Complexity of Dependencies in Discourse: Are Dependencies in Discourse More Complex than in Syntax?

Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh
University of Pennsylvania
aleewk,rjprasad,joshi,nikhild@seas.upenn.edu

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

Abstract

This paper investigates the complexity of dependencies at the discourse level, in particular the dependencies between discourse connectives and their arguments. Our study is based on data from the Penn Discourse Treebank (PDTB) and is therefore an exploration into the ways treebanks can inform linguistic issues. We observe that, unlike in syntax, there is more uncertainty and flexibility with regards to the location and extent of discourse arguments. This leads to a variety of possible patterns of dependencies between pairs of discourse relations, including nested, crossed and a range of other non-tree-like configurations. Nevertheless, our main conclusion is that the types of discourse dependencies are highly restricted since the more complex cases can be factored out by appealing to discourse notions like *anaphora* and *attribution*. We conjecture that the complexity of dependencies is far more restricted at the discourse level as compared to the syntactic level.

1 Introduction

In this paper, we present some preliminary results concerning the complexity of dependencies at the discourse level with respect to discourse connectives and their arguments. The basis of our study is data from the Penn Discourse Treebank (PDTB), a corpus of 1 million words (from the same text as the Penn Treebank corpus) which is being annotated for all discourse connectives (explicit and implicit) and their arguments. We will present some details of the PDTB corpus in Section 2, which will be adequate for the purpose of this paper.

The PDTB project has its origin in the idea that the machinery used at the syntactic level for describing dependencies could be carried over to the discourse level. Lexicalized Tree-Adjoining Grammar (LTAG) is a system for describing the dependencies between lexical anchors of elementary trees and their arguments, using two language independent composition operations of substitution and adjunction. D-LTAG is a lexicalized approach to discourse relations ([12, 3], for example). Lexicalization here means that each elementary tree in D-LTAG is anchored by a discourse connective (explicit or implicit) corresponding to discourse relations and has slots for arguments from other parts of the text.

LTAG, used at the sentence level, has been extensively studied. It characterizes nested as well as certain classes of crossed dependencies (projective and certain classes of non-projective dependencies, to use the terminology of dependency grammar representations), leading to the so-called mildly context-sensitive languages. Naturally, the following question arises: Is the complexity of dependencies at the level of discourse (with respect to discourse connectives and their arguments) at the same level as at the sentence level, as characterized by LTAG?

This question sets the stage for our investigations. Although the final release of the corpus has not taken place yet, by now there is quite a substantial amount of data from the corpus available in order to embark on a preliminary study. The PDTB project began with the D-LTAG representations in mind, as described in Section 2. However, the annotation guidelines were subsequently made as theory-independent as possible so that the corpus would be usable by a wide range of users. Further, the annotations are in terms of text spans and not necessarily coincident with phrases at the syntactic level.

Given this background, in Section 3 we discuss the various types of dependencies we have observed in the PDTB. A number of issues arise due to the fact that, although at the discourse level, the a-rity of predicates (discourse connectives) is two and exactly two, the extent and location of the arguments (i.e., the text spans corresponding to the arguments) is far more uncertain in contrast to predicates at the syntactic level, where the a-rities of the predicates are not necessarily fixed but the arguments are quite local and their extents are relatively fixed. Further, two discourse connectives can share an argument (a text span) either completely or partially, or the arguments of one connective can interleave with the arguments of another connective. These considerations lead to a variety of possible patterns of complex dependencies as described with examples from the corpus in Section 3.

However, after analyzing a range of cases from the corpus, it can be argued that the complexity of dependencies is quite limited. Our main conclusion (discussed in detail in Section 4) is that although a whole range of complex dependencies are possible, many of these can be factored out. The actual types of valid dependencies observed in the data are highly restricted, especially when it is recognized that: i) one of the arguments of the so-called adverbial connectives is always *anaphoric* (a claim that has been extensively investigated in [12]); or ii) *attribution* within an argument belongs to a different component of discourse and is not considered part of the discourse structure. In future work, we intend to pursue this issue cross-linguistically. Our conjecture is that the complexity of dependencies is far more restricted at the discourse level as compared to the syntactic level, even for languages whose complexity at the syntactic level is much higher than English.[1]

2 The Penn Discourse TreeBank (PDTB)

The Penn Discourse Treebank [10] contains annotations of discourse relations and their arguments on the 1 million word Wall Street Journal (WSJ) corpus. Following the approach towards discourse structure and discourse semantics in D-LTAG [12], the PDTB annotates semantic or informational relations holding between two (and only two) Abstract Objects (AOs), expressed either explicitly via lexical items or implicitly via adjacency. For the former, the lexical items anchoring the relation are annotated as Explicit connectives. For the latter, the implicit inferrable relations are annotated by inserting an Implicit connective that best expresses the inferred relation. In Example (1), the subordinating conjunction *since* is an Explicit connective anchoring a TEMPORAL relation between the event of the earthquake hitting and a state where no music is played by a certain woman. (The 4-digit number in parentheses at the end of examples is the WSJ file number of the text.)

(1) *She hasn't played any music* since **the earthquake hit**. (0766)

An example of a relation inferred due to adjacency is given in (2), where the CAUSAL relation between the AOs denoted by the two adjacent sentences is annotated with *because* as the Implicit connective.

(2) *Also unlike Mr. Ruder, Mr. Breeden appears to be in a position to get somewhere with his agenda*. Implicit=BECAUSE (CAUSE) **As a former White House [...], he is savvy in the ways of Washington**. (0955)

Explicit connectives are identified from three grammatical classes: subordinating conjunctions (e.g., *because*, *when*), coordinating conjunctions (e.g., *and*, *or*),

[1] Czech would be an excellent language to test our hypothesis. We are currently discussing with the Prague Dependency Treebank (PDT) group to see if a portion of their corpus could be annotated for discourse connectives and their arguments in the style of the PDTB. Interestingly, in a recent study of the complexity of syntactic dependencies, it has been observed that 99.89% of the dependencies in the PDT fall into the class of the so-called well-nested dependencies, which includes many crossing dependencies. The class of well-nested dependencies is related to the derivations in LTAG [6, 1].

and discourse adverbials (e.g., *however*, *otherwise*). Arguments of connectives are simply labelled Arg2 for the argument appearing in the clause syntactically bound to the connective, and Arg1 for the other argument. In the examples in this section, Arg1 appears in italics while Arg2 appears in bold.

In addition to the argument structure of discourse relations, the PDTB also annotates the *attribution* of relations (both explicit and implicit) as well as of each of their arguments. A variety of cases can be distinguished depending on the *attribution source* of the discourse relation or its arguments; i.e., whether the relation or arguments are ascribed to the writer of the text or to someone other than the writer. A full description of attribution in the PDTB can be found in [8], [2] and [9]. As an example, in (3), the relation and Arg2 are attributed to the writer, but Arg1 is attributed to some other speaker, here Mr. Green.

(3) When **Mr. Green won a $240,000 verdict in a land condemnation case against the state in June 1983**, he says *Judge O.Kicki unexpectedly awarded him an additional $100,000*. (0267)

The first release (April 2006) of the Penn Discourse Treebank, PDTB 1.0 [8], is freely available from `http://www.seas.upenn.edu/~pdtb`. For a comparison of the PDTB with related efforts towards discourse annotation, see [5].

3 Complexity of Dependencies

Before describing the more unusual types of dependencies that we have come across in the PDTB, it should be pointed out that in the majority of cases within the corpus, a pair of discourse relations enter into a very “normal” structural relationship with one another, that is, they are tree structures. There are two types of these pairs of tree structures - independent relations and full embeddings. The former is a very trivial case, but we discuss it here for the sake of completeness. Independent relations refer to the very common situation where one discourse relation simply follows another in sequence, with their argument spans being entirely independent of one another (see Figure 1a.). An example is shown below[2]:

(4) **The securities-turnover tax has been long criticized by the West German financial community** BECAUSE **it tends to drive securities trading and other**

[2] Note that our notational convention here is different from the one used in Section 2 because we have to show the arguments for pairs of relations instead of individual relations. So here, connectives are in boxes, the arguments of the first connective will be shown in boldface, and the arguments of the second connective will be in italics. If some portion of the arguments of the first connective overlap with some portion of the arguments of the second, the overlapping spans will be in both boldface and italics. We shall see numerous cases of such overlaps in subsequent examples. For convenience, we have included a table under each example to clarify the various arguments of each connective.

banking activities out of Frankfurt into rival financial centers, especially London, where trading transactions isn't taxed. *The tax has raised less than one billion marks ($545.3 million) annually in recent years,* BUT *the government has been reluctant to abolish the levy for budgetary concerns*. (0302)

Conn	Arg1	Arg2
BECAUSE	The securities-turnover...community	it tends...isn't taxed
BUT	The tax has raised...recent years	the government...concerns

In (4), since the relation headed by BECAUSE fully precedes the relation headed by BUT, the two discourse relations are essentially independent of one another and there is no overlap or crossing of any kind.

A "fully embedded" structure is another common one encountered. This is where one discourse relation is entirely realized as one of the arguments of another discourse connective (see Figure 1b). An example of this is shown below:

(5) **The drop in earnings had been anticipated by most Wall Street analysts**, BUT ***the results were reported*** AFTER ***the market closed***. (1221)

Conn	ARG1	ARG2
BUT	The drop...analysts	the results...market closed
AFTER	the results were reported	the market closed

In (5), the AFTER relation is wholly embedded as the ARG2 of the BUT relation. This full-embedding type of structure is a common occurrence in syntax. For example, a clause can be embedded within a higher clause, serving as the argument to the higher predicate. This is a simple tree structure and its ubiquitous presence in discourse is not surprising.

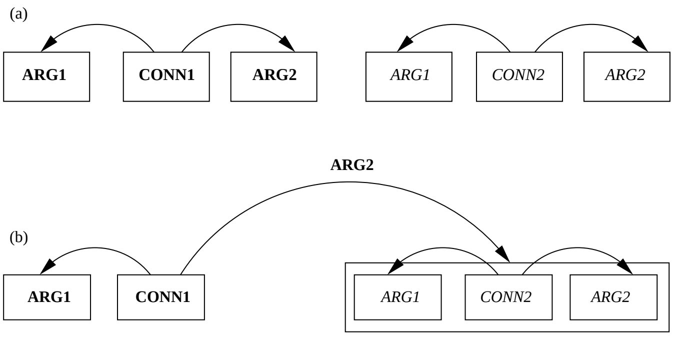

Figure 1: Two tree structures: (a) independent relations; (b) full embedding

We now turn our attention to the more unusual dependencies that appear in the corpus. For this study, we identified and searched for four types of non-tree-like dependencies, as enumerated below:

(i) **Shared argument**: two connectives share the same argument span (Fig.2a)

(ii) **Properly contained argument**: the argument span of a connective is fully contained within a larger argument span of another connective (Fig.2b)

(iii) **Pure crossing**: the argument node of one connective is interspersed within the arguments of another connective (Fig. 2c)

(iv) **Partially overlapping arguments**: the argument span of one connective partially overlaps the argument span of another connective (Fig.2d)

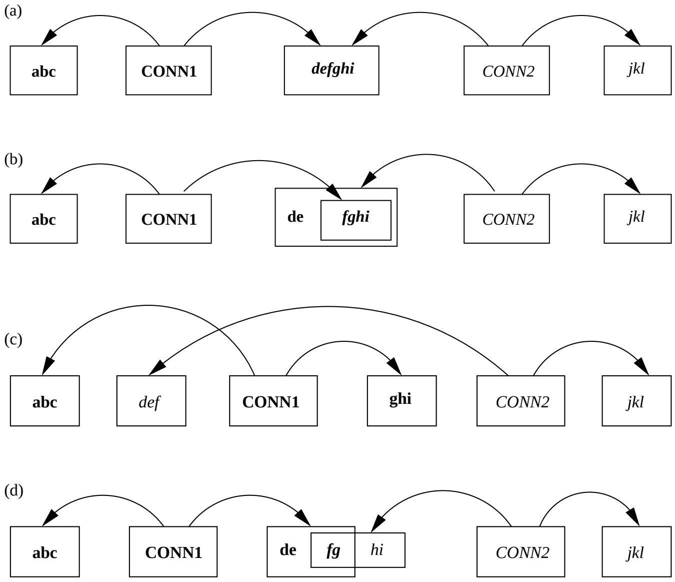

Figure 2: Four types of non-tree-like dependency structures: (a) Shared argument; (b) Properly contained argument; (c) Pure crossing; (d) Partially overlapping arguments. The small-case letters "abc...jkl" represent sequentially ordered strings in the discourse.

Each of these four types will be discussed and exemplified in the following subsections. A fuller discussion will follow in Section 4.

3.1 Shared argument

This refers to two connectives which share the exact same text span for one of their arguments. The example below illustrates this scenario:

(6) **In times past, life-insurance salesmen targeted heads of household, meaning men,** **BUT** ***ours is a two-income family and accustomed to it***. *SO* *if anything happened to me, I'd want to leave behind enough so that my 33-year-old husband would be able to pay off the mortgage and some other debts.* (1574)

Conn	ARG1	ARG2
BUT	In times past...meaning men	ours is...accustomed to it
SO	ours is...accustomed to it	if anything happened...other debts

In this case, the text span of the Arg2 of BUT is exactly the span for the Arg1 of SO, hence the argument is shared.

3.2 Properly contained argument

In this case, the span of one argument of a connective is fully embedded within a larger span of text which constitutes the argument of another connective:

(7) Japanese retail executives say the main reason they are reluctant to jump into the fray in the U.S. is **that – unlike manufacturing – retailing is extremely sensitive to local cultures and life styles**. **IMPLICIT-FOR EXAMPLE** *The Japanese have watched* ***the Europeans and Canadians stumble in the U.S. market***, *AND* *they fret that business practices that have won them huge profits at home won't translate into success in the U.S.* (0814)

Conn	ARG1	ARG2
FOR EXAMPLE	that – unlike...life styles	the Europeans...market
AND	The Japanese...market	they fret...in the U.S.

In (7), the clause in bold and italics "the Europeans and Canadians stumble in the U.S. market" belongs to both the implicit connective FOR EXAMPLE as well as the explicit connective AND. This clause is the exact Arg2 of FOR EXAMPLE, but it combines with some other text, namely "The Japanese have watched" to form the Arg1 of AND. Hence, the second argument of FOR EXAMPLE is properly contained within a larger text span that constitutes the first argument of AND.

3.3 Pure crossing

In purely crossing structures, the argument(s) of one connective is interleaved within the arguments of the other connective. (8) shows such a structure:

(8) **Under these deals, the RTC sells just the deposits and the healthy assets**. *These "clean-bank" transactions leave the bulk of bad assets, mostly real estate, with the government,* **to be sold** **LATER**. *In these four,* *FOR INSTANCE*, *the RTC is stuck with $4.51 billion in bad assets.* (2348)

Conn	ARG1	ARG2
LATER	Under these deals...healthy assets	to be sold
FOR INSTANCE	These “clean-bank”...government	In these four...assets

Here, the Arg1 of FOR INSTANCE is interspersed between the two arguments of LATER while Arg2 of FOR INSTANCE comes after the second argument of LATER, hence the interleaving.

3.4 Partially overlapping arguments

Partially overlapping arguments are cases where two arguments of different connectives mutually share only portions of their text spans:

(9) **He (Mr. Meek) said** ***the evidence pointed to wrongdoing by Mr. Keating "and others***," ***ALTHOUGH*** ***he didn’t allege any specific violation***. Richard Newsom, a California state official who last year examined Lincoln’s parent, American Continental Corp., said *he* *ALSO* *saw evidence that crimes had been committed*. (0335)

Conn	ARG1	ARG2
ALTHOUGH	He said the evidence...and others	he didn’t allege...violation
ALSO	the evidence...violation	he saw....committed

In (9) above, Arg1 of ALTHOUGH includes the higher verb of saying (i.e., “He said”), whereas Arg1 of ALSO begins at the lower clause (i.e.. “the evidence...”) and extends further to include the adjunct ALTHOUGH clause. These two arguments share only a portion of their spans, with each having some ‘leftover’ material not part of the intersected spans. Their arguments thus partially overlap.

4 Discussion

If the various dependency structures discussed in the previous section are all taken into account, the complexity of dependencies at the discourse level will be increased quite significantly. The existence of pure crossing dependencies would indeed mean that there exist complex structures in discourse that we do not find at the syntactic level in English. Moreover, there is as yet no syntactic parallel to the phenomena of partially overlapping arguments or properly contained arguments that we seem to observe in the PDTB.

Nevertheless, we argue here that at most, only the shared argument structures and most likely only a subset of structures with properly contained arguments should be considered part of the discourse structure. Two factors support this conclusion. Firstly, the other two types of non-tree-like dependencies discussed - pure

Shared argument	$\sim$1400	7.5%
Properly contained argument	$\sim$400	2%
Pure crossing	24	.12%
Partially crossing	4	negligible

Figure 3: Frequency count for the four non-tree-like structures under discussion (percentage figures are based on the approximately 20000 annotated tokens in the PDTB)

crossing dependencies and partially overlapping arguments - can be explained by appealing to non-structural phenomena in discourse, particularly *anaphora* and *attribution*. A similar argument has been made elsewhere ([11], [7]) to argue against the crossing structures presented as evidence for a graph-based representation of discourse [13]. Secondly, figure 3, which lists the approximate frequency counts for all four structures discussed in this paper, shows relatively high counts for structures containing a shared argument or a properly contained argument. On the other hand, the count for the remaining two structures are low, and may simply be due to annotation noise if not ruled out by anaphora or attribution. In what follows, we discuss each structure in turn, classifying each as either rare or common.

4.1 Rare discourse structures

4.1.1 Pure crossing

One generalization from our observation thus far is that with purely crossing dependencies, at least one of the connectives (and often both) are *discourse adverbials*. The class of discourse adverbials would seem to include, among others, connectives that specify temporal conjunctions (“then”, “later”, etc.), or connectives presupposing shared knowledge of a generalization or set (“for example”, “also”, “first...second”, etc.) [4]. It has been argued that discourse adverbials take only one of their arguments structurally, the other argument being anaphoric [12]. This contrasts with other classes of connectives like subordinating and coordinating conjunctions which take both their arguments structurally, allowing the semantics of the relation to be derived compositionally.

In the PDTB, no explicit distinction is made between these sub-classes of discourse connectives. Arguments for both structural and anaphoric connectives are annotated in the same way, since the corpus aims to be as theory-independent as possible in order to be useful to people working within different frameworks. Nevertheless, if we accept the distinction between structural and anaphoric connectives and claim that one of the arguments of a discourse adverbial is not specified structurally, the existence of crossing dependencies in the PDTB is rendered spurious and is indeed a by-product of its theory-neutral approach to annotation.

4.1.2 Partially overlapping arguments

There is a very limited set of cases with partially overlapping arguments in the corpus, which suggests that this kind of pattern might not be integral to discourse structure. The few cases that exist might well be amenable to reanalysis, if not treated as aberrations. Moreover, even if we accept the annotations as they stand, it seems that one of the ‘leftover’ materials in partially overlapping arguments is a higher verb of *attribution*. For instance, in Example (9) seen above, the higher verb “He said” is included as an argument of ALTHOUGH, but excluded as an argument of ALSO. The PDTB framework sees discourse relations (associated with an explicit or implicit connective) as holding between two *abstracts objects*, such as events, states, etc. On the other hand, attribution relates a proposition to an agent/individual entity, not to another proposition, event, etc. Hence, if we take the strict view that only abstract objects are involved in discourse relations and we leave the analysis and processing of attribution to a different component of discourse, we can essentially excise the attribution spans from the rest of the argument span. Doing so with partially overlapping arguments seems to leave us with a structure having only a properly contained argument.

4.2 Common discourse structures

4.2.1 Shared arguments

Shared arguments are ubiquitous in the PDTB, with approximately 1400 instances accounted for out of 20000 annotated tokens in the corpus (Fig. 3). There is indeed no plausible way to appeal to the notion of attribution to rule out a shared argument node which is coextensive and an integral part of two discourse relations. Moreover, it appears that two structural connectives can share an argument node, so we cannot appeal to the notion of discourse anaphora to rule out such cases either. Thus, we argue that a theory of discourse structure must account for shared argument nodes, i.e. a node with multiple ancestors. This kind of structure is not unique to discourse. Shared nodes are even encountered at the level of syntax, and many frameworks introduce trace elements to deal with such cases.

4.2.2 Properly contained argument

Structures with a properly contained argument are also plentiful in the PDTB (see Fig.3, where approximately 400 instances have been identified). In a number of cases that we examined, the so-called ‘leftover’ material of the larger argument text span (once the span of the contained argument is subtracted) is a verb of attribution not unlike the case with partially overlapping arguments just discussed. Nevertheless, there are many more cases where we cannot factor out the dependencies due to attribution. Here is an example:

(10) *The pound, which had been trading at about $1.6143 in New York prior to Mr. Lawson's announcement, sank more than two cents to $1.5930, prompting* ***the Federal Reserve Bank to buy pounds for dollars***. **The Fed's move, HOWEVER, only proved a stopgap to the pound's slide** and the Fed intervened for a second time at around $1.5825 [...]. *MEANWHILE, dollar trading was relatively uninspired throughout the session, according to dealers.* (0769)

Conn	ARG1	ARG2
HOWEVER	the Federal Reserver...dollars	The Fed's...pound's slide
MEANWHILE	The pound...pounds for dollars	dollar trading...dealers

Example (10) shows another properly contained argument: the Arg1 of HOWEVER appears within the Arg1 of MEANWHILE. The 'leftover' material once we subtract the properly contained argument is a quite complicated hodgepodge of clauses: "The pound, which had been trading at about $1.6143 in New York prior to Mr. Lawson's announcement, sank more than two cents to $1.5930, prompting". This certainly cannot be written off as an attribution. Further study will be needed to classify the various types of 'leftover' materials in this particular structure.

5 Conclusion

In this paper, we examined a range of dependencies observed between connectives and their arguments in the PDTB corpus. We noted that there were a number of unusual structures which have pure crossing dependencies, partially overlapping arguments, a properly contained argument or a shared argument. However, by appealing to notions such as *anaphora* and *attribution*, and by roughly estimating how common or uncommon each of these unusual structures are in the PDTB, we argued that pure crossing dependencies, partially overlapping arguments and a subset of structures containing a properly contained argument should not be considered part of the discourse structure. The actual types of dependencies observed in the data are therefore highly restricted. Discourse structure is hence likely less complex than syntactic structure.

In the future, we plan to: i) continue this current work and reexamine our results after the final release of the PDTB corpus in April 2007; ii) carry out more formal investigations of these classes of dependencies, comparable to the studies in [6]; iii) carry out cross-linguistic studies of dependencies at the discourse level.

References

[1] Manueal Bodirsky, Marco Kuhlmann, and Mathias Mohl. Well-nested drawings as models of syntactic structure. In *Proc. of the 10th Conference on Formal Grammar and 9th Meeting of Mathematics of Language*, 2005.

[2] Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Attribution and the (non)-alignment of syntactic and discourse arguments of connectives. In *Proc. of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, 2005.

[3] Kate Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind Joshi, and Bonnie Webber. D-LTAG system: Discourse parsing with a lexicalized tree adjoining grammar. *Journal of Logic, Language and Information*, 12(3), 2003.

[4] Katherine Forbes-Riley, Bonnie Webber, and Aravind Joshi. Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, 23:55–106, 2006.

[5] Aravind Joshi, Rashmi Prasad, and Bonnie Webber. Discourse annotation: Discourse connectives and discourse relations. Tutorial at the Association for Computational Linguistics. See http://www.seas.upenn.edu/ pdtb/papers/joshi-etal06-discourse-annotation-tutorial.ppt, 2006.

[6] Marco Kuhlmann and Joakim Nivre. Mildly non-projective dependency structure. In *Proc. of the TAG+8 Workshop*, 2006.

[7] Eleni Miltsakaki. Some thoughts on attribution. Presentation at the Seminar on Discourse Structure (CIS 639), IRCS, University of Pennsylvania, 2005.

[8] The PDTB-Group. The Penn Discourse TreeBank 1.0 Annotation Manual. Technical Report IRCS-06-01, IRCS, University of Pennsylvania, 2006.

[9] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. Annotating attribution in the Penn Discourse TreeBank. In *Proc. of the ACL Workshop on Sentiment and Subjectivity in Text*, 2006.

[10] Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, and Bonnie Webber. Annotation and data mining of the Penn Discourse Treebank. In *Proc. of the ACL Workshop on Discourse Annotation*, 2004.

[11] Bonnie Webber. Accounting for discourse relations: Constituency and dependency. In Mary Dalrymple Miriam Butt and Tracy King, editors, *Intelligent Linguistic Architectures*, pages 339–360. CSLI Publications, 2006.

[12] Bonnie Webber, Aravind Joshi, Mathew Stone, and Alistair Knott. Anaphora and Discourse Structure. *Computational Linguistics*, 29(4), 2003.

[13] Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 32(2):249–287, 2005.