Penn Discourse Treebank: Building a Large Scale Annotated Corpus Encoding DLTAG-based Discourse Structure and Discourse Relations

Cassandre Creswell, Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi
Institute of Research in Cognitive Science
University of Pennsylvania
creswell@babel.ling.upenn.edu
forbesk@unagi.cis.upenn.edu
elenimi@linc.cis.upenn.edu
rjprasad@linc.cis.upenn.edu
joshi@linc.cis.upenn.edu

Bonnie Webber
School of Informatics
University of Edinburgh
bonnie@inf.ed.ac.uk

Abstract

Large scale annotated corpora have played a critical role in speech and natural language research. However, while existing annotated corpora such as the Penn Treebank have been highly successful at the sentence-level, we also need large-scale annotated resources that reliably encode key aspects of discourse. In this paper, we detail (1) our plans for building the Penn Discourse Treebank (PDTB), (2) our preliminary annotation work, and (3) the results to date of our efforts. Annotation in the PDTB will focus on coherence relations associated with discourse connectives, including their argument structure and anaphoric links, thus exposing a clearly defined level of discourse structure and supporting the extraction of a range of inferences associated with discourse connectives.

1 Introduction

Large annotated corpora have played a critical role in speech and natural language research, enabling the derivation of statistical information and its integration with linguistic knowledge. This has led to both scientific and technological advances, for example in robust parsing and automatic extraction of relations and coreference, and in their applications to information extraction, question answering, summarization, and machine translation. The Penn Treebank (Marcus et al., 1993) is an outstanding example of such a resource with worldwide impact on natural language processing. However, while the Penn Treebank has been highly successful, it only contains sentence-level annotation. We also need significant annotated resources that encode key aspects of discourse. Although there have been some attempts to construct resources encoding discourse related information, they have been small scale and they have not been integrated or linked with corpora annotated at the sentence-level.

In this paper, we present details of our plans for building the Penn Discourse Treebank (PDTB) as a large corpus in which coherence relations associated with discourse connectives have been reliably annotated, including their argument structure and anaphoric links. The annotation is grounded in the framework for a Lexicalized Tree-Adjoining Grammar for Discourse (DLTAG) (Webber et al., 1999; Webber et al., to appear). The annotated corpus is aimed at exposing a clearly defined level of discourse structure and supporting the extraction of a range of inferences associated with discourse connectives.

The paper is organized as follows. In Section 2, we describe the DLTAG framework on which the annotations are based. In Section 3, we discuss the general goals of the PDTB project, focussing on the aspects of discourse structure and semantics that we will annotate the corpus with. Section 4 describes some of the applications for which the PDTB will prove useful, especially when con-

sidered together with other annotations carried out independently on the same corpus. In Section 5, we describe an initial implementation of a "shallow" discourse parsing tool that can be used during the annotation process to obtain a discourse parse. Section 6 describes an experiment with annotating connectives and their arguments. In Section 7, we describe our efforts in annotating lexico-syntactic indicators of semantic features of the arguments of anaphoric connectives, to support the development of resolution algorithms. Finally, we discuss related work on annotating discourse information in corpora in Section 8 before concluding in Section 9.

2 The Framework: A Lexicalized Tree-Adjoining Grammar for Discourse

In a Lexicalized Tree-Adjoining Grammar (TAG) for discourse (Webber et al., 1999; Webber et al., to appear), discourse structure is regarded as being created by a composition of elementary trees anchored by discourse connectives. Analogous to predicates in the sentence-level grammar, a DLTAG posits that discourse connectives are predicates at the discourse level. Arguments of these discourse connectives are clauses which can be realized *structurally* or *anaphorically*. An LTAG contains two kinds of elementary trees: *initial* trees, which encode basic predicate-argument dependencies, and *auxiliary* trees, which introduce recursion, allowing for modification and/or elaboration of the elementary trees. All structural composition is achieved with two operations, *substitution* and *adjunction*. In DLTAG, clauses connected by a subordinating conjunction form an initial tree whose compositional semantics is determined by the semantic requirements of the subordinate conjunction (the predicate) on its arguments (the clauses). Auxiliary trees are used for providing further information through adjunction. They can be anchored by adverbials, by conjunctions like *and*, or may have no lexical realization. Furthermore, a discourse predicate may take all its arguments structurally, as in the case of subordinating conjunctions, or anaphorically, by making use of events or situations available from the previous discourse, as in the case of *then*.

3 The Penn Discourse Treebank

Our discourse-level annotation is produced as stand-off annotation, which will link to the Penn Treebank syntactic annotations as well as the predicate-argument annotations, called the Proposition Bank or PropBank (Kingsbury and Palmer, 2002).

Annotated in the PDTB are the three major types of explicit discourse connectives given in (A)-(C) below. A fourth type consists of the empty connective, described in (D), reflecting the fact that although all (or almost all) sentences in a text contribute to its discourse structure, not all sentences will have their structural connection to the text realized explicitly as a lexical connective.

(A) **Subordinate conjunctions:** Subordinate conjunctions form a uniform class of predicates. Each connective has two structural arguments, not necessarily in a fixed order. For example, the two arguments of *because* can appear in different orders – (1) versus (2):

(1) John failed the exam *because* he was lazy.

(2) *Because* he was lazy, John failed the exam.

(B) **Coordinate conjunctions:** Like subordinate conjunctions, coordinate conjunctions, such as *and*, take their arguments structurally, and contribute to a compositional derivation of the interpretation of the units joined by the conjunction.

(C) **Adverbial connectives:** Adverbial connectives also take two arguments. However, only one of these comes structurally, from the clause containing the adverbial. The second argument of these connectives is identified anaphorically. For example, in (3) below, the discourse adverbial *then* in (3d) takes one structural argument, *he found out that he was broke*, and one anaphoric argument, resolved to *he ordered two cases of the '97*. Adverbial connectives may appear sentence initially, medially or finally. The position of the adverbial connective in the sentence affects the scope of the connective and is often

associated with the information structure of the sentence.

(3) a. On the one hand, John loves Barolo.
b. So he ordered 3 cases of the '97.
c. On the other hand, he had to cancel the order
d. because he *then* found out that he was broke.

(D) **Empty connectives:** Empty connectives are predicates which realize a coherence relation but do not appear lexically in the discourse. In (4) below, (4a) and (4b) are connected via a causal relation in the absence of a lexical connective. The DLTAG framework is unique in providing structural descriptions for these covert discourse connectives. In DLTAG, the two clauses are connected at the discourse level by a tree with a null anchor. This structural description will prove very useful to the annotators who will be able to annotate the arguments of the null predicate on the DLTAG parse output.

(4) a. You should not lend John any books.
b. He never returns them.

Initially we start with a set of tags corresponding to (1) the types of connectives (structural, anaphoric and null) and their positions (initial, medial, and final), and (2) the positions of the arguments of the connective. An argument of a connective can be, for example, an embedded clause, the preceding sentence, or the immediately preceding discourse. Eventually, the annotated arguments will also be linked to the appropriate segments in the PTB and PropBank within the standoff annotation architecture. We will also annotate certain additional kinds of semantic information associated with connectives, as well as lexico-syntactic information that provides evidence for semantics (Section 7). Specification of the frames associated with the argument structures of connectives, including anaphoric links, will also help the annotators during the annotation process by letting them judge quickly and accurately the relevant roles played by the surrounding context for each connective in a discourse corpus, enabling them to distinguish the arguments for a connective from the surrounding clauses. This is very similar to providing argument-adjunct frames for a verb to the annotators of PropBank.

4 Applications

PTB has been widely used in natural language processing with applications ranging from parsing, information extraction, question-answering, summarization, machine translation, generation systems, as well as in corpus based studies in linguistics and psycholinguistics. The PropBank, which is under construction and which is integrated with Penn Treebank, is expected to have at least the same range of applications, with the quality and coverage of the results going well beyond what has been so far possible with the use of Penn Treebank alone. This is due to the added "knowledge" incorporated in the PropBank. Since the PDTB will provide a substantial level of discourse structure information, together with Penn Treebank and PropBank, PDTB will raise the bar very substantially with respect to the quality and coverage achieved in the applications mentioned above.

A simple example will make our point. At present one technique in question answering (QA) involves the QA system returning the sentences R from one or more documents that express the same predicate-argument relations as the query. Systems may also return other sentences which mention entities that are coreferential with those in R. With smoothing, this collection of sentences can serve as a useful response to the initial query. The PDTB will allow an additional set of relevant sentences to be retrieved – those which are arguments of connectives in R including the anaphoric links associated with connectives, thus producing an even richer response to the query. It might even be possible to use the same architecture as the one developed for entity coreference in a question-answer system (Morton, 2000).

5 Parsing Discourse

(Forbes et al., 2001) presents a preliminary implementation of a discourse parsing system for a lexicalized Tree-Ajoining Grammar for discourse, specifying the integration of sentence and discourse level processing. The system is based on the assumption that the compositional aspects of semantics at the discourse-level parallel those at the sentence-level. (Just as at the sentence-level, inferential semantics and anaphor resolution must also be computed, but this is not done compositionally.) The integration of sentence and discourse processing is achieved through both the sentence and discourse grammars being based on LTAG, with the same parser working at both levels. The discourse parsing system therefore allows a smooth transition from the derivation of sentence structure to the derivation of discourse structure.

This preliminary implementation takes a discourse (such as a small segment of the Penn Treebank corpus) as input and first parses the sentences individually and then extracts clauses (i.e., the clausal derivations) from the sentence parses. Next, the same parser (Sarkar, 2000) treats the clauses as units (corresponding to the lexical anchors in LTAG) and parses the discourse with a DLTAG grammar whose elementary trees are anchored by the discourse connectives and whose arguments are the units corresponding to clauses. The lexicon associated with the discourse grammar includes all connectives and the type of tree(s) they anchor. Subordinate and coordinate conjunctions are associated with trees in which both arguments come structurally. Adverbial connectives are associated with trees that take one structural argument, with the other argument identified via an anaphoric link. For more detail on the architecture, see (Forbes et al., 2001).

As an example of a DLTAG derivation, Figure 1 shows the output of the system on the short extract from Section 21 of the Penn Treebank WSJ corpus, given in Example 5. (The discourse connectives in the text are shown in bold.)

(5) a. The pilots could play hardball by noting they are crucial to any sale or restructuring because they can refuse to fly the airplanes.

b. **If** they were to insist on a low bid of, say $200 a share, the board mighn't be able to obtain a higher offer from other bidders **because** banks might hesitate to finance a transaction the pilots oppose.

c. **Also**, **because** UAL chairman Stephen Wolf and other UAL executives have joined the pilots' bid, the board might be able to exclude him from its deliberations in order to be fair to other bidders.

To avoid the problem of getting too many sentential derivations for the long and complex sentences typically found in this corpus, we used the single derivations produced by LEXTRACT (Xia et al., 2000), which takes the Treebank and Treebank-specific information and produces derivation trees for the sentences annotated in the Treebank.

We plan to use the discourse parsing system during the annotation process to obtain a discourse parse for the Penn Treebank sentences. While the parse provided by the tool will most likely have errors, it will be easier for the annotators to edit this parse rather than to construct the entire discourse by hand. This is analogous to the use of "shallow parsing" in the construction of the Penn Treebank.

6 Location of Anaphoric Arguments: An Experiment

(Creswell et al., 2002) reports on an annotation study of nine connectives (*so*, *therefore*, *as a result*, *in addition*, *also*, *moreover*, *nevertheless*, *yet* and *whereas*). For each connective, seventy-five tokens (a total of 675 tokens) were extracted from a variety of corpora: Brown, Wall Street Journal, Switchboard and 58 transcribed oral histories from the online Social Security Administration Oral History Archives (SSA).[1] The 675 tokens were split into three groups and annotated by three annotators (225 tokens per annotator). Each token was annotated with tags that encoded information about (a) the connective's left argument (ARG), and (b) the clause containing the connective (CONN). Both ARG and CONN were annotated with a REF tag that encoded an ID number which was the same for both in a single token. ARG was further tagged with a TYPE tagset that identified the size of the argument. The tags under TYPE were: MAIN if the argument was contained

[1]The Brown, Wall Street Journal and Switchboard corpora are available from LDC, http://www.ldc.upenn.edu. The SSA corpus is available at http://www.ssa.gov/history/orallist.html

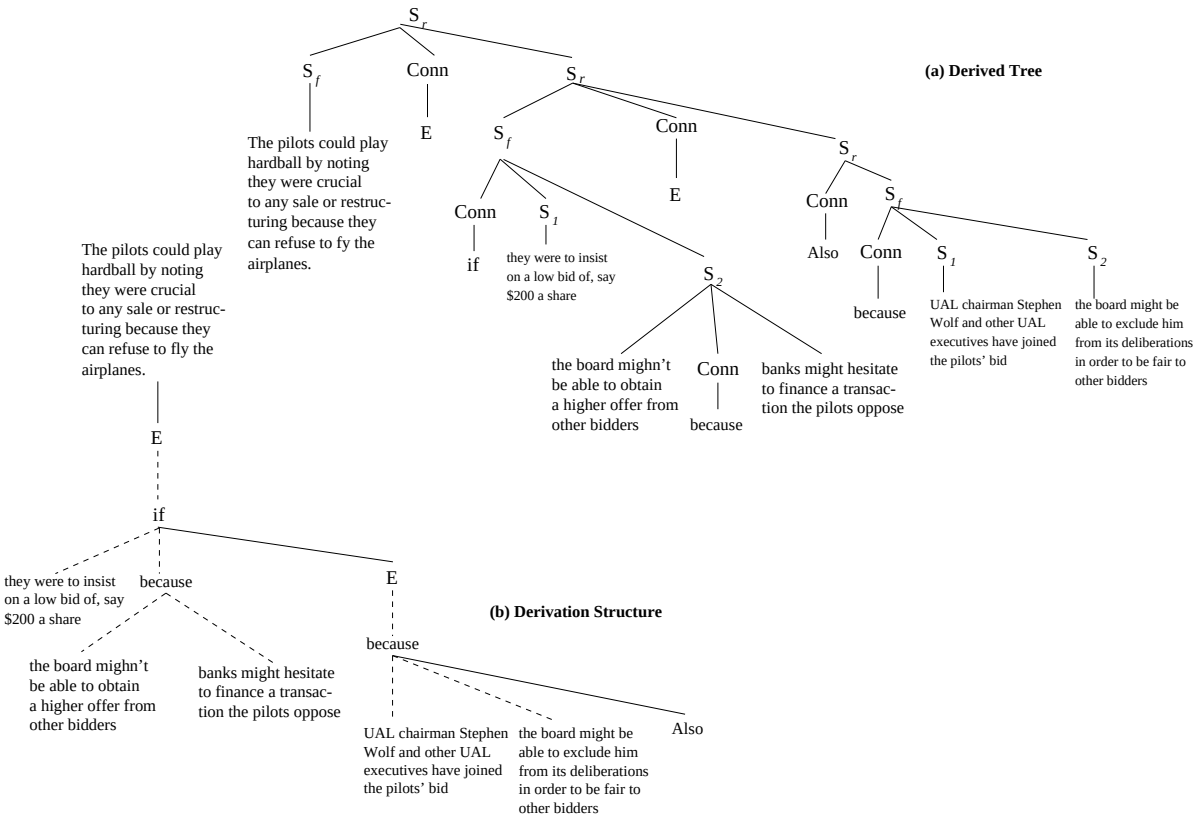

Figure 1: Derived Tree (a) and Derivation Structure (b) for WSJ Discourse in (5)

in a full sentence (including subordinate clauses); MAIN-MULT if the argument was contained in a sequence of sentences; SUB if the argument was contained in a subordinate clause; and XP if the argument was contained in a phrasal constituent. The ARG tagset will be useful in computing statistics on the type of antecedent of anaphoric connectives, and for formulating constraints for anaphora resolution.

The clause containing the connective, CONN, was annotated with two tagsets: COMB and POS. COMB was used to identify punctuation (PERIOD, COMMA, etc.), coordinating conjunctions (AND and BUT), and adverbial connectives (YET, SO, et.) that can co-occur with the connective. Information about co-occurrence with punctuation and other connectives will, among other things, be useful for determining the structural descriptions of individual connectives. POS was used to identify the position of the connective in its clause (INITIAL, MEDIAL, FINAL). The POS tagset will be useful for formulating constraints relevant to the *Information Structure* of the clause, which has been found to be relevant to anaphor resolution (Kruijff-Korbayová and Webber, 2001). The complete set of tags is given in Table 1. In subsequent discourse annotation of the Penn Treebank, the values of these features can be extracted automatically.

Preliminary results from this annotation show that for most of the connectives included in the study, there is a strong tendency for the left argument to be identified locally (in the structural sense) – either in the immediately preceding sentence or in immediately preceding sequence of sentences, in most cases the preceding paragraph. Most notably, *so* (which appears only in initial position) always takes a sentence or a sequence of sentences as its left argument, confirming its appropriate treatment as a structural connective. *In addition*, *yet*, *moreover*, *as a result* and *also*, tend to take their left argument locally but they demonstrate a larger syntactic variety of potential arguments such as subordinate clauses or phrasal

ARG		
	REF	ID #
	TYPE	MAIN= sentence MAIN-MULT= multiple sentences SUB = subordinate clause XP= phrasal constituent (NONE)= no left argument
CONN		
	REF	ID #
	COMB	PERIOD COMMA COLON SEMI-COLON DASH 'AND' 'BUT' CONN
	POS	INITIAL MEDIAL FINAL

Table 1: Annotation tagsets

constituents. *So*, *nevertheless* and *moreover* are likely to take larger discourse segments as arguments. Larger discourse segments appear to lead to vagueness in resolving anaphora. For example, it was often difficult to determine the extent of the left-hand argument of *nevertheless*, which could also be a phrasal intra-sentential constituent (XP). The connective *therefore* often takes its left-hand argument from a subordinate clause. Finally, we found it useful to make special tags for combinations with a complementiser (COMP) and a subordinate conjunction (SUB). *As a result*, for example, quite often appears in complement clauses, which creates ambiguity in the interpretation.

7 Semantic Properties of Anaphoric Arguments

The annotation reported in (Creswell et al., 2002) included only surface syntactic features. However, in order to understand what licenses the use of a particular anaphoric discourse connective and hence resolve it, we also need to characterize the semantic properties of its arguments and their lexico-syntactic realization. This is the subject of (Miltsakaki et al., 2003), which focuses on the adverbial connective *instead*.

Instead as a discourse connective conveys the fact that the interpretation of its matrix clause (its structural argument) is an alternative to something in the previous discourse that admits or invites alternatives (its anaphoric argument). We examined 100 successive instances of sentence initial *instead*, (a) identifying the text containing the source of the argument; (b) computing inter-annotator agreement; (c) annotating lexico-syntactic features that appeared to correlate with the existence of alternatives; and (d) quantifying the frequency of appearance of these features in the identified arguments. The set of features included clausal negation, presence of a monotone-decreasing quantifier on subject or object, presence of a modal auxiliary, and conditionality. We also kept a record of the verbs in the argument clause and in the higher clause for embedded arguments to enable characterization of predicates that give rise to alternatives.

For as many as 67% of the tokens, the anaphoric argument realized at least one of the negation/md-quantifier/modality/conditionality features. In an additional 27%, the semantics of the argument's main verb or the higher verb embedding the argument's main verb admits alternative situations or events (e.g., *expect*, *want*, *deny*, etc.). In sum, for a total of 94% of tokens, we were able to characterize features of the arguments that could be automatically extracted from existing annotations and used to help resolve these anaphoric arguments. In the remaining cases, the annotated features were absent, meaning that the set is incomplete.

Our annotated features can be automatically extracted from the Penn Treebank syntactic annotation. Thus, for a future expanded study, we will be able to layer our annotation of discourse arguments on top of the existing Penn Treebank corpus and then automatically extract the relevant features with minimal effort. In fact, this layered annotation on an existing public resource is crucial for any empirical investigation of theoretical linguistic issues as it facilitates considerably research on the contribution of lexico-syntactic properties of sentence and discourse level predicates to discourse meaning.

8 Related Work

Efforts to annotate discourse structure started as a way of providing empirical justification for high-level theories of discourse structure (Grosz and Sidner, 1986; Moser and Moore, 1996). Although much time and energy was devoted to the work

(Di Eugenio et al., 1998), the results have not been widely used, though remaining a resource for the future.

The work closest to our own is the resource developed by Marcu (Marcu, 2000) based on Rhetorical Structure Theory (RST). The principles of RST (Mann and Thompson, 1988) include: (1) that adjacent units of discourse are related by a single rhetorical relation that accounts for the semantic or pragmatic (intentional) sense associated with their adjacency; (2) that units so related form larger units that participate in rhetorical relations with units that they themselves are adjacent to; and (3) that in many, but not all, such juxtapositions, one of the units (the satellite) provides support for the other (the nucleus), which then appears to be the basis for rhetorical relations that the larger unit participates in.

Given these principles, the two main features of RST annotation are (1) demarcation of the elementary discourse units that participate in relations and (2) labeling of those relations. The two are not independent. For example, a relation (attribution) postulated between the specification of a speech act (e.g., *Riordan said*) and its content specifed as direct or indirect speech (e.g., *We must expand the vision of our party*) means that a subject-verb fragment must be marked as an elementary discourse unit if the object of the verb is direct or indirect speech.

Marcu's RST-annotated corpus differs from the current annotation effort in three main ways: First, the RST-annotated corpus does not indicate the basis for a rhetorical relation being annotated between two elementary or derived units. Even though there is a strictly ordered protocol to follow in assigning rhetorical relations, the corpus contains no record of either the particular basis on which a rule from the protocol such as *If the relation is one of Explanation, assign relation Explanation*, is taken to hold, or why the conditions for earlier rules in the protocol were taken to fail to hold.

In the PDTB, we have undertaken to annotate all and only the arguments of discourse connectives – adverbials, prepositional phrases and conjunctions. As such, the basis for each coherence relation is the higher-order predicate associated with the connective, to which the discourse units involved serve as arguments. (The precise semantic nature of that relation may be ambiguous - e.g., whether the relation conveyed by *then* is one of temporal ordering or logical consequence. But existing lexico-syntactic annotations and annotations of clausal predicate-argument relations currently in progress in the PropBank project (Kingsbury and Palmer, 2002) will provide a solid basis for disambiguation efforts.)

Secondly, the discourse relation holding between units has to be inferred, using semantic and pragmatic information, in cases where an overt connective is missing from the discourse. While the RST-annotated corpus records inferred relations, it omits any indication of what was used in inferring them. The PDTB annotation scheme takes two steps towards remedying this omission: (a) it is built on top of the DLTAG parse, which provides structural descriptions for the empty connectives and (b) the DLTAG parse links up to sentence-level syntactic and semantic annotation for each sentence. Identifying the empty connectives and accessing sentence-level syntactic and semantic information are crucial steps towards an automated inference of discourse relations in the absence of lexically realized connectives.

Finally, RST annotation of elementary discourse units, derived discourse units and rhetorical relations bear the entire burden of supporting language technology algorithms derived from the RST annotated corpus. The PDTB annotation effort will be an additional layer on top of text already annotated with syntactic structure (PTB) and predicate-argument relations (PropBank). These layers will be linked, and both their presence and their linkage will provide a richer substrate for the development and evaluation of practical algorithms.

We are not downplaying the importance of having an annotated corpus of coherence relations associated with adjacent discourse units. But we believe that the task of producing such a corpus can be made easier by having already identified the higher-order predicate-argument relations associated with explicit discourse connectives. They can then be factored into the calculation or removed from the calculation, as appropriate (Webber et al.,

to appear).

9 Conclusions

This paper has described practical and empirical work done in the basic framework of lexicalised discourse structure presented in (Webber et al., 1999; Webber et al., to appear), including our new NSF-supported effort to create a Discourse Treebank, as a resource for developing further applications of Language Technology.

Acknowledgements

This work has been supported in part by EPSRC Grant GR/M75129/01.

References

Cassandra Creswell, Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2002. The Discourse Anaphoric Properties of Connectives. In *Proceedings of the 4th Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2002)*, pages 45–50, Lisbon, Portugal.

Barbara Di Eugenio, Pamela W. Jordan, Johanna D. Moore, and Richmond H. Thomason. 1998. An Empirical Investigation of Proposals in Collaborative Dialogues. In *Proceedings of COLING/ACL'98*, pages 325–329, Montreal, Canada.

Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind Joshi, and Bonnie Webber. 2001. DLTAG system - Discourse Parsing with a Lexialized Tree Adjoining grammar. In *Proceedings of the ESSLLI-2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics*, pages 17–36, Helsinki, Finland.

Barbara Grosz and Candace Sidner. 1986. Attentions, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to Propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Canary Islands, Spain.

Ivana Kruijff-Korbayová and Bonnie Webber. 2001. Information Structure and the Semantics of 'otherwise'. In *Proceedings of ESSLLI 2001: Workshop on Information Structure, Discourse Structure and Discourse Semantics*, pages 61–78, Helsinki, Finland.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory. Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281.

Daniel Marcu. 2000. The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics*, 26(3):395–448.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Eleni Miltsakaki, Cassandre Creswell, Katherine Forbes, Aravind Joshi, and Bonnie Webber. 2003. Anaphoric Arguments of Discourse Connectives: Semantic Properties of Antecedents versus non-antecedents. In *Proceedings of the Workshop on the Computational Treatment of Anaphora: 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, Budapest, Hungary.

Thomas Morton. 2000. Coreference for NLP Applications. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong.

Megan Moser and Johanna Moore. 1996. Toward a Synthesis of Two Accounts of Discourse Structure. *Computational Linguistics*, 22(3):409–419.

Anoop Sarkar. 2000. Practical Experiments in Parsing Using Tree Adjoining Grammars. In *Proceedings of the Fifth Workshop on Tree-Adjoining Grammars, Paris, France, May 25–27*.

Bonnie Webber, Alistair Knott, Mathew Stone, and Aravind Joshi. 1999. Discourse Relations: A Structural and Presuppositional Account using Lexicalised TAG. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, College Park, MD.*, pages 41–48.

Bonnie Webber, Aravind Joshi, Mathew Stone, and Alistair Knott. to appear. Anaphora and Discourse Structure. *Computational Linguistics, 2003.*

Fei Xia, Martha Palmer, and Aravind Joshi. 2000. A Uniform Method of Grammar Extraction and its Applications. In *Proc. of the Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, Hong Kong.