A Dependency Treebank for English

Owen Rambow*, Cassandre Creswell*, Rachel Szekely†, Harriet Taber†, Marilyn Walker♠
rambow@unagi.upenn.edu

*University of Pennsylvania
Philadelphia, PA, USA
†The Graduate Center, The City University of New York
New York, NY, USA
♠AT&T Labs – Research
Florham Park, NJ, USA

Abstract

This paper presents the syntactic annotation level of a project aimed at providing a small dialog corpus with multiple levels of annotation. The syntactic annotation is based on dependency syntax. We outline the reasons for choosing dependency, and show the syntactic annotation for some constructions. We finish by describing the current state of the project.

1. Introduction

As part of the ISLE project, we have undertaken to richly annotate a small corpus of dialogs with multiple layers of annotation (orthography, intonation, syntax, NP coreference, information structure, dialog structure, discourse structure).[1] In this paper, we report on the syntactic annotation. The initial corpus consists of 13,000 words of dialogs between a human travel agent and about 30 different human customers, collected and transcribed at Carnegie-Mellon University as part of the DARPA Communicator project. Since we are interested in developing a model of the travel agent, we have only annotated her side of the dialog. While the speech is spontaneous and shows the usual signs of spontaneous speech (in particular disfluencies), the genre is goal-directed and professional, and it is very familiar to the travel agent; therefore, the disfluencies are limited.

We determined to annotate syntax independently of all other layers of annotation, starting from the speech transcription. Following the example of the Prague Dependency Treebank (PDT) for Czech (Böhmová et al., 2001; Hajic et al., 2001), we chose a dependency annotation rather than phrase-structure.

We justify our choice of dependency annotation in more detail in Section 2.1.. In the remainder of Section 2., we present the basic ideas of our annotation scheme. In Section 4., we discuss some more syntactic constructions in detail. We present the annotation procedure in Section 5., and discuss the current state of the project in Section 6.. We conclude with a discussion of future work.

2. Syntactic Annotation

2.1. Dependency Rather than Phrase-Structure

In designing the annotation, we aimed for a simple representation that is useful for any syntactic study of this corpus, independently of the application (such as parsing or generation). While corpus-based approaches to parsing do not pose any specific constraints on the syntactic representation used (as long as it is learnable), corpus-based approaches to generation do, since not all syntactic representations are suitable as a starting point for generation. In particular, the syntactic representation must be sufficiently "deep": a phrase structure representation already encodes word order[2] and as such is not useful as the starting point for generation. In corpus-based work on generation that uses the Penn Tree Bank (PTB), such as (Bangalore and Rambow, 2000), it is necessary to derive more abstract levels of representation from the annotation using heuristics. We would like to avoid the need for heuristic in future uses of this corpus.

The level of representation that many, if not most, applications actually need is a representation of lexical predicate-argument structure, where the lexical predicates are augmented with information such as tense, aspect or definiteness. This representation provides a useful interface to a (lexical-)semantic or conceptual representation. Many different syntactic theories provide such a level of representation, such as the f-structure of LFG (Kaplan and Bresnan, 1982), the SUBCAT of some versions of HPSG (Pollard and Sag, 1994), the Deep-Syntactic Representation of MTT (Mel'čuk, 1988), the Tectogrammatical Representation of FGD (Sgall et al., 1986), or (in a restricted sense) the derivation tree of TAG (Joshi and Schabes, 1991). Though of course all of these levels of representation differ in important linguistic aspects, they have in common that they do not use phrase structure, and instead can be represented as a dependency structure, i.e., a tree in which the nodes are labeled with the lexical predicates.[3]

[1]The work reported in this paper was funded by an ISLE grant to the University of Pennsylvania.

[2]Note that this is true even if we assume a phrase-structure representation which does not assume that nodes are ordered: the phrase structure imposes a restriction on word order through the phrase grouping. In particular, the interesting choices in languages such as English, for example topicalization or dative-shift, must have already been made in a phrase-structure representation. Of course, if the phrase-structure representation is fully flat, these remarks do not apply, but then it is equivalent to a dependency representation.

[3]We leave aside the issue of coindexation of arguments in

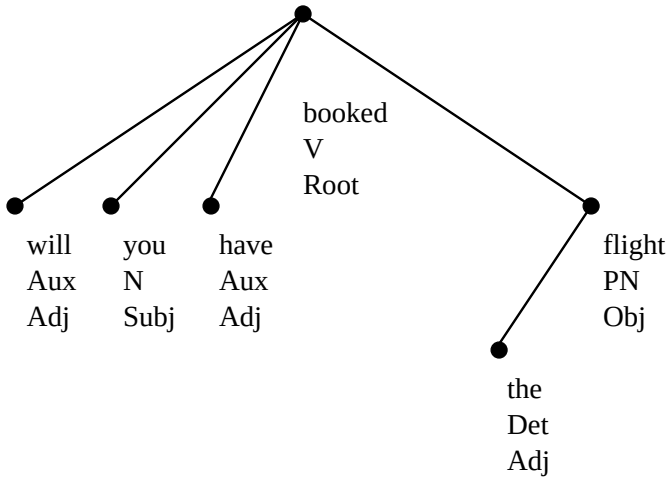

Figure 1: Auxiliaries as dependents: analysis for *will you have booked the flight*. The first line is the word, the second the part-of-speech, the third shows the surface role (SRole). In this example, the deep role (DRole) is always the same as the surface role.

Note that the second annotation standard for the Penn Treebank (PTB2) (Marcus et al., 1994) also includes a predicate-argument structure. However, the annotation is superimposed on the phrase structure which is the core of the PTB (Marcus et al., 1993). Many applications, including most recent parsers based on the PTB such as (Magerman, 1995; Collins, 1997), continue to use heuristics-based translation schemes from the phrase structure to lexical dependency ("head percolation table"). The use of such heuristics suggests that PTB2 annotation did not meet the needs of applications. (Note also that while parsing *a priori* does not place any condition on the syntactic annotation, the predicate-argument structure is useful in improving performance.) We would like to avoid the need for heuristics in using the corpus.

The question arises whether phrase structure is necessary at all in a corpus, and whether its absence does not represent a constraint on its usefulness. We observe that on its own, the fact that the standard for parser evaluation is based on phrase structure is hardly a good reason for advocating the continued use of phrase structure in corpora. However, if phrase structure is needed for a particular project or application, it is possible to derive the needed customized phrase structure from the dependency (lexical predicate-argument) structure along with the surface word order. In fact, it can be seen as the object of the study of syntax to establish such a mapping. Different theories will use different formal (or informal) means for defining this mapping and will define different mappings; LFG and TAG can be seen as proposing very explicit (though different) ways of relating phrase structure and lexical predicate-argument structure. Xia and Palmer (2001) discuss several algorithms for deriving phrase structure form dependency, embodying different theories of phrase structure, and compare them to the theory used (rather implicitly) in the PTB.

cases such as control, as it is done in LFG and other formalisms.

2.2. An Augmented Monostratal Analysis

While we follow the Prague Dependency Treebank (PDT) in using dependency annotation, we use a single level of annotation, and we do not use a separate level of representation for surface syntax. (In this respect we also differ from other dependency-based linguistic theories, such as MTT.) The main reason is that we wish to make the annotation process easier, and because the relevant surface syntax can be retrieved from the underlying syntax, the word order, and some additional information. (In the PDT, the tectogrammatical or deep level of annotation is partially derived automatically, and some of the remaining manual tasks – in particular the inclusion of empty argument nodes – we propose to perform as part of our monostratal annotation. So we are not suggesting that using a monostratal level will necessarily take us only half the time a bistratal annotation would.)

The annotation is a direct representation of lexical predicate-argument structure. Arguments and adjuncts are dependents of their predicate. We attach all function words to their lexical heads. For example, auxiliaries are dependents on the main lexical verb, rather than the inverse (see Figure 1). As a result, we do not show morphological subject-verb agreement as happening between two nodes which are in a direct dependency relationship; however, it is a simple matter to match up a surface subject and the finite auxiliary of the same verb, and thus the representation can easily be transformed into a more surface-oriented one. (Though this step is a bit more complex in subject-to-subject raising cases, see Section 4.2..) In Section 4. we discuss some constructions in more detail.

In addition, in cases in which argument structure has been changed (by passivization, or by use of expletive subjects as in *there*-constructions) we annotate for separate deep- and surface-syntactic roles. Redundantly, we also annotate for the process that resulted in the divergence between deep- and surface-syntactic role. For example, in a passive, the surface subject has surface role SUBJECT, but

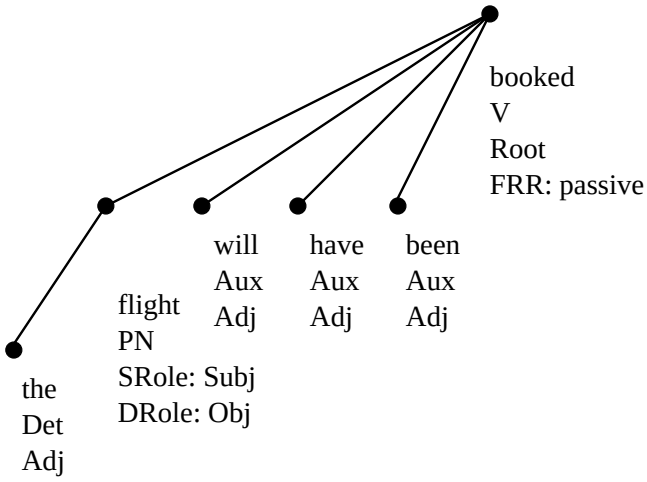

Figure 2: Passive voice example: analysis for *the flight will have been booked*. The first line is the word, the second the part-of-speech, the third shows the surface role (SRole), except for the word *flight*, for which the deep and surface roles are both given explicitly. In all other nodes of this example, the deep role is always the same as the surface role. The root node is marked as passive in the FRR feature; this feature is not set for all other nodes.

deep role OBJECT. The verb which has undergone passivization is so annotated directly. See Figure 2.

We do not explicitly annotate "movement" – the displacement of a node from its usual position with respect to its governor. Instead, we annotate the dependency arc as if the element were in place, even if this results in a non-projective structure. In the case of *wh*-movement, there is no need for the use of traces or a similar mechanism if both dependency and word-order information is available. In the case of subject-to-subject raising, the surface structure is a little more complex to retrieve, since it requires the identification of the matrix verb as a raising verb. However, this is signaled by the absence of a surface subject (and the moved embedded subject). Finally, recall that we analyze passivization lexically and not as a movement-type phenomenon (unlike the PTB, for example).

2.3. Morpho-Syntactic Features

Each node is annotated with the following morpho-syntactic features:

- The **wordform**.
- The **root** form of the word.
- The **POS** according to a simplified tag set that identifies the lexical category (noun, verb, adjective, adverb, preposition, conjunction, determiner, auxiliary) and several other items (punctuation, symbol, speech disfluency, miscellaneous). Note that unlike the PTB tag set, we distinguish verbs from auxiliaries, and we do not lump all forms of *to* to one category.
- The morphological **Features** determine the inflected form, given the root. Note that given the **POS** and **Features** features, the PTB POS can be determined.
- **SRole**. This is the surface-syntactic role, as determined by agreement and word order. Possible values are Subj, Obj, Obj2, PObj, PObj2, Adj, and Root. "Adj" covers all adjuncts, and also the relation between a function word and its governor. Of course, this relation is rather different linguistically from a true adjunct relation, but the distinction can be easily made based on the daughter node's POS feature.
- **DRole**. This refers to the deep-syntactic role and is only filled in if it differs from the surface syntactic role.
- **FRR** or functional relation reassignment. This must be filled in if the deep and surface roles differ. Possible values are None, Pass(ive), Erg(ative), Pred(icative), and There. "Pred" is given to adjectives, nouns and prepositions which are used predicatively (i.e., which acquire a subject), while "There" refers to the *there*-construction.

In addition, we have two fields, **Comment** and **Check** which allow the annotators to leave notes and to mark their progress.

3. Annotating a Speech-Based Corpus

The annotators work directly off the transcribed speech and do not have access to the speech files. The transcription has cleaned up speech disfluencies (including filled pauses and other non-meaningful sounds). The transcription, however, has added only minimal punctuation – mainly turn-final periods or question marks. For example, here is one turn with all the punctuation the annotators see:

> One thirty pm okay take one second it's departing Seattle at one thirty pm you'll arrive into Pittsburgh at eight fifty nine at night and that's US Airways flight one one five and for the rental car the least expensive I have with budget would be one sixty four for the entire time okay did you need a hotel .

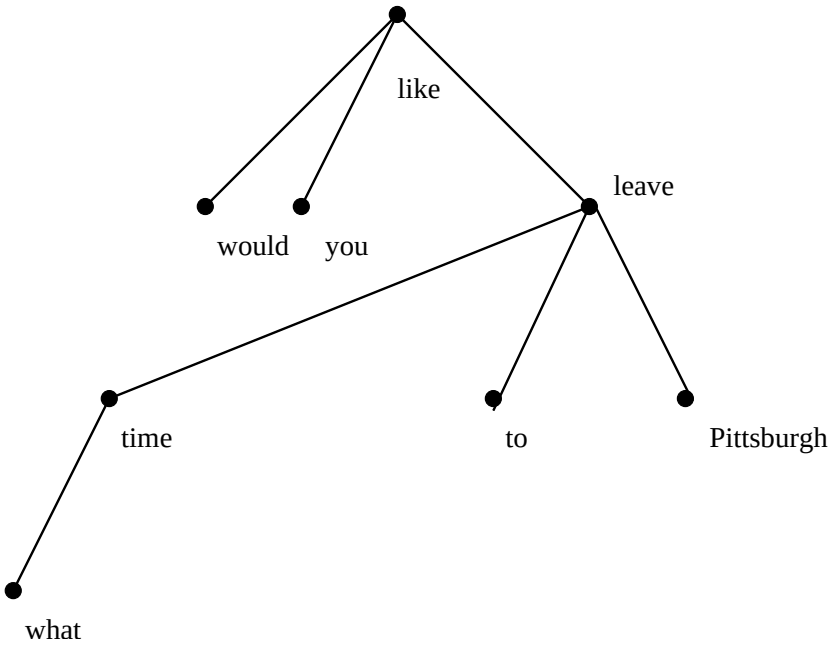

Figure 3: *Wh*-movement: analysis for *what time would you like to leave Pittsburgh.*

The annotators do not add punctuation, and the separate clauses are simply added as adjuncts to the preceding clause's main verb. While this approach does not correspond to standard orthographic conventions, we have chosen it in order not to have to make any assumptions about the role of punctuation. In particular, we wanted to avoid a situation in which punctuation was considered a reflection of intonation (for example, a period reflecting a deeper fall followed by a longer pause, while a comma is used when the pause is shorter). We believe it is more useful to supply the actual intonational features as well and thus allow for different conclusions, given the syntax and the intonation.

4. Some More Constructions

4.1. Long-Distance *wh*-Movement

We do not annotate *wh*-movement, be it local or long distance. The *wh*-constituent is simply dependent on its deep governor. In case of long-distance movement, this creates non-projective constructions, which is not a particular problem, either conceptually or for the Graph tool. An example is shown in Figure 3.

The most common projective surface-syntactic analysis involves reattaching the moved constituent to the matrix finite auxiliary, which, in a matrix *wh*-question, must exist. It is clear that it is easy to identify the matrix finite auxiliary. In cases of embedded long distance *wh*-questions, the preferred projective analysis is less appealing (attachment to the finite main verb to the left of the subject, as in *I know what airline you want to fly*). If we want to derive such an analysis, we can easily identify the verb to which we need to attach the *wh*-constituent positionally: it is the first finite verb to its right which is an ancestor.

4.2. Subject-to-Subject Raising

As in long-distance *wh*-movement, we annotate as if the movement had not occurred, and we attach the subject to the lower verb (of whose predicate it is an argument). This is shown in Figure 4. As a result of this annotation, the matrix verb *seem* has no subject – it is generally assumed to be a single-argument predicate, its clausal complement, with perhaps a prepositional object as well (*it seems to us that...*). Note that as in the case of *wh*-movement, there is no direct dependency link between the surface subject *it* and the matrix verb *seems* (which has no subject). It is unclear whether a bistratal representation can adequately capture the complexities of these constructions, since different syntactic processes can interact. In *What do you think seems to John to have been done*, the *what* can be argued to have surface-syntactic relations with *do* (licensing the auxiliary), *seems* (agreement), and perhaps *have* (or *to*), in addition to being the deep-syntactic direct object of *done*. This kind of interaction of passive, raising, and *wh*-movement can be captured in a movement-and-trace-based analysis as done in the PTB; we adopt essentially the same approach, but without (explicit) traces.

4.3. Control

In control verbs, we add the empty subject of the embedded verb ("PRO" in the Chomskyan theories) as a new node, whose word is the empty word *e*. Since it is an empty word, its location does not matter – in Figure 5, it is shown in the same position as the matrix verb, but it does not matter. We do not add the coindexation at this stage – we intend to add all anaphoric links, both intra- and intersentential in a separate annotation stage with specialized tools.

4.4. VP-Ellipsis

Our choice of always having the main verb as head of a clause poses a problem in cases of VP-ellipsis, in which the main verb and its non-subject arguments have been elided. For example, in *Does the hotel have a pool? Yes, it does [have a pool]* we can also elide *[have a pool]* and obtain

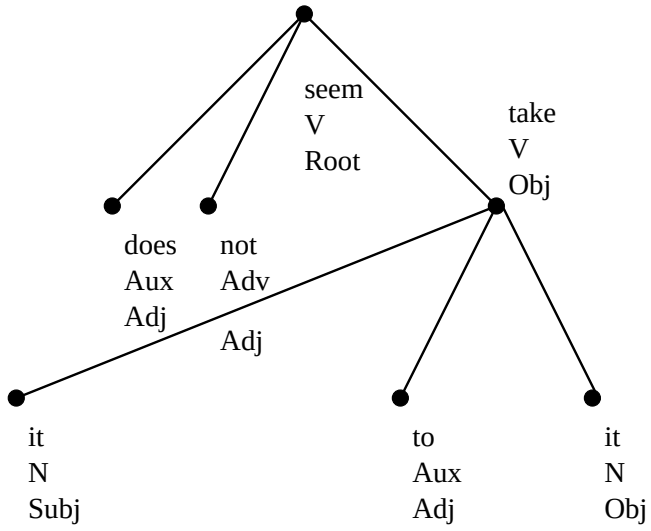

Figure 4: Raising: analysis for *it does not seem to take it*

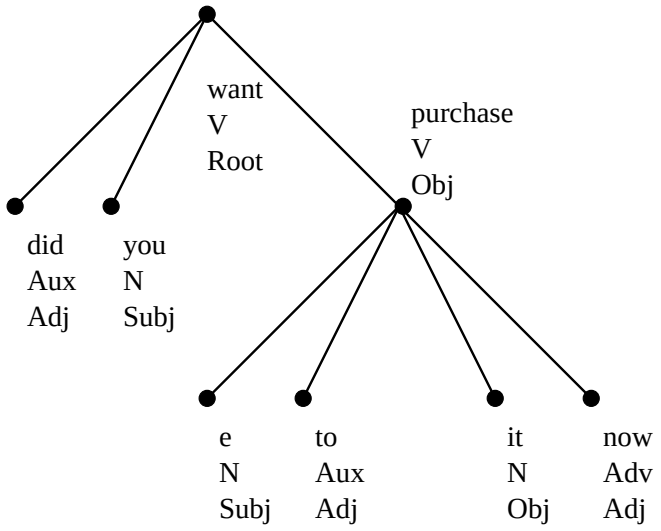

Figure 5: Control: analysis for *did you want to purchase it now*

Yes, it does. To solve the problem, we postulate an empty main verb which represents the locus of the anaphor. An example is shown in Figure 6.

5. Annotation Procedure

The procedure for annotation is as follows.

- There is an on-line manual. We also trained the annotators. with a 4-hour in-person session. The first dialog (about 250 words) was a training annotation with copious feedback.
- The dialogs are parsed with a dependency parser, the Supertagger and Lightweight Dependency Analyzer of Bangalore and Joshi (1999). This parser was trained on the Wall Street Journal and the language is quite different (questions, *wh*-words, disfluencies), so the quality of the parses is sometimes fairly low.
- The annotators correct the output of the parser using GRAPH, a graphical environment developed by the PDT effort and kindly made available to us by the PDT group. The structural analysis is annotated by drag-and-drop.
- In addition, the annotators correct the features using a pop-up menu. For those features that have a small range of possible values, the value is chosen by clicking. There is also a comment field to let the annotators make comments which are useful for the manual development. The feature values have also been filled in by the parser (and its morphological analyzer); this performs well in particular on the morphological features.
- Finally, a small program automatically checks for inconsistencies (e.g., if surface and deep roles differ, then there must be an explanation in the form of a value for the FRR feature; a preposition always needs

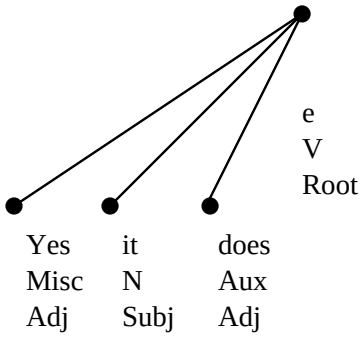

Figure 6: VP-ellipsis: analysis for *Yes, it does* in response to a question such as *Does the hotel have a pool?*.

an object; and so on). The output of this program lets the annotators catch small errors.

6. Results so Far and Annotator Agreement

So far, we have annotated nearly a third of the corpus. After the initial training phase, the annotation rate has been about 30 words an hour. We hope that this rate will further improve as the annotators become more familiar with the annotation scheme; furthermore, we intend to train the parser on the initial annotations, so that the quality of the parsed corpus will increase (and thus the work required from the annotators decrease).

We performed an inter-annotator agreement analysis on one dialog with 272 words (not counting punctuation). Since we added punctuation, there were a total of 272 dependency arcs. Two annotators disagreed on 16 of those (for an agreement of 94%), 3 of these disagreements related to the attachment of the sentence-final period. The differences in dependency analysis relate to compound nouns (*US Air*), traditional PP attachment ambiguities (*US Air has a non-stop flight at 7 pm* and *a preference for a price range on that*), adverbial phrase attachment ambiguity (*[earlier than that] on US Air the only flight before that on US Air is at twelve twenty*), an erroneous analysis of the set phrase *you're welcome*, and three occurrences of an erroneous low attachment of a determiner in a compound noun. (Note that some disagreements can lead to more than one differing dependency arcs.) Apart from the erroneous analyses, it seems that the disagreements are not easily resolvable and represent true problems of analysis in this domain. We attempt to achieve consistency by developing heuristics for cases of ambiguity (such as low attachment).

In addition, we have seen that in choosing features, the POS and Feats features achieve agreement of 93–94%, SRole 95% (mainly errors of oversight, and disagreement resulting from differences in dependency structures), while all other features achieve agreement at or above 99%.

7. Conclusion

In conclusion, we think that dependency is a good way to annotate English syntax, since the more abstract representation (as compared to phrase structure) makes the task of the annotators and of the designers of the manual easier, and hopefully will allow for a broader use of the corpus, without the need to use heuristics to get at the information that is really needed. Any other representation can be derived from it as needed. We also think that the annotation of small, domain-specific corpora is useful as the PTB's restricted genre limits its usefulness in completely different domains. In future work, we hope to show that using even a syntactically annotated in-domain corpus is beneficial.

In future work, we will add other levels of annotation to the same corpus. We will definitely annotate for NP coreference, and intend to also annotate some form of information-status information such as topic-focus articulation.

8. References

Srinivas Bangalore and Aravind Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–266.

Srinivas Bangalore and Owen Rambow. 2000. Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2001. The Prague Dependency Treebank: Three-Level Annotation Scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, July.

Jan Hajic, Eva Hajicová, Martin Holub, Petr Pajas, Petr Sgall, Barbora Vidová-Hladká, and Veronika ezníková. 2001. The current status of the prague dependency treebank. In *LNAI 2166*, LNAI 2166, pages 11–20. Springer Verlag, Berlin, Heidelberg, New York.

Aravind K. Joshi and Yves Schabes. 1991. Tree-adjoining grammars and lexicalized grammars. In Maurice Nivat and Andreas Podelski, editors, *Definability and Recognizability of Sets of Trees*. Elsevier.

Ronald M. Kaplan and Joan W. Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In J. W. Bresnan, editor, *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, Mass., December.

David Magerman. 1995. Statistical decision-tree models for parsing. In *33rd Meeting of the Association for Computational Linguistics (ACL'95)*.

Mitchell M. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19.2:313–330, June.

M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*.

Igor A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, New York.

Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

P. Sgall, E. Hajicova, and J. Panevova. 1986. *The meaning of the sentence and its semantic and pragmatic aspects*. Reidel, Dordrecht.

Fei Xia and Martha Palmer. 2001. Converting dependency structure to phrase structures. In *hlt2001*, pages 61–65.