX 2)))
(SUBJ ((CAT PRO) (ROOT *WE) (INDEX 1)))
(ADJUNCT (*MULTIPLE* ((CAT ADV) (ROOT *LATER) (INDEX 6))
              ((PP-OBJ ((CAT N-PROP) (ROOT *SANDY) (INDEX 5)))
                        (CAT PREP) (ROOT *WITH) (INDEX 4))))
(SUBCAT *SUBJ) (CAT V) (VFORM INF) (ROOT *PLAY) (INDEX 3))
```

Figure 12: The input sentence is correctly analyzed in the first pass, and no further passes are performed.

```

Input:
    you want a cookie ?

Ambiguous morphosyntactic analysis (ambiguous POS):
    pro|you v|want det|a n|cookie ?

Disambiguated morphosyntactic analysis (disambiguated POS):
    pro|you v|want det|a n|cookie ?

Pass 1 (using disambiguated POS):
    (Parse failed)

Pass 2 (using ambiguous POS):
    (Parse failed)

Pass 3 (allowing insertion of auxiliary):
    ((COUNTER 1)
     (MOOD *INTERROGATIVE)
     (AUX ((DUMMY +)))
     (SUBJ ((CAT PRO) (ROOT *YOU) (INDEX 1)))
     (OBJECT ((DET ((CAT DET) (ROOT *A) (INDEX 3)))
              (CAT N) (ROOT *COOKIE) (INDEX 4)))
     (SUBCAT *SUBJ-OBJ) (ROOT *WANT) (CAT V) (INDEX 2))

```

Figure 13: The input sentence is a question with a missing auxiliary, not covered by the grammar. Parsing fails in the first and second passes. A successful analysis is obtained in the third pass, with the insertion of an auxiliary.

Pass	POS Ambiguity	Insertion	Skipping
1	None	None	None
2	Limited	None	None
3	Limited	Auxiliary	None
4	Limited	NP	None
5	Limited	Auxiliary and NP	None
6	Limited	None	One word

Table 1: Coverage and ambiguity settings for different passes of parsing

Correct Analyses	
Pass 1 (unambiguous POS, no robustness)	115
Pass 2 (ambiguous POS)	29
Pass 3 (insertion of AUX)	3
Pass 4 (insertion of NP)	2
Pass 5 (insertion of AUX and NP)	4
Pass 6 (one word skipping)	4
Total	157 (78.5%)

Table 2: Contribution of each pass to correct analyses

Incorrect Analyses		
Lack of grammar coverage	Insertion	7
	Over-generation	5
POS tag error		4
Transcription error		1
Total		17 (8.5%)

Table 3: Causes of errors in incorrect analyses

No Analysis found	
Lack of grammar coverage	19
Lack of knowledge	5
Transcription error	1
Ungrammatical utterance	1
Total	26 (13%)

Table 4: Causes of parsing failures