

Representing Discourse Coherence: A Corpus-Based Study

Florian Wolf*
University of Cambridge

Edward Gibson**
Massachusetts Institute of Technology

This article aims to present a set of discourse structure relations that are easy to code and to develop criteria for an appropriate data structure for representing these relations. Discourse structure here refers to informational relations that hold between sentences in a discourse. The set of discourse relations introduced here is based on Hobbs (1985).

We present a method for annotating discourse coherence structures that we used to manually annotate a database of 135 texts from the Wall Street Journal and the AP Newswire. All texts were independently annotated by two annotators. Kappa values of greater than 0.8 indicated good interannotator agreement.

We furthermore present evidence that trees are not a descriptively adequate data structure for representing discourse structure: In coherence structures of naturally occurring texts, we found many different kinds of crossed dependencies, as well as many nodes with multiple parents. The claims are supported by statistical results from our hand-annotated database of 135 texts.

1. Introduction

An important component of natural language discourse understanding and production is having a representation of discourse structure. A coherently structured discourse here is assumed to be a collection of sentences that are in some relation to each other. This article aims to present a set of discourse structure relations that are easy to code and to develop criteria for an appropriate data structure for representing these relations.

There have been two kinds of approaches to defining and representing discourse structure and coherence relations. These approaches differ with respect to what kinds of discourse structure they are intended to represent. Some accounts aim to represent the intentional-level structure of a discourse; in these accounts, coherence relations reflect how the role played by one discourse segment with respect to the interlocutors' intentions relates to the role played by another segment (e.g., Grosz and Sidner 1986). Other accounts aim to represent the informational structure of a discourse; in these accounts, coherence relations reflect how the meaning conveyed by one discourse segment relates to the meaning conveyed by another discourse segment (e.g., Hobbs 1985; Marcu 2000; Webber et al. 1999). Furthermore, accounts of discourse structure vary greatly with respect to how many discourse relations they assume, ranging from 2 (Grosz and Sidner 1986) to over 400 different coherence relations (reported in Hovy and

* Computer Laboratory and Genetics Department, Cambridge, CB3 0FD, U.K.
E-mail: Florian.Wolf@cl.cam.ac.uk

** Department of Brain and Cognitive Sciences, Cambridge, MA 02139. E-mail: egibson@mit.edu.

Maier [1995]). However, Hovy and Maier (1995) argue that, at least for informational-level accounts, taxonomies with more relations represent subtypes of taxonomies with fewer relations. This means that different informational-level-based taxonomies can be compatible with each other; they differ with respect to how detailed or fine-grained a manner they represent informational structures of texts. Going beyond the question of how different informational-level accounts can be compatible with each other, Moser and Moore (1996) discuss the compatibility of rhetorical structure theory (RST) (Mann and Thompson 1988) with the theory of Grosz and Sidner (1986). However, note that Moser and Moore (1996) focus on the question of how compatible the claims are that Mann and Thompson (1988) and Grosz and Sidner (1986) make about intentional-level discourse structure.

In this article, we aim to develop an easy-to-code representation of informational relations that hold between sentences or other nonoverlapping segments in a discourse monologue. We describe an account with a small number of relations in order to achieve more generalizable representations of discourse structures; however, the number is not so small that informational structures that we are interested in are obscured. The goal of the research presented is not to encode intentional relations in texts. We consider annotating intentional relations too difficult to implement in practice at this time. Note that we do not claim that intentional-level structure of discourse is not relevant to a full account of discourse coherence; it just is not the focus of this article.

The next section describes in detail the set of coherence relations we use, which are mostly based on Hobbs (1985). We try to make as few a priori theoretical assumptions about representational data structures as possible. These assumptions are outlined in the next section. Importantly, however, we do not assume a tree data structure to represent discourse coherence structures. In fact, a major result of this article is that trees do not seem adequate to represent discourse structures.

This article is organized as follows. Section 2 describes the procedure we used to collect a database of 135 texts annotated with coherence relations. Section 3 describes in detail the descriptive inadequacy of tree structures for representing discourse coherence, and Section 4 provides statistical evidence from our database that supports this claim. Section 5 offers some concluding remarks.

2. Collecting a Database of Texts Annotated with Coherence Relations

This section describes (1) how we defined discourse segments, (2) which coherence relations we used to connect discourse segments, and (3) how the annotation procedure worked.

2.1 Discourse Segments

There is agreement that discourse segments should be nonoverlapping spans of text. However, there is disagreement in the literature about how to define discourse segments (cf. the discussion in Marcu [2000]). Whereas some argue that discourse segments should be prosodic units (Hirschberg and Nakatani 1996), others argue for intentional units (Grosz and Sidner 1986), phrasal units (Lascarides and Asher 1993; Longacre 1983; Webber et al. 1999), or sentences (Hobbs 1985).

For our database, we mostly adopted a clause-unit-based definition of discourse segments. We chose this method of segmenting discourse because it was easy to use.

Table 1

Contentful conjunctions used to illustrate coherence relations.

<i>Cause-effect</i>	because; and so
<i>Violated expectation</i>	although; but; while
<i>Condition</i>	if ... (then); as long as; while
<i>Similarity</i>	and; (and) similarly
<i>Contrast</i>	by contrast; but
<i>Temporal sequence</i>	(and) then; first, second, ...; before; after; while
<i>Attribution</i>	according to ...; ... said; claim that ...; maintain that ...; stated that ...
<i>Example</i>	for example; for instance
<i>Elaboration</i>	also; furthermore; in addition; note (furthermore) that; (for, in, on, against, with, ...) which; who; (for, in, on, against, with, ...) whom
<i>Generalization</i>	in general

However, we also assumed that contentful coordinating and subordinating conjunctions (cf. Table 1) can delimit discourse segments.

Note that we did not classify *and* as delimiting discourse segments if it was used to conjoin nouns in a conjoined noun phrase, like *dairy plants and dealers* in example (1) (from wsj_0306; Wall Street Journal 1989 corpus [Harman and Liberman 1993]) or if it was used to conjoin verbs in a conjoined verb phrase, like *snowed and rained* in example (2) (constructed):

- (1) Milk sold to the nation's dairy plants and dealers averaged \$14.50 for each hundred pounds.
- (2) It snowed and rained all day long.

We classified periods, semicolons, and commas as delimiting discourse segments. However, in cases like example (3) (constructed), in which they conjoin a complex noun phrase, commas were not classified as delimiting discourse segments.

- (3) John bought bananas, apples, and strawberries.

We furthermore treated attributions (*John said that . . .*) as discourse segments. This was empirically motivated. The texts used here were taken from news corpora, and there, attributions can be important carriers of coherence structures. For instance, consider a case in which some source *A* and some source *B* both comment on some event *X*. It should be possible to distinguish between a situation in which source *A* and source *B* make basically the same statement about event *X* and a situation in which source *A* and source *B* make contrasting comments about event *X*. Note, however, that we treated cases like example (4) (constructed) as one discourse segment and not as two separate ones (. . . *cited* and *transaction costs . . .*). We separated attributions only if the attributed material was a complementizer phrase, a sentence, or a group of sentences. This is not the case in example (4): The attributed material is a complex NP (*transaction costs from its 1988 recapitalization*).

- (4) The restaurant operator cited transaction costs from its 1988 recapitalization.

2.2 Discourse Segment Groupings

Adjacent discourse segments could, in our approach, be grouped together. For example, discourse segments were grouped if they all stated something that could be attributed to the same source (cf. section 2.3 for a definition of attribution coherence relations). Furthermore, discourse segments were grouped if they were topically related. For example, if a text discussed inventions in information technology, there could be groups of a few discourse segments each talking about inventions by specific companies. There might also be subgroups, consisting of several discourse segments each, talking about specific inventions at specific companies. Thus, marking groups could determine a partially hierarchical structure for the text.

Other examples of discourse segment groupings included cases in which several discourse segments described an event or a group of events that all occurred before another event or another group of events described by another (group of) discourse segments. In those cases, what was described by a group of discourse segments was in a temporal sequence relation with what was described by another (group of) discourse segments (cf. section 2.3 for a definition of temporal-sequence coherence relations). Note furthermore that in cases in which one topic required one grouping and a following topic required a grouping that was different from the first grouping, both groupings were annotated.

Unlike approaches such as the TextTiling algorithm (Hearst 1997), ours allowed partially overlapping groups of discourse segments. The idea behind this option was to allow groupings of discourse segments in which a transition discourse segment belonged to the previous as well as the following group. However, the option was not used by the annotators (i.e., in our database of 135 hand-annotated texts, there were no instances of partially overlapping discourse segment groups).

2.3 Coherence Relations

As pointed out in section 1, we aim to develop a representation of informational relations between discourse segments. Note one difference between schema-based approaches (McKeown 1985) and coherence relations as we used them: Whereas schemas are instantiated from information contained in a knowledge base, coherence relations as we used them do not make (direct) reference to a knowledge base.

There are a number of different informational coherence relations, dating back, in their basic definitions, to Hume, Plato, and Aristotle (cf. Hobbs 1985; Hobbs et al. 1993; Kehler 2002). The coherence relations we used are mostly based on Hobbs (1985); below we describe each coherence relation we used and note any differences between ours and Hobbs's (1985) set of coherence relations (cf. Table 2 for an overview of how our set of coherence relations relates to the set of coherence relations in Hobbs [1985]).

The kinds of coherence relations we used include *cause-effect* relations, as in example (5) (constructed), in which discourse segment 1 states the cause for the effect that is stated in discourse segment 2:

- (5) Cause-effect
1. There was bad weather at the airport
 2. and so our flight got delayed.

Our *cause-effect* relation subsumed the *cause* as well as the *explanation* relation in Hobbs (1985). A *cause* relation holds if a discourse segment stating a cause occurs

before a discourse segment stating an effect; an *explanation* relation holds if a discourse segment stating an effect occurs before a discourse segment stating a cause. We encoded this difference by adding a direction to the *cause-effect* relation. In a graph, this can be represented by a directed arc going from cause to effect.

Another kind of causal relation is *condition*. Hobbs (1985) does not distinguish *condition* relations from either *cause* or *explanation* relations. However, we felt that it might be important to distinguish between a causal relation describing an actual causal event (*cause-effect*, cf. above), on the one hand, and a causal relation describing a possible causal event (*condition*, cf. below), on the other hand. In example (6) (constructed), discourse segment 2 states an event that will take place if the event described by discourse segment 1 also takes place:

- (6) Condition
1. If the new software works,
 2. everyone should be happy.

In a third type of causal relation, the *violated expectation* relation (also *violated expectation* in Hobbs [1985]), a causal relation between two discourse segments that normally would be present is absent. In example (7) (constructed), discourse segment 1 normally would be a cause for everyone's being happy; this expectation is violated by what is stated by discourse segment 2:

- (7) Violated expectation
1. The new software worked great,
 2. but nobody was happy.

Other possible coherence relations include *similarity* (*parallel* in Hobbs [1985]) or *contrast* (also *contrast* in Hobbs [1985]) relations, in which similarities or contrasts are determined between corresponding sets of entities or events, such as between discourse segments 1 and 2 in example (8) (constructed) and discourse segments 1 and 2 in example (9) (constructed), respectively:

- (8) Similarity
1. The first flight to Frankfurt this morning was delayed.
 2. The second flight arrived late as well.

- (9) Contrast
1. The first flight to Frankfurt this morning was delayed.
 2. The second flight arrived on time.

Discourse segments might also *elaborate* (also *elaboration* in Hobbs [1985]) on other sentences, as in example (10) (constructed), in which discourse segment 2 elaborates on discourse segment 1:

- (10) Elaboration
1. A probe to Mars was launched from the Ukraine this week.
 2. The European-built "Mars Express" is scheduled to reach Mars by late December.

Discourse segments can provide examples for what is stated by another discourse segment. In example (11) (constructed), discourse segment 2 states an *example*

(*exemplification* in Hobbs [1985]) for what is stated in discourse segment 1:

- (11) Example
1. There have been many previous missions to Mars.
 2. A famous example is the Pathfinder mission.

Hobbs (1985) also includes an *evaluation* relation, as in example (12) (from Hobbs [1985]), in which discourse segment 2 states an evaluation of what is stated in discourse segment 1. We decided to call such relations *elaborations*, since we found it too difficult in practice to reliably distinguish *elaborations* from *evaluations* (according to our annotation scheme, in example (12), what is stated in discourse segment 2 elaborates on what is stated in discourse segment 1):

- (12) Elaboration (labeled as evaluation in Hobbs [1985])
1. (A story.)
 2. It was funny at the time.

Unlike Hobbs (1985), we did not have a separate *background* relation as in example (13) (modified from Hobbs [1985]), in which what is stated in discourse segment 1 provides the background for what is stated in discourse segment 2. As with the *evaluation* relation, we found the *background* relation too difficult to reliably distinguish from *elaboration* relations (according to our annotation scheme, in example (13), what is stated in discourse segment 1 elaborates on what is stated in discourse segment 2):

- (13) Elaboration (labeled as background in Hobbs [1985])
1. T is the pointer to the root of a binary tree.
 2. Initialize T.

In a *generalization* relation, as in example (14) (constructed), one discourse segment (here discourse segment 2) states a generalization for what is stated by another discourse segment (here discourse segment 1):

- (14) Generalization
1. Two missions to Mars in 1999 failed.
 2. There are many missions to Mars that have failed.

Furthermore, discourse segments can be in an *attribution* relation, as in example (15) (constructed), in which discourse segment 1 states the source of what is stated in discourse segment 2 (cf. [Bergler 1991] for a more detailed semantic analysis of *attribution* relations):

- (15) Attribution
1. John said that
 2. the weather would be nice tomorrow.

Hobbs (1985) does not include an *attribution* relation. However, we decided to include *attribution* as a relation because, as pointed out in section 2.1, the texts we annotated are taken from news corpora. There, attributions can be important carriers of coherence structures.

In a *temporal sequence* relation, as in example (16) (constructed), one discourse segment (here discourse segment 1) states an event that takes place before another event stated by the other discourse segment (here discourse segment 2):

- (16) Temporal Sequence
1. First, John went grocery shopping.
 2. Then he disappeared in a liquor store.

In contrast to *cause-effect* relations, there is no causal relation between the events described by the two discourse segments. The *temporal sequence* relation is equivalent to the *occasion* relation in Hobbs (1985).

The *same* relation, illustrated by example (17) (constructed), is an epiphenomenon of assuming contiguous distinct elements of text (Hobbs [1985] does not include a *same* relation). A *same* relation holds if a subject NP is separated from its predicate by an intervening discourse segment. For instance, in example (17), discourse segment 1 is the subject NP of a predicate in discourse segment 3, and so there is a *same* relation between discourse segments 1 and 3; discourse segment 1 is the first and discourse segment 3 is the second segment of what is actually one single discourse segment, separated by the intervening discourse segment 2, which is in an *attribution* relation with discourse segment 1 (and therefore also with discourse segment 3, since discourse segments 1 and 3 are actually one single discourse segment):

- (17) Same
1. The economy,
 2. according to some analysts,
 3. is expected to improve by early next year.

Table 2 provides an overview of how our set of coherence relations relates to the set of coherence relations in Hobbs (1985).

We distinguish between asymmetrical or directed relations, on the one hand, and symmetrical or undirected relations, on the other hand (Mann and Thompson 1988; Marcu 2000). *Cause-effect*, *condition*, *violated expectation*, *elaboration*, *example*, *generalization*, *attribution*, and *temporal sequence* are asymmetrical or directed relations, whereas *similarity*, *contrast*, and *same* are symmetrical or undirected relations. In asymmetrical or directed relations, the directions of relations are as follows:

- *Cause-effect*: from the discourse segment stating the cause to the discourse segment stating the effect
- *Condition*: from the discourse segment stating the condition to the discourse segment stating the consequence
- *Violated expectation*: from the discourse segment stating the cause to the discourse segment describing the absent effect
- *Elaboration*: from the elaborating discourse segment to the elaborated discourse segment
- *Example*: from the discourse segment stating the example to the discourse segment stating the exemplified
- *Generalization*: from the discourse segment stating the special case to the discourse segment stating the general case

Table 2

Correspondence between the set of coherence relations in Hobbs (1985) and our set of coherence relations.

Hobbs (1985)	Our annotation scheme
<i>Occasion</i>	<i>Temporal sequence</i>
<i>Cause</i>	<i>Cause-effect</i> : cause stated first, then effect; directionality indicated by directed arcs in a coherence graph
<i>Explanation</i>	<i>Cause-effect</i> : effect stated first, then cause; directionality indicated by directed arcs in a coherence graph
—	<i>Condition</i>
<i>Evaluation</i>	<i>Elaboration</i>
<i>Background</i>	<i>Elaboration</i>
<i>Exemplification</i> : example stated first, then general case; directionality indicated by directed arcs in a coherence graph	<i>Example</i>
<i>Exemplification</i> : general case stated first, then example; directionality indicated by directed arcs in a coherence graph	<i>Generalization</i>
<i>Elaboration</i>	<i>Elaboration</i>
<i>Parallel</i>	<i>Similarity</i>
<i>Contrast</i>	<i>Contrast</i>
<i>Violated expectation</i>	<i>Violated expectation</i>
—	<i>Attribution</i>
—	<i>Same</i>

- *Attribution*: from the discourse segment stating the source to the attributed statement
- *Temporal sequence*: from the discourse segment stating the event that happened first to the discourse segment stating the event that happened second

This definition of directionality is related to Mann and Thompson’s (1988) notion of nucleus and satellite nodes (where the nodes can represent [groups of] discourse segments): For asymmetrical or directed relations, the directionality is from satellite to nucleus node; by contrast, symmetrical or undirected relations hold between two nucleus nodes.

Note also that in our annotation project, we decided to annotate a coherence relation either if there was a coherence relation between the complete content of two discourse segments, or if there was a relation between parts of the content of two discourse segments. Consider the following example (from ap890104-0003; AP Newswire corpus; [Harman and Liberman 1993]):

- (18) 1. *a*[Difficulties have arisen] *b*[in enacting the accord for the independence of Namibia]
 2. for which SWAPO has fought many years,

For this example we would annotate an *elaboration* relation from discourse segment 2 to discourse segment 1 (discourse segment 2 provides additional details about the accord

mentioned in discourse segment 1), although the relation actually only holds between discourse segment 2 and the second part of discourse segment 1, indicated by brackets.

Although it is beyond the scope of the current project, future research should investigate annotations with discourse segmentations that allow annotating relations only between parts of discourse segments that are responsible for a coherence relation. For example, consider example (19) (from ap890104-0003; AP Newswire corpus [Harman and Liberman 1993]), in which brackets indicate how more-fine-grained discourse segments might be marked:

- (19) 1. a [for which] b [SWAPO] c [has fought many years,]
 2. referring to the acronym of the South-West African Peoples
 Organization nationalist movement.

In our current project, we annotated an *elaboration* relation from discourse segment 2 to discourse segment 1 (discourse segment 2 provides additional details, the full name, for *SWAPO*, which is mentioned in discourse segment 1). A future, more detailed, annotation of coherence relations could then annotate this *elaboration* relation to hold only between discourse segment 2 and the word *SWAPO* in discourse segment 1.

2.4 Coding Procedure

To code the coherence relations of a text, we used a procedure consisting of three steps. In the first step, a text was segmented into discourse segments (cf. section 2.1).

In the second step, adjacent discourse segments that were topically related were grouped together. The criteria for this step are described in section 2.2.

In the third step, coherence relations (cf. section 2.3) were determined between discourse segments and groups of discourse segments. Each previously unconnected (group of) discourse segment(s) was tested to see whether it connected to any of the (groups of) discourse segments that had already been connected to the already existing representation of discourse structure.

In order to help determine the coherence relation between (groups of) discourse segments, the annotators judged which, if any, of the contentful coordinating conjunctions in Table 1 resulted, when used, in the most acceptable passage (cf. Hobbs 1985; Kehler 2002). If using a contentful conjunction to connect (groups of) discourse segments resulted in an acceptable passage, this was used as evidence that the coherence relation corresponding to the mentally inserted contentful conjunction held between the (groups of) discourse segments under consideration. This mental exercise was done only if there was not already a contentful coordinating conjunction that disambiguated the coherence relation.

The following list (which was also used by the annotators to guide them in their task) shows in more detail how the annotations were carried out:

1. Segment the text into discourse segments:
 - (a) Insert segment boundaries at every period that marks a sentence boundary (i.e., not at periods such as those in *Mrs.* or *Dr.*).
 - (b) Insert segment boundaries at every semicolon and colon that marks a sentence or clause boundary.
 - (c) Insert segment boundaries at every comma that marks a sentence or clause boundary; do not insert segment boundaries at commas that conjoin complex noun or verb phrases.

- (d) Insert segment boundaries at every quotation mark, if there is not already a segment boundary based on steps (a)–(c).
 - (e) Insert segment boundaries at the contentful coordinating conjunctions listed in Table 1, if there is not already a segment boundary based on steps (a)–(d). For *and*, do not insert a segment boundary if it is used to conjoin verbs or nouns in a conjoined verb or noun phrase.
2. Generate groupings of related discourse segments:
- (a) Group contiguous discourse segments that are enclosed by pairs of quotation marks.
 - (b) Group contiguous discourse segments that are attributed to the same source.
 - (c) Group contiguous discourse segments that belong to the same sentence (marked by periods, commas, semicolons, or colons).
 - (d) Group contiguous discourse segments that are topically centered on the same entities or events.
3. Determine coherence relations between discourse segments and groups of discourse segments. For each previously unconnected (group of) discourse segment(s), test whether it connects to any of the (groups of) discourse segments that have already been connected to the already existing representation of discourse structure. Use the following steps for each decision:
- (a) Use pairs of quotation marks as a signal for *attribution*.
 - (b) For pairs of (groups of) discourse segments that are already connected with one of the contentful coordinating conjunctions from Table 1, choose the coherence relation that corresponds to the coordinating conjunction.
 - (c) For pairs of (groups of) discourse segments that are not connected with one of the contentful coordinating conjunctions from Table 1:
 - i. Mentally connect the (groups of) discourse segments with one of the coordinating conjunctions from Table 1 and judge whether the resultant passage sounds acceptable.
 - ii. If the passage sounds acceptable, choose the coherence relation that corresponds to the coordinating conjunction selected in step (c.i).
 - iii. If the passage does not sound acceptable, repeat step (c.i) until an acceptable coordinating conjunction is found.
 - iv. If the passage does not sound acceptable with any of the coordinating conjunctions from Table 1, assume that the (groups of) discourse segments under consideration are not related by a coherence relation.
 - (d) Iterative procedure for steps (a) and (b):
 - i. Start with any of the unambiguous coordinating conjunctions from Table 1 (*because, although, if . . . then, . . . said, for example*).

Table 3
Statistics for texts in our database.

	Number of words	Number of discourse segments
Mean	545	61
Minimum	161	6
Maximum	1,409	143
Median	529	60

- ii. If none of the unambiguous coordinating conjunctions results in an acceptable passage, use the more ambiguous coordinating conjunctions (*and, but, while, also, etc.*).
- (e) Important distinctions for steps (2) and (3) (this is based on practical issues that came up during the annotation project):
 - i. *Example versus elaboration*: An *example* relation sets up an additional entity or event (the example), whereas an *elaboration* relation provides more details about an already introduced entity or event (the one on which one elaborates).
 - ii. *Cause-effect versus temporal sequence*: Both *cause-effect* and *temporal sequence* describe a temporal order of events (in *cause-effect*, the cause has to precede the effect). However, only *cause-effect* relations have a causal relation between what is stated by the (groups of) discourse segments under consideration. Thus, if there is a causal relation between the (groups of) discourse segments under consideration, assume *cause-effect* rather than *temporal sequence* (cf. Lascarides and Asher 1993).

2.5 Annotators

The annotators for the database were MIT undergraduate students who worked in our lab as research students. For training, the annotators received a manual that described the background of the project, discourse segmentation, coherence relations and how to recognize them, and how to use the annotation tools that we developed in our lab (Wolf et al. 2003). The first author of this article provided training for the annotators. Training consisted of explaining the background of the project and the annotation method and of annotating example texts (these texts are not included in our database). Training took 8–10 hours in total, distributed over five days of a week. After completing the training, annotators worked independently.

2.6 Statistics on Annotated Database

In order to evaluate hypotheses about appropriate data structures for representing coherence structures, we have collected a database of 135 texts from the Wall Street Journal 1987–1989 (30 texts) and the AP Newswire 1989 (105 texts) (both from Harman and Liberman [1993]) in which the relations between discourse segments have been labeled with the coherence relations described above. Table 3 shows statistics for this database.

Steps 2 (discourse segment grouping) and 3 (coherence relation annotation) of the coding procedure described in section 2.4 were performed independently by two annotators. For step 1 (discourse segmentation), a pilot study on 10 texts showed that agreement on this step, as determined by $\text{number of common segments} / (\text{number of common segments} + \text{number of differing segments})$, was never below 90%. Therefore, all 135 texts were segmented by two annotators together, resulting in segmentations that both annotators could agree on.

In order to determine interannotator agreement for step 2 of the coding procedure for the database of annotated texts, we calculated kappa statistics (Carletta 1996). We used the following procedure to construct a confusion matrix: First, all groups marked by either annotator were extracted. Annotator 1 had marked 2,616 groups, and annotator 2 had marked 3,021 groups in the whole database. The groups marked by the annotators consisted of 536 different discourse segment group types (for example, groups that included the first two discourse segments of each text were marked 31 times by both annotators; groups that included the first three discourse segments of each text were marked 6 times by both annotators). Therefore, the confusion matrix had 536 rows and columns. For all annotations of the 135 texts, the agreement was 0.8449, per chance agreement was 0.0161, and kappa was 0.8424. Annotator agreement did not differ as a function of text length, arc length, or kind of coherence relation (all χ^2 values < 1).

We also calculated kappa statistics to determine interannotator agreement for step 3 of the coding procedure.¹ For all annotations of the 135 texts, the agreement was 0.8761, per chance agreement was 0.2466, and kappa was 0.8355. Annotator agreement did not differ as a function of text length ($\chi^2 = 1.27, p < 0.75$), arc length ($\chi^2 < 1$), or kind of coherence relation ($\chi^2 < 1$). Table 4 shows the confusion matrix for the database of 135 annotated texts that was used to compute the kappa statistics. The table shows, for example, that much of the interannotator disagreement seems to have been driven by disagreement over how to annotate *elaboration* relations (in the whole database, annotator 1 marked 260 elaboration relations where annotator 2 marked no relation; annotator 2 marked 467 elaboration relations where annotator 1 marked no relation).

The only other comparable discourse annotation project that we are currently aware of is that of Carlson, Marcu, and Okurowski (2002).² Since they use trees and split the annotation process into different substeps than those in our procedure, their annotator agreement figures are not directly comparable to ours. Furthermore, note that Carlson and her colleagues do not report annotator agreement figures for their database as a whole, but for different subsets of four to seven documents that were each annotated by different pairs of annotators. For discourse segmentation, they report kappa values ranging from 0.951 to 1.00; for annotation of discourse tree spans, their kappa values ranged from 0.778 to 0.929; for annotation of coherence relation nuclearity (whether a node in a discourse tree is a nucleus or a satellite, cf. section 2.3 for the definition of these terms), kappa values ranged from 0.695 to 0.882; for assigning types of coherence relations, they reported kappa values ranging from 0.624 to 0.823.

1 Note that interannotator agreement for step 3 was influenced by interannotator agreement for step 2. For example, one annotator might mark a group of discourse segments 2 and 3, whereas the second annotator might not mark that group of discourse segments. If the first annotator then marks, for example, a *cause-effect* coherence relation between discourse segment 4 and the group of discourse segments 2 and 3, whereas the second annotator marks a *cause-effect* coherence relation between discourse segment 4 and discourse segment 3, this would count as a disagreement. Thus, our measure of interannotator agreement for step 3 is conservative.

2 Note that Miltsakaki et al. (2004) report results on annotating connectives but not on annotating whole discourse structures.

Table 4

Confusion matrix of annotations for the database of 135 annotated texts. *contr* = contrast; *expv* = violated expectation; *ce* = cause-effect; *none* = no coherence relation; *gen* = generalization; *cond* = condition; *examp* = example; *ts* = temporal sequence; *attr* = attribution; *elab* = elaboration; *sim* = similarity.

		Annotator 2													
Annotator 1		<i>contr</i>	<i>expv</i>	<i>ce</i>	<i>none</i>	<i>gen</i>	<i>cond</i>	<i>examp</i>	<i>ts</i>	<i>attr</i>	<i>elab</i>	<i>same</i>	<i>sim</i>	Sum	Percentage
<i>contr</i>		383	11	0	34	0	0	0	2	0	0	0	0	430	4.47
<i>expv</i>		4	113	0	7	0	0	0	0	0	0	0	0	124	1.29
<i>ce</i>		0	0	446	14	0	0	0	0	0	5	0	0	465	4.83
<i>none</i>		66	24	42	0	0	2	27	16	6	467	1	64	715	7.43
<i>gen</i>		0	0	0	1	21	0	0	0	0	1	0	0	23	0.24
<i>cond</i>		0	0	0	2	0	127	0	1	0	1	0	0	131	1.36
<i>examp</i>		0	0	1	18	0	0	219	0	0	3	0	0	241	2.51
<i>ts</i>		1	1	2	7	0	0	0	214	0	1	0	0	226	2.35
<i>attr</i>		0	0	0	5	0	0	0	0	1,387	0	0	0	1,392	14.47
<i>elab</i>		0	0	17	260	0	3	0	3	0	3,913	1	0	4,197	43.63
<i>same</i>		0	0	2	5	0	0	0	1	0	0	530	1	539	5.60
<i>sim</i>		7	0	3	43	0	0	0	6	0	0	3	1,074	1,136	11.81
Sum		461	149	513	396	21	132	246	243	1,393	4,391	535	1,139		
Percentage		4.79	1.55	5.30	4.12	0.20	1.37	2.56	2.53	14.50	45.60	5.56	11.80		

3. Data Structures for Representing Coherence Relations

In order to represent the coherence relations between discourse segments in a text, most accounts of discourse coherence assume tree structures (Britton 1994; Carlson, Marcu, and Okurowski 2002; Corston-Oliver 1998; Longacre 1983; Grosz and Sidner 1986; Mann and Thompson 1988; Marcu 2000; Polanyi and Scha 1984; Polanyi 1996; Polanyi et al. 2004; van Dijk and Kintsch 1983; Walker 1998); some accounts do not allow crossed dependencies but appear to allow nodes with multiple parents (Lascarides and Asher 1991).³ Other accounts assume less constrained graphs that allow crossed dependencies as well as nodes with multiple parents (e.g., Bergler 1991; Birnbaum 1982; Danlos 2004; Hobbs 1985; McKeown 1985; Reichman 1985; Zukerman and McConachy 1995; for dialogue structure, Penstein Rose et al. 1995).

Some proponents of tree structures assume that trees are easier to formalize and to derive than less constrained graphs (Marcu 2000; Webber et al. 2003). We demonstrate that in fact many coherence structures in naturally occurring texts cannot be adequately represented by trees. Therefore we argue for less constrained graphs in which nodes represent discourse segments and labeled directed arcs represent the coherence relations that hold between these discourse segments as an appropriate data structure for representing coherence.

Some proponents of more general graphs argue that trees cannot account for a full discourse structure that represents informational, intentional, and attentional discourse relations. For example, Moore and Pollack (1992) point out that rhetorical structure theory (Mann and Thompson 1988) has both informational and intentional coherence relations but then forces annotators to decide on only one coherence relation between any two discourse segments. Moore and Pollack argue that often there is an informational as well as an intentional coherence relation between two discourse segments, which then presents a problem for RST, since only one of the relations can be annotated. Instead, Moore and Pollack propose allowing more than one coherence relation between two discourse segments, which violates the tree constraint of not having nodes with multiple parents.

Reichman (1985) argues that tree-based story grammars are not sufficient to account for discourse structure. Instead, she argues that in order to account for the intentional structure of discourse, more general data structures are needed. We argue that the same is true for the informational structure of discourse.

Moore and Pollack (1992), Moser and Moore (1996), and Reichman (1985) argue that trees are insufficient for representing informational, intentional, and attentional discourse structure. Note, however, that the focus of our work is on informational coherence relations, not on intentional relations. That does not mean that we think that attentional or intentional structure should not be part of a full account of discourse structure. Rather, we would like to argue that whereas the above accounts argue against trees for representing informational, intentional, and attentional discourse structure together, we argue that trees are not even descriptively adequate to describe just informational discourse structure by itself.

3 Although Lascarides and Asher (1991) do not explicitly disallow crossed dependencies, they argue that when a discourse structure is being constructed, the right frontier of an already existing discourse structure is the only possible attachment point for a new incoming discourse segment (cf. also Polanyi 1996; Polanyi and Scha 1984; Webber et al. 1999). This constraint on building discourse structures effectively disallows crossed dependencies.

Some accounts of informational discourse structure do not assume tree structures (e.g., Bergler [1991] and Hobbs [1985] for monologue and Penstein Rose et al. [1995] for dialogue structure). However, none of these accounts provides systematic empirical support for using more general graphs rather than trees. Providing a systematic empirical study of whether trees are descriptively adequate for representing discourse coherence is the goal of this article.

There are also accounts of informational discourse structure that argue for trees as a “backbone” for discourse structure but allow certain violations of tree constraints (crossed dependencies or nodes with multiple parents). Examples of such accounts include Webber et al. (1999) and Knott (1996). Similarly to our approach, Webber et al. (1999) investigated informational coherence relations. The kinds of coherence relations they used are basically the same as those that we used (cf. also Hobbs 1985). However, they argue for a tree structure as a backbone for discourse structure but have also addressed violations of tree structure constraints. In order to accommodate violations of tree structure constraints (in particular, crossed dependencies), Webber et al. (1999) argue for a distinction between “structural” discourse relations, on the one hand, and “nonstructural” or “anaphoric” discourse relations on the other hand. Structural discourse relations are represented within a lexicalized tree-adjointing grammar framework, and the resultant structural discourse structure is represented by a tree. However, more recently, Webber et al. (2003) have argued that structural discourse structure should allow nodes with multiple parents, but no crossed dependencies. It is unclear, however, why Webber et al. (2003) allow one kind of tree constraint violation (nodes with multiple parents) but not another (crossed dependencies).

Note that there seems to be a problem with the definition of “structural” versus “nonstructural” discourse structure in Webber et al. (1999): According to Webber et al. (1999), nonstructural discourse relations are licensed by anaphoric relations and can be involved in crossed dependencies. However, Webber et al. (1999) also argue that one criterion for nonstructural coherence relations is that they can cross (non)structural coherence relations. Since this definition of “nonstructural” appears to be circular, it is necessary to find an independent way to validate the difference between structural and nonstructural coherence relations. Knott (1996) might provide a way to empirically formalize the claims in Webber et al. (1999), or at least claims that seem to be very similar to those in Webber et al. (1999): Based on the observation that he cannot identify characteristic cue phrases for *elaboration* relations (e.g., *because* would be a characteristic cue phrase for *cause-effect*), Knott argues that *elaboration* relations are more permissive than other types of coherence relations (e.g., *cause-effect*, *similarity*, *contrast*). As a consequence, Knott argues, *elaboration* relations would be better described in terms of focus structures (cf. Grosz and Sidner 1986), which Knott argues are less constrained, than in terms of rhetorical relations (cf. Hobbs 1985; Mann and Thompson 1988), which Knott argues are more constrained. This hypothesis makes testable empirical claims: *Elaboration* relations should in some way pattern differently from other coherence relations. We come back to this issue in sections 4.1 and 4.2.

In this article we present evidence against trees as a data structure for representing discourse coherence. Note, though, that the evidence does not support the claim that discourse structures are completely arbitrary. The goal of our research program is to first determine which constraints on discourse structure are empirically viable. To us, the work we present here seems to be the crucial first step in avoiding arbitrary constraints on inferences for building discourse structures. In other words, the point we wish to make here is that although there might be other constraints on possible discourse annotations that will have to be identified in future research, tree structure constraints

do not seem to be the right kinds of constraints. This appears to be a crucial difference between approaches like Knott’s (1996), Marcu’s (2000), or Webber et al.’s (2003), on the one hand, and our approach, on the other hand. The goal of the former approaches seems to be to first specify a set of constraints on possible discourse annotations and then to annotate texts with these constraints in mind.

The following two sections illustrate problems with trees as a representation of discourse coherence structures. Section 3.1 shows that the discourse structures of naturally occurring texts contain crossed dependencies, which cannot be represented in trees. Another problem for trees, in addition to crossed dependencies, is that many nodes in coherence graphs of naturally occurring texts have multiple parents. This is shown in section 3.2. Because of these problems for trees, we argue for a representation such as chain graphs (cf. Frydenberg 1989; Lauritzen and Wermuth 1989), in which directed arcs represent asymmetrical or directed coherence relations and undirected arcs represent symmetrical or undirected coherence relations (this is equivalent to arguing for directed graphs with cycles). For all the examples in sections 3.1 and 3.2, chain-graph-based analyses are given. RST analyses are given only for those examples that are also annotated by Carlson, Marcu, and Okurowski (2002) (in those cases, the RST analyses are those provided by Carlson, Marcu, and Okurowski).

3.1 Crossed Dependencies

Consider the text passage in example (20) (modified from SAT practice materials):

- (20) 1. Schools tried to teach students history of science.
 2. At the same time they tried to teach them how to think logically and inductively.
 3. Some success has been reached in the first of these aims.
 4. However, none at all has been reached in the second.

Figure 1 shows the coherence graph for example (20). Note that the arrowheads of the arcs represent directionality for asymmetrical relations (*elaboration*) and bidirectionality for symmetrical relations (*similarity, contrast*).

The coherence structure for example (20) can be derived as follows:

- *Contrast* relation between discourse segments 1 and 2: Discourse segments 1 and 2 describe teaching different things to students.
- *Contrast* relation between discourse segments 3 and 4: Discourse segments 3 and 4 describe varying degrees of success (some vs. none).
- *Elaboration* relation between discourse segments 3 and 1: Discourse segment 3 provides more details (the degree of success) about the teaching described in discourse segment 1.

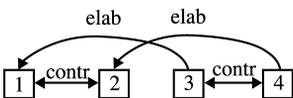


Figure 1
 Coherence graph for example (20). *contr* = *contrast*; *elab* = *elaboration*.

- *Elaboration* relation between discourse segments 4 and 2: Discourse segment 4 provides more details (the degree of success) about the teaching described in discourse segment 2.

In the resultant coherence structure for (20), there is a crossed dependency between $\{3, 1\}$ and $\{4, 2\}$.

In order to be able to represent a structure like the one for (20) in a tree without violating validity assumptions about tree structures (Diestel 2000), one might consider augmenting a tree either with feature propagation (Shieber 1986) or with a coindexation mechanism (Chomsky 1973). There is a problem, however, with both feature propagation and coindexation mechanisms: Both the tree structure itself and the features and coindexations as well represent the same kind of information (coherence relations). It is unclear how a dividing line could be drawn between tree structures and their augmentation. That is, it is unclear how one could decide which part of a text coherence structure should be represented by the tree structure and which part should be represented by the augmentation. Other areas of linguistics have faced this issue as well. Researchers investigating data structures for representing intrasentential structure, for instance, generally fall into two groups. One group tries to formulate principles that allow representation of some aspects of structure in the tree itself and other aspects in some augmentation formalism (e.g., Chomsky 1973; Marcus et al. 1994). Another group argues that it is more parsimonious to assume a unified dependency-based representation that drops the tree constraints of allowing no crossed dependencies (e.g., Brants et al. 2002; Skut et al. 1997; König and Lezius 2000). Our approach falls into the latter group. As we point out, there does not seem to be a well-defined set of constraints on crossed dependencies in discourse structures. Without such constraints, it does not seem viable to represent discourse structures as augmented tree structures.

An important question is how many different kinds of crossed dependencies occur in naturally occurring discourse. If there are only a very limited number of different structures with crossed dependencies in natural texts, one could make special provisions to account for these structures and otherwise assume tree structures. Example (20), for instance, has a listlike structure. It is possible that listlike examples are exceptional in natural texts. However, there are many other naturally occurring nonlistlike structures that contain crossed dependencies. As an example of a nonlistlike structure with a crossed dependency (between $\{4, 2\}$ and $\{3, 1-2\}$), consider example (21) (constructed):

- (21)
1. Susan wanted to buy some tomatoes
 2. and she also tried to find some basil
 3. because her recipe asked for these ingredients.
 4. The basil would probably be quite expensive at this time of the year.

The coherence structure for (21), shown in Figure 2, can be derived as follows:

- *Similarity* relation between 1 and 2: 1 and 2 both describe shopping for grocery items.
- *Cause-effect* relation between 3 and 1–2: 3 describes the cause for the shopping described by 1 and 2.

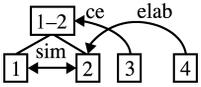


Figure 2
Coherence graph for example (21). *sim* = similarity; *ce* = cause-effect; *elab* = elaboration.

- *Elaboration* relation between 4 and 2: 4 provides details about the basil in 2.

Example (22), (from ap890109-0012; AP Newswire 1989 corpus [Harman and Liberman 1993]) has a similar structure:

- (22) 1. The flight Sunday took off from Heathrow Airport at 7:52pm
 2. and its engine caught fire 10 minutes later,
 3. the Department of Transport said.
 4. The pilot told the control tower he had the engine fire under control.

The coherence structure for example (22) can be derived as follows:

- *Temporal sequence* relation between 1 and 2: 1 describes the takeoff that happens before the engine fire described by 2 occurs.
- *Attribution* relation between 3 and 1-2: 3 mentions the source of what is said in 1-2.
- *Elaboration* relation between 4 and 2: 4 provides more detail about the engine fire in 2.

The resulting coherence structure, shown in Figure 3, contains a crossed dependency between {4, 2} and {3, 1-2}.

Consider example (23) (from wsj.0655; Wall Street Journal 1989 corpus [Harman and Liberman 1993]):

- (23) 1. _{1a}[Mr. Baker’s assistant for inter-American affairs,] _{1b}[Bernard Aronson,]
 2. while maintaining
 3. that the Sandinistas had also broken the cease-fire,
 4. acknowledged:
 5. “It’s never very clear who starts what.”

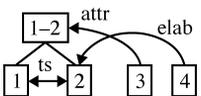


Figure 3
Coherence graph for example (22). *ts* = temporal sequence; *attr* = attribution; *elab* = elaboration.

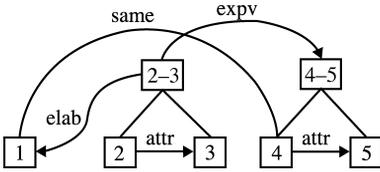


Figure 4
Coherence graph for example (23). *expv* = violated expectation; *elab* = elaboration; *attr* = attribution.

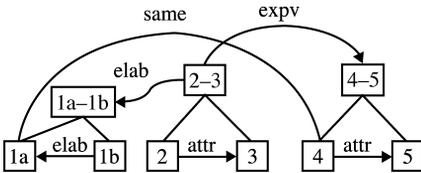


Figure 5
Coherence graph for example (23) with discourse segment 1 split into two segments. *expv* = violated expectation; *elab* = elaboration; *attr* = attribution.

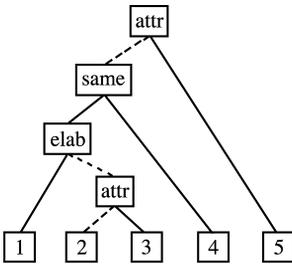


Figure 6
Tree-based RST annotation for example (23) from Carlson, Marcu, and Okurowski (2002). Broken lines represent the start of asymmetric coherence relations; continuous lines represent the end of asymmetric coherence relations; symmetric coherence relations have two continuous lines (cf. section 2.3). *attr* = attribution; *elab* = elaboration.

The annotations based on our annotation scheme with the discourse segmentation based on the segmentation guidelines in Carlson, Marcu, and Okurowski (2002) are presented in Figure 4, and those with the discourse segmentation based on our segmentation guidelines from section 2.1 are presented in Figure 5. Figure 6 shows a tree-based RST annotation for example (23) from Carlson, Marcu, and Okurowski (2002). The only difference between our approach and that of Carlson, Marcu, and Okurowski with respect to how example (23) is segmented is that Carlson and her colleagues assume discourse segment 1 to be one single segment. By contrast, based on our segmentation guidelines, discourse segment 1 would be segmented into two segments (because of the comma that does not separate a complex NP or VP), 1a and 1b, as indicated by the brackets in example (24):⁴

⁴ Based on our segmentation guidelines, the complementizer *that* in discourse segment 3 would be part of discourse segment 2 instead (cf. (15)). However, since this would not make a difference in terms of the resulting discourse structure, we do not provide alternative analyses with *that* as part of discourse segment 2 instead of discourse segment 3.

(24) _{1a}[Mr. Baker's assistant for inter-American affairs,] _{1b}[Bernard Aronson,]

The coherence structure for example (23) can be derived as follows:

- If discourse segment 1 is segmented into 1a and 1b (following our discourse segmentation guidelines), *elaboration* relation between 1a and 1b: 1b provides additional detail (a name) about what is stated in 1a (Mr. Baker's assistant).
- *Same* relation between 1 (or 1a) and 4: The subject NP in 1 (*Mr. Baker's assistant*) is separated from its predicate in 4 (*acknowledged*) by intervening discourse segments 2 and 3 (and 1b in our discourse segmentation).
- *Attribution* relation between 2 and 3: 2 states the source (the elided Mr. Baker) of what is stated in 3.
- *Elaboration* relation between the group of discourse segments 2 and 3 and discourse segment 1 (or the group of discourse segments 1a and 1b in our discourse segmentation): 2 and 3 state additional detail (a statement about a political process) about what is stated in 1 (or 1a and 1b) (Mr. Baker's assistant).
- *Attribution* relation between 4 (and by virtue of the *same* relation, also 1 or 1a) and 5: 4 states the source (Mr. Baker's assistant) of what is stated in 5.
- *Violated expectation* relation between the group of discourse segments 2 and 3 and the group of discourse segments 4 and 5: Although Mr. Baker's assistant acknowledges cease-fire violations by one side (discourse segments 2 and 3), he acknowledges that it is in fact difficult to clearly blame one side for cease-fire violations (discourse segments 4 and 5).

The resulting coherence structure, shown in Figure 5 (discourse segmentation from Carlson, Marcu, and Okurowski [2002]) and Figure 6 (our discourse segmentation), contains a crossed dependency: The *same* relation between discourse segment 1 and discourse segment 4 crosses the *violated expectation* relation between the group of discourse segments 2 and 3 and the group of discourse segments 4 and 5.

Figure 6 represents a tree-based RST annotation for example (23) from Carlson, Marcu, and Okurowski (2002); in Figure 6, dashed lines represent the start of asymmetric coherence relations and continuous lines mark the end of asymmetric coherence relations; symmetric coherence relations have two continuous lines (cf. section 2.3 for the distinction between symmetric and asymmetric coherence relations and for the directions of asymmetric coherence relations). Carlson, Marcu, and Okurowski (2002) do not provide descriptions of how they derived tree-based RST structures for their examples that are used in this article. Therefore, instead of discussing how the tree-based RST structures were derived, we show comparisons of the RST structure and our chain-graph-based structure; the comparison for (23) is provided in Table 5. Note in particular that the RST structure for example (23) does not represent the *violated expectation* relation between 2–3 and 4–5; that relation could not be annotated without violating the tree constraint of not allowing crossed dependencies.

Table 5

Comparison for example (23) of tree-based RST structure from Carlson, Marcu, and Okurowski (2002) and our chain-graph-based structure.

Tree-based RST structure	Our chain-graph-based structure
(1a and 1b are one discourse segment)	<i>Elaboration</i> between 1a and 1b
<i>Same</i> between 1–2 and 4	<i>Same</i> between 1 (or 1a) and 4
<i>Attribution</i> between 1 and 2	<i>Attribution</i> between 1 and 2
<i>Elaboration</i> between 2–3 and 1	<i>Elaboration</i> between 2–3 and 1 (or 1a and 1b)
<i>Attribution</i> between 1–4 and 5	<i>Attribution</i> between 4 and 5
(no relation)	<i>Violated expectation</i> between 2–3 and 4–5

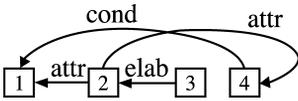


Figure 7

Coherence graph for example (25). *cond* = condition; *attr* = attribution; *elab* = elaboration.

3.2 Nodes with Multiple Parents

In addition to including crossed dependencies, many coherence structures of natural texts include nodes with multiple parents. Such nodes cannot be represented in tree structures. Consider example (25) (from ap890103 = 0014; AP Newswire 1989 corpus [Harman and Liberman 1993]).

- (25)
1. “Sure I’ll be polite,”
 2. promised one BMW driver
 3. who gave his name only as Rudolf.
 4. “As long as the trucks and the timid stay out of the left lane.”

The coherence structure for example (25) can be derived as follows:

- *Attribution* relation between 2 and 1 and 2 and 4: 2 states the source of what is stated in 1 and 4, respectively.
- *Elaboration* relation between 3 and 2: 3 provides additional detail (the name) about the BMW driver in 2.
- *Condition* relation between 4 and 1: 4 states the BMW driver’s condition for being polite, stated in 1.⁵ This *condition* relation is also indicated by the phrase “as long as.”

In the resultant coherence structure for example (25), node 1 has two parents—one *attribution* and one *condition* ingoing arc (cf. Figure 7).

⁵ A cultural reference: In Germany, when driving on a highway, it is only lawful to pass on the left side. Thus, Rudolf is essentially saying that he will be polite as long as the trucks and the timid do not keep him from passing other cars.

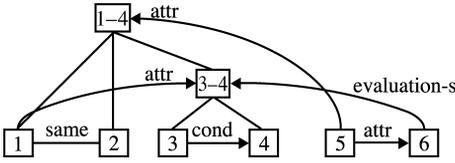


Figure 8

Coherence graph for example (26). Additional coherence relation used (from Carlson, Marcu, and Okurowski [2002]): *evaluation-s* = the situation presented in the satellite assesses the situation presented in the nucleus (*evaluation-s* would be *elaboration* in our annotation scheme). *attr* = *attribution*; *cond* = *condition*.

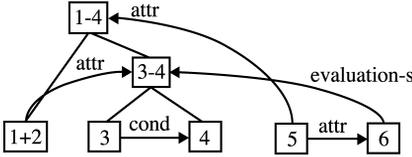


Figure 9

Coherence graph for example (26) with discourse segments 1 and 2 merged into one single discourse segment. Additional coherence relation used (from Carlson, Marcu, and Okurowski [2002]): *evaluation-s* = the situation presented in the satellite assesses the situation presented in the nucleus (*evaluation-s* would be *elaboration* in our annotation scheme). *attr* = *attribution*; *cond* = *condition*.

As another example of a discourse structure that contains nodes with multiple parents, consider the structure of example (26) (from wsj_0655; Wall Street Journal 1989 corpus [Harman and Liberman 1993]):

- (26) (they in 4 and 6 = Contra supporters; this is clear from the whole text wsj_0655)
1. "The administration should now state
 2. that
 3. if the February election is voided by the Sandinistas
 4. they should call for military aid,"
 5. said former Assistant Secretary of State Elliott Abrams.
 6. "In these circumstances, I think they'd win."

Our annotations are shown in Figures 8 (discourse segmentation from Carlson, Marcu, and Okurowski [2002]) and 9 (our discourse segmentation); Carlson et al.'s (2002) tree-based RST annotation is shown in Figure 10. The only difference between our annotation and that of Carlson, Marcu, and Okurowski is that we do not assume two separate discourse segments for 1 and 2; 1 and 2 are one discourse segment in our annotation (represented by the node 1+2 in Figure 9). Note also that in discourse segment 3 of example (23) "that" is not in a separate discourse segment; it is unclear why in example (26), "that" is in a separate discourse segment (discourse segment 2) and not part of discourse segment 3. The discourse structure for example (26) can be derived as follows:

1. According to our discourse segmentation guidelines (cf. section 2.1), 1 and 2 should be one single discourse segment: Therefore either *same* relation between 1 and 2 (cf. Figure 8), or merge 1 and 2 into one single discourse segment, 1+2 (cf. Figure 9).

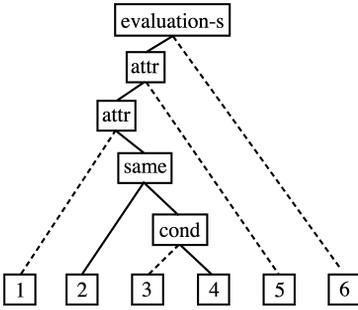


Figure 10

Tree-based RST annotation for example (26) from Carlson, Marcu, and Okurowski (2002). Broken lines represent the start of asymmetric coherence relations; continuous lines represent the end of asymmetric coherence relations; symmetric coherence relations have two continuous lines (cf. section 2.3). Additional coherence relation used (from Carlson, Marcu, and Okurowski [2002]): *evaluation-s* = the situation presented in the satellite assesses the situation presented in the nucleus (*evaluation-s* would be *elaboration* in our annotation scheme). *attr* = *attribution*; *cond* = *condition*.

2. *Attribution* relation between 1 or 1+2 and 3–4: 1 or 1+2 state the source (the administration) of what is stated in 3–4.
3. *Condition* relation between 3 and 4: 3 states the condition for what is stated in 4 (the *condition* relation is also signaled by the cue phrase *if* in 3).
4. *Attribution* relation between 5 and 1–4: 5 states the source of what is stated in 1–4.
5. *Attribution* relation between 5 and 6: 5 states the source of what is stated in 6.
6. *Evaluation-s*⁶ relation between 6 and 3–4: 3–4 state what is evaluated by 6—the Contra supporters should call for military aid, and if the February election is voided (group of discourse segments 3–4), the Contra supporters might win (discourse segment 6). Note that in our annotation scheme, the *evaluation-s* relation would be an *elaboration* relation (6 provides additional detail about 3–4: Elliott Abrams’s opinion on the Contras’ chances of winning).

In the resultant coherence structure for example (26), node 3–4 has multiple parents or ingoing arcs: one *attribution* ingoing arc and one *evaluation-s* ingoing arc (cf. Figures 8 and 9).

Table 6 presents a comparison of the RST annotation and our chain-graph-based annotation for (26). Note in particular that the *attribution* relation between 5 and 6 cannot be represented in the RST tree structure. Note furthermore that the RST tree contains an *evaluation-s* relation between 6 and 1–5. However, this *evaluation-s* relation seems to hold rather between 6 and 3–4: What is being evaluated is a chance for the Contras to win

6 The relation *evaluation-s* is part of the annotation scheme in Carlson, Marcu, and Okurowski (2002) but not part of our annotation scheme. In an *evaluation-s* relation, the situation presented in the satellite assesses the situation presented in the nucleus (Carlson, Marcu, and Okurowski 2002). An *evaluation-s* relation would be an *elaboration* relation in our annotation scheme.

Table 6

Comparison for (26) of tree-based RST structure (from Carlson, Marcu, and Okurowski (2002) and our chain-graph-based structure.

Tree-based RST structure	Our chain-graph-based structure
<i>Same</i> between 2 and 3–4	<i>Same</i> between 1 and 2, or merging of 1 and 2 to 1+2
<i>Attribution</i> between 1 and 2–4	<i>Attribution</i> between 1 or 1+2 and 3–4
<i>Condition</i> between 3 and 4	<i>Condition</i> between 3 and 4
<i>Attribution</i> between 5 and 1–4 (no relation)	<i>Attribution</i> between 5 and 1–4 <i>Attribution</i> between 5 and 6
<i>Evaluation-s</i> between 6 and 1–5	<i>Evaluation-s</i> between 6 and 3–4

a military conflict under certain circumstances. But a coherence relation between 6 and 3–4 could not have been annotated in a tree structure.

4. Statistics

We performed a number of statistical analyses on our annotated database to test our hypotheses. Each set of statistics was calculated for both annotators separately. However, since the statistics for both annotators were never different from each other (as confirmed by significant R^2 s > 0.9 or by χ^2 s > 1), we report only the statistics for one annotator in the following sections.

An important question is how frequent the phenomena discussed in the previous sections are. The more frequent they are, the more urgent the need for a data structure that can adequately represent them. The following sections report statistical results on crossed dependencies (section 4.1) and nodes with multiple parents (section 4.2).

4.1 Crossed Dependencies

The following sections report counts on crossed dependencies in the annotated database of 135 texts (cf. section 1). Section 4.1.1 reports results on the frequency of crossed dependencies, section 4.1.2 reports results concerning the question of what types of coherence relations tend to be involved in crossed dependencies, and section 4.1.3 reports results on the arc lengths of coherence relations involved in crossed dependencies. Section 4.1.4 provides a short summary of the statistical results on crossed dependencies.

4.1.1 Frequency of Crossed Dependencies. In order to track the frequency of crossed dependencies for the coherence structure graph of each text, we counted the minimum number of arcs that would have to be deleted in order to eliminate crossed dependencies in the coherence structure. Figure 11 illustrates this process. The example graph depicted in the figure contains the following crossed dependencies: $\{1, 3\}$ crosses with $\{2, 4\}$, $\{3, 5\}$ with $\{2, 4\}$, and $\{5, 7\}$ with $\{6, 8\}$. By deleting $\{2, 4\}$, two crossed dependencies can be eliminated: the crossing of $\{1, 3\}$ with $\{2, 4\}$ and the crossing of $\{3, 5\}$ with $\{2, 4\}$. By deleting either $\{5, 7\}$ or $\{6, 8\}$ the remaining crossed dependency between $\{5, 7\}$ and $\{6, 8\}$ can be eliminated. Therefore two edges would have to be deleted from the graph in Figure 11 in order to make it free of crossed dependencies.

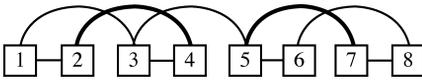


Figure 11
Example graph with crossed dependencies.

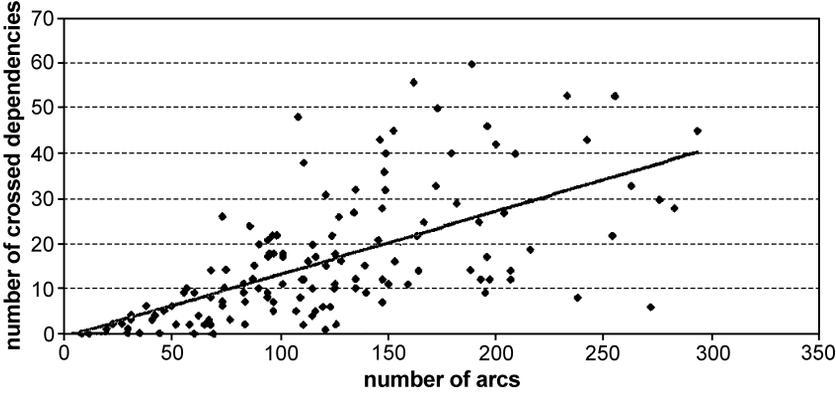


Figure 12
Correlation between number of arcs and number of crossed dependencies.

Table 7
Percentages of arcs to be deleted in order to eliminate crossed dependencies in the database texts.

Mean	12.5
Minimum	0
Maximum	44.4
Median	10.9

Table 7 shows the results of the counts. On average for the 135 annotated texts, 12.5% of arcs in a coherence graph have to be deleted in order to make the graph free of crossed dependencies. Seven texts out of the 135 had no crossed dependencies. The mean number of arcs for the coherence graphs of these texts was 36.9 (minimum: 8, maximum: 69, median: 35). The mean number of arcs for the other 128 coherence graphs (those with crossed dependencies) was 125.7 (minimum: 20, maximum: 293, median: 115.5). Thus, the graphs with no crossed dependencies had significantly fewer arcs than the graphs that had crossed dependencies ($\chi^2(1) = 15,330.35$ (Yates’s correction for continuity applied), $p < 10^{-6}$). This is a likely explanation for why these seven texts had no crossed dependencies.

More generally, linear regressions show a correlation between the number of arcs in a coherence graph and the number of crossed dependencies. The more arcs a graph has, the higher the number of crossed dependencies ($R^2 = 0.39$, $p < 10^{-4}$; cf. Figure 12). The same linear correlation holds between text length and number of crossed dependencies: The longer a text, the more crossed dependencies are in its coherence structure graph (for text length in discourse segments: $R^2 = .29$, $p < 10^{-4}$; for text length in words: $R^2 = .24$, $p < 10^{-4}$).

4.1.2 Types of Coherence Relations Involved in Crossed Dependencies. In addition to the question of how frequent crossed dependencies are, another question is whether

Table 8

Percentages of arcs to be deleted in order to eliminate crossed dependencies.

Coherence relation	Percentage of coherence relations participating in crossed dependencies	Percentage of overall coherence relations	Factor (= overall/crossed dependencies)
<i>Same</i>	1.13	17.21	15.23
<i>Condition</i>	0.05	0.28	5.59
<i>Attribution</i>	1.93	6.31	3.27
<i>Temporal sequence</i>	0.94	1.56	1.66
<i>Generalization</i>	0.24	0.34	1.40
<i>Contrast</i>	5.84	7.93	1.36
<i>Cause-effect</i>	1.13	1.53	1.35
<i>Violated expectation</i>	0.61	0.82	1.40
<i>Elaboration</i>	50.52	37.97	0.71
<i>Example</i>	4.43	3.15	1.34
<i>Similarity</i>	33.18	22.91	0.69

there are certain types of coherence relations that participate more or less frequently in crossed dependencies than other types of coherence relations. For an arc to participate in a crossed dependency, it must be in the set of arcs that would have to be deleted from a coherence graph in order to make that graph free of crossed dependencies (cf. the procedure outlined in section 4.1.1). In other words, the question is whether the frequency distribution over types of coherence relations is different for arcs participating in crossed dependencies compared to the overall frequency distribution over types of coherence relations in the whole database.

Figure 13 shows that the overall distribution over types of coherence relations participating in crossed dependencies is not different from the distribution over types of coherence relations overall. This is confirmed by the results of a linear regression, which show a significant correlation between the two distributions of percentages ($R^2 = 0.84$, $p < .0001$). Note that the overall distribution includes only arcs with length greater than one, since arcs of length one cannot participate in crossed dependencies.

However, there are some differences for individual coherence relations. Some types of coherence relations occur considerably less frequently in crossed dependencies than overall in the database. Table 8 shows the data from Figure 13 ranked by the factor of “percentage of overall coherence relations” by “percentage of coherence relations participating in crossed dependencies.” The proportion of *same* relations, for instance, is 15.23 times greater, and the percentage of *condition* relations is 5.59 times greater, overall in the database than in crossed dependencies. We do not yet understand the reason for these differences and plan to address this question in future research.

Another way of testing whether certain coherence relations contribute more than others to crossed dependencies is to remove coherence relations of a certain type from the database and then count the remaining number of crossed dependencies. For example, it is possible that the number of crossed dependencies is reduced once all *elaboration* relations are removed from the database. Table 9 shows that by removing all *elaboration* relations from the database of 135 annotated texts, the percentage of coherence relations involved in crossed dependencies is reduced from 12.5% to 4.96% of the remaining coherence relations. That percentage is reduced even further, to 0.84%, by removing all *elaboration* and *similarity* relations from the database. These numbers seem to be partial support for Knott’s (1996) hypothesis: Knott argued that *elaboration* relations are less

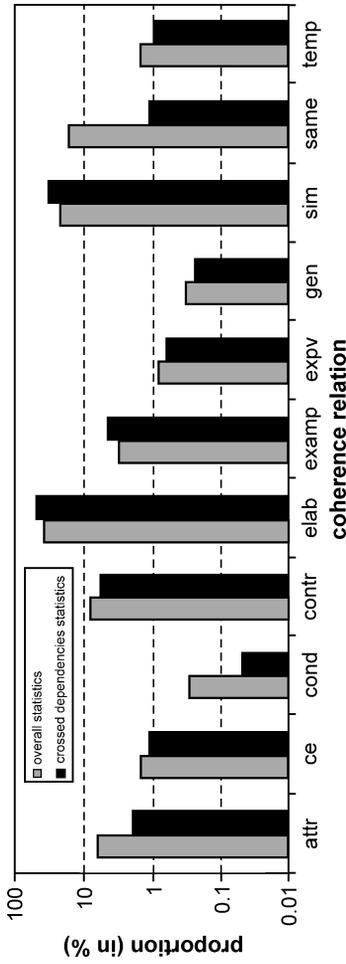


Figure 13

Distributions over types of coherence relations. For each condition (“overall statistics” and “crossed-dependencies statistics”), the sum over all coherence relations is 100; each bar in each condition represents a fraction of the total of 100 in that condition. The *y*-axis uses a \log_{10} scale. *attr* = attribution; *ce* = cause-effect; *cond* = condition; *contr* = contrast; *elab* = elaboration; *examp* = example; *expv* = Violated expectation; *gen* = generalization; *sim* = similarity; *ts* = temporal sequence.

Table 9

Effect of removing different types of coherence relations on the percentage of coherence relations involved in crossed dependencies.

Coherence relation removed	Remaining percentage of coherence relations involved in crossed dependencies			
	Mean	Min	Max	Median
<i>Same</i>	13.08	0	44.44	11.39
<i>Condition</i>	12.63	0	45.28	10.89
<i>Attribution</i>	13.44	0	44.86	11.36
<i>Temporal sequence</i>	12.53	0	44.44	10.87
<i>Generalization</i>	12.53	0	44.44	10.84
<i>Contrast</i>	11.88	0	46.15	9.86
<i>Cause-effect</i>	12.67	0	49.47	11.03
<i>Violated expectation</i>	12.51	0	44.44	10.87
<i>Elaboration</i>	4.96	0	47.47	1.23
<i>Example</i>	12.08	0	44.44	9.89
<i>Similarity</i>	7.32	0	24.56	7.04
<i>Elaboration and similarity</i>	0.84	0	10.68	0.00

constrained than other types of coherence relations (cf. the discussion of Knott [1996] in section 3).

However, there is a possible alternative hypothesis to Knott’s (1996). In particular, *elaboration* relations are very frequent (37.97% of all coherence relations; cf. Table 8). It is possible that removing *elaboration* relations from the database reduces the number of crossed dependencies only because a large number of coherence relations are removed when *elaborations* are removed. In other words, an alternative hypothesis to that of Knott (1996) is that the lower number of crossed dependencies is just due to less-dense coherence graphs (i.e., the less dense coherence graphs are, the lower the chance for crossed dependencies). We tested this hypothesis by correlating the percentage of coherence relations removed with the percentage of crossed dependencies that remain after removing a certain type of coherence relation.⁷ Figure 14 shows that the higher the percentage of removed coherence relations, the lower the percentage of coherence relations becomes that are involved in crossed dependencies. This correlation is confirmed by a linear regression ($R^2 = 0.7697, p < .0005$; after removing the *elaboration* data point: $R^2 = 0.4504, p < .05$; these linear regressions do not include the data point *elaboration + similarity*). Thus, although removing certain types of coherence relations reduces the number of crossed dependencies, it results in a very impoverished representation of coherence structure (i.e., after removing all *elaboration* and all *similarity* relations, only 39.12% of all coherence relations would still be represented [cf. Table 8]; the figure is 52.13% based on the distribution over coherence relations including those with absolute arc length one [cf. Table 11]).

With respect to Knott’s (1996) hypothesis, note that leaving out *elaboration* relations still leaves the proportion of remaining crossed dependencies at 4.96% (cf. Table 9).

⁷ Note that the percentages of removed coherence relations do not include coherence relations of absolute arc length one, since removing those coherence relations cannot have any influence on the number of crossed dependencies (coherence relations of absolute arc length one cannot be involved in crossed dependencies). Thus, the percentages of coherence relations removed in Figure 14 are from the third column of Table 8.

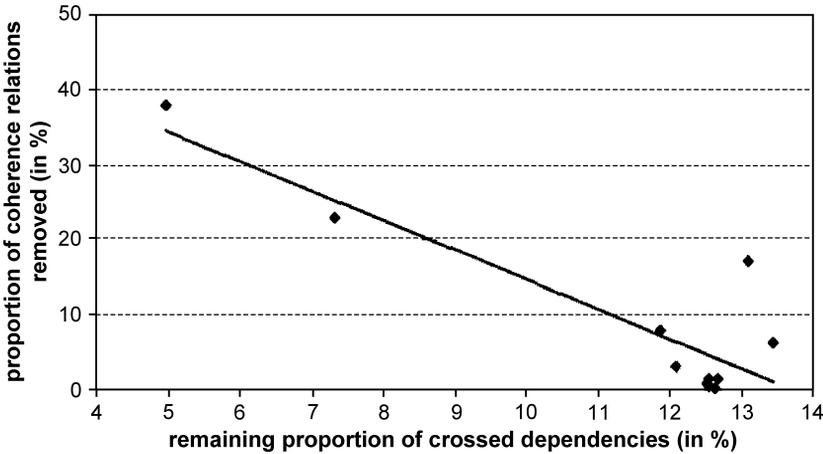


Figure 14

Correlation between removed percentage of overall coherence relations and remaining percentage of crossed dependencies. Note that the data point for *elaboration + similarity* is not included in the figure. $R^2 = 0.7699$, $p < .0005$.

In order to further reduce the proportion of remaining crossed dependencies, it is necessary to remove *similarity* relations in addition to removing *elaboration* relations (cf. Table 9). This is a pattern of results that is not predicted by any literature that we are aware of (including Knott [1996], among others, although he predicts these results partially). We believe this issue should be addressed in future research.

4.1.3 Arc Lengths of Coherence Relations Involved in Crossed Dependencies. Another question is how great the distance typically is between discourse segments that participate in crossed dependencies, or how great the arc length is for coherence relations that participate in crossed dependencies.⁸ It is possible, for instance, that crossed dependencies primarily involve long-distance arcs and that more local crossed dependencies are disfavored. However, Figure 15 shows that the distribution over arc lengths is practically identical for the overall database and for coherence relations participating in crossed dependencies (linear regression: $R^2 = 0.937$, $p < 10^{-4}$), suggesting a strong locality bias for coherence relations overall as well as for those participating in crossed dependencies.⁹ The arc lengths are normalized in order to take into account the varying length of texts. Normalized arc length is calculated by dividing the absolute length of an arc by the maximum length that that arc could have, given its position in its text. For example, if there is a coherence relation between discourse segment 1 and discourse segment 4 in a text, the raw distance between them would be three. If these discourse segments are part of a text that has five discourse segments total (i.e., 1 to 5),

⁸ The distance between two discourse segments is not measured in terms of how many coherence links one has to follow from any discourse segment x to any discourse segment y to which discourse segment x is related via a coherence relation. Instead, distance is measured in terms of the number of intervening discourse segments. Thus, distance between nodes reflects linear distance between two discourse segments in a text. For example, the distance between a discourse segment 1 and a discourse segment 4 would be three.

⁹ The arc length distribution for the database overall does not include arcs of (absolute) length one, since such arcs cannot participate in crossed dependencies.

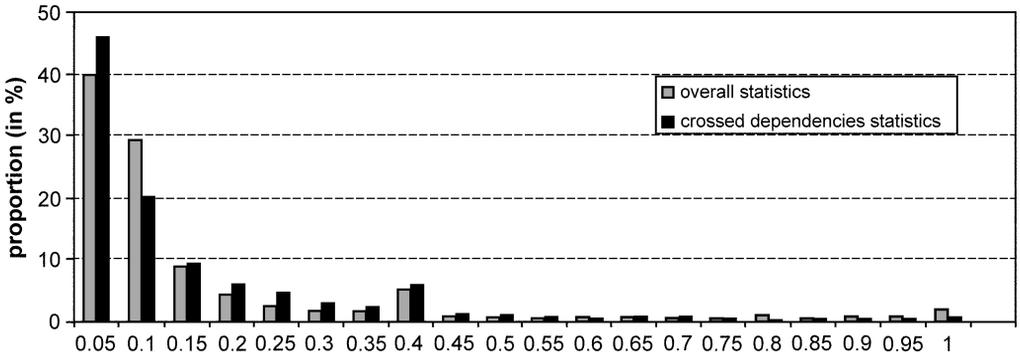


Figure 15
 Comparison of normalized arc length distributions. For each condition (“overall statistics” and “crossed-dependencies statistics”), the sum over all coherence relations is 100; each bar in each condition represents a fraction of the total of 100 in that condition.

the normalized distance would be $3/4 = 0.75$ (because four would be the maximum possible length of an arc that originates in discourse segment 1 or 4, given that the text has five discourse segments in total).

4.1.4 Summary of Crossed-Dependencies Statistics. Taken together, the statistical results on crossed dependencies suggest that crossed dependencies are too frequent to be ignored by accounts of coherence. Furthermore, the results suggest that any type of coherence relation can participate in a crossed dependency. However, there are some cases in which knowing the type of coherence relation that an arc represents can be informative as to how likely that arc is to participate in a crossed dependency. The statistical results reported here also suggest that crossed dependencies occur primarily locally, as evidenced by the distribution over lengths of arcs participating in crossed dependencies.

4.2 Nodes with Multiple Parents

Section 3.2 provided examples of coherence structure graphs that contain nodes with multiple parents. In addition to crossed dependencies, nodes with multiple parents are another reason why trees are inadequate for representing natural language coherence structures. The following sections report statistical results from our database on nodes with multiple parents. As in the previous section on crossed dependencies, we report results on the frequency of nodes with multiple parents (section 4.2.1), the types of coherence relations ingoing to nodes with multiple parents (section 4.2.2), and the arc length of coherence relations ingoing to nodes with multiple parents (section 4.2.3).

Table 10
 In-degree of nodes in the overall database.

Mean	1.60
Minimum	1
Maximum	12
Median	1

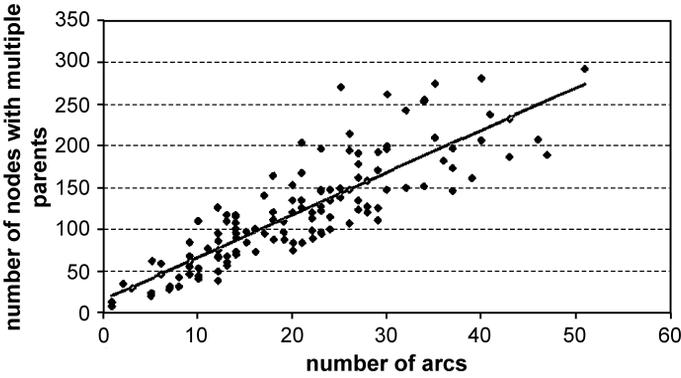


Figure 16
Correlation between number of arcs and number of nodes with multiple parents.

Section 4.2.4 provides a short summary of the statistical results on nodes with multiple parents.

4.2.1 Frequency of Nodes with Multiple Parents. We determined the frequency of nodes with multiple parents by counting the number of nodes with in-degree greater than one. We assume nodes with in-degree greater than one in a graph to be the equivalent of nodes with multiple parents in a tree. The results of our count indicated that 41.22% of all nodes in the database have an in-degree greater than one. In addition to counting the number of nodes with in-degree greater than one, we determined the mean in-degree of the nodes in our database. Table 10 shows that the mean in-degree (= mean number of parents) of all nodes in the investigated database of 135 texts is 1.6. As for coherence relations involved in crossed dependencies (cf. section 4.1.1), a linear regression showed a significant correlation between the number of arcs in a coherence graph and the number of nodes with multiple parents (cf. Figure 16; $R^2 = 0.7258$, $p < 10^{-4}$; for text length in discourse segments: $R^2 = .6999$, $p < 10^{-4}$; for text length in words: $R^2 = .6022$, $p < 10^{-4}$). The proportion of nodes with in-degree greater than one and the mean in-degree of the nodes in our database suggest that even if a mechanism could be derived for representing crossed dependencies in (augmented) tree graphs, nodes with multiple parents present another significant problem for trees representing coherence structures.

4.2.2 Types of Coherence Relations Ingoing to Nodes with Multiple Parents. As with crossed dependencies, an important question is whether there are certain types of coherence relations that are more or less frequently ingoing to nodes with multiple parents than other types of coherence relations. In other words, the question is whether the frequency distribution over types of coherence relations is different for arcs ingoing to nodes with multiple parents compared to the overall frequency distribution over types of coherence relations in the whole database. Figure 17 shows that the overall distribution over types of coherence relations ingoing to nodes with multiple parents is not different from the distribution over types of coherence relations overall.¹⁰ This is confirmed by the results of a linear regression, which show

¹⁰ Note that, unlike in section 4.1.2, the distribution over coherence relations for all coherence relations includes arcs with length one, since there was in this case no reason to exclude them.

Table 11
Proportion of coherence relations.

Coherence relation	Percentage of coherence relations ingoing to nodes with multiple parents	Percentage of overall coherence relations	Factor (= overall/ingoing to nodes with multiple parents)
<i>Attribution</i>	7.38	12.68	1.72
<i>Cause-effect</i>	2.63	4.19	1.59
<i>Temporal sequence</i>	1.38	2.11	1.53
<i>Condition</i>	0.83	1.21	1.46
<i>Violated expectation</i>	0.90	1.13	1.26
<i>Generalization</i>	0.17	0.21	1.22
<i>Contrast</i>	6.72	7.62	1.13
<i>Same</i>	10.72	9.74	0.91
<i>Similarity</i>	20.22	20.79	1.03
<i>Elaboration</i>	45.83	38.13	0.83
<i>Example</i>	3.20	2.19	0.68

a significant correlation between the two distributions of percentages ($R^2 = 0.967$, $p < 10^{-4}$).

Unlike for crossed dependencies (cf. Table 8), there are no big differences for individual coherence relations. Table 11 shows the data from Figure 17, ranked by the factor of “percentage of overall coherence relations” by “percentage of coherence relations ingoing to nodes with multiple parents.”

As for crossed dependencies, we also tested whether removing certain kinds of coherence relations reduced the mean in-degree (number of parents) and/or the percentage of nodes with in-degree greater than one (more than one parent). Table 12 shows that removing all *elaboration* relations from the database reduces the mean in-degree of nodes from 1.60 to 1.238 and the percentage of nodes with in-degree greater than one from 41.22% to 20.29%. Removing all *elaboration* as well as all *similarity* relations reduces these numbers further to 1.142 and 11.24%, respectively. As Table 12 also shows, removing other types of coherence relations does not lead to as great a reduction in the mean in-degree and the percentage of nodes with in-degree greater than one.

However, as with crossed dependencies (cf. section 4.1.2), we also tested whether the reduction in nodes with multiple parents could simply be due to removing more and more coherence relations (i.e., the less dense a graph is, the smaller the chance that there are nodes with multiple parents). We correlated the percentage of coherence relations removed with the mean in-degree of the nodes after removing different types of coherence relations.¹¹ Figure 18 shows that the higher the percentage of removed coherence relations, the lower the mean in-degree of the nodes in the database becomes. This correlation is confirmed by the results of a linear regression ($R^2 = 0.9455$, $p < 10^{-4}$; after removing the *elaboration* data point: $R^2 = 0.8310$, $p < .0005$; note that these linear regressions do not include the data point *elaboration + similarity*). We also correlated

¹¹ Note that in the correlations in this section, the proportions of removed coherence relations include coherence relations of absolute arc length one, because removing these coherence relations also has an effect on the mean in-degree of nodes and the proportion of nodes with in-degree greater than one. Thus, the proportions of coherence relations removed in Figure 18 and in Figure 19 are from the third column of Table 11.

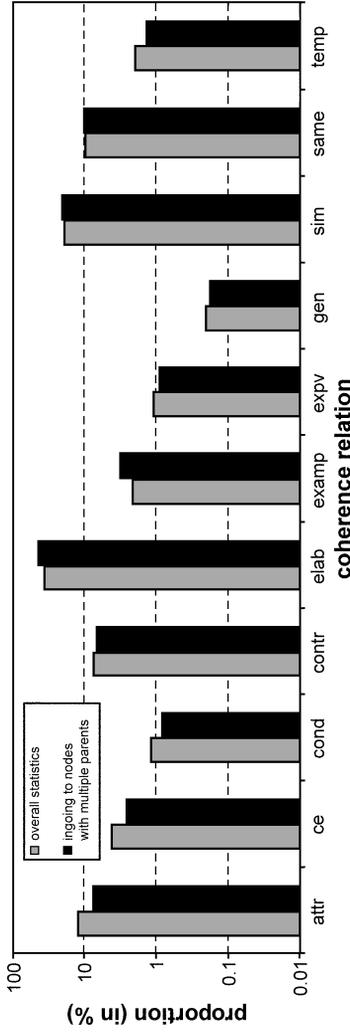


Figure 17

Distributions over types of coherence relations. For each condition (“overall statistics” and “ingoing to nodes with multiple parents”), the sum over all coherence relations is 100; each bar in each condition represents a fraction of the total of 100 in that condition. The y-axis uses a log₁₀ scale. *attr* = attribution; *ce* = cause-effect; *cond* = condition; *contr* = contrast; *elab* = elaboration; *examp* = example; *expv* = violated expectation; *gen* = generalization; *sim* = similarity; *ts* = temporal sequence.

Table 12

Effect of removing different types of coherence relations on the mean in-degree of nodes and on the percentage of nodes with in-degree greater than 1.

Coherence relation removed	In-degree of nodes				Percentage of nodes with in-degree > 1
	Mean	Min	Max	Median	
<i>Same</i>	1.519	1	12	1	35.85
<i>Condition</i>	1.599	1	12	1	41.01
<i>Attribution</i>	1.604	1	12	1	41.18
<i>Temporal sequence</i>	1.599	1	12	1	41.12
<i>Generalization</i>	1.600	1	12	1	41.16
<i>Contrast</i>	1.569	1	12	1	39.45
<i>Cause-effect</i>	1.599	1	12	1	41.14
<i>Violated expectation</i>	1.598	1	12	1	40.96
<i>Elaboration</i>	1.238	1	11	1	20.29
<i>Example</i>	1.574	1	11	1	40.37
<i>Similarity</i>	1.544	1	12	1	36.25
<i>Elaboration and similarity</i>	1.142	1	11	1	11.24

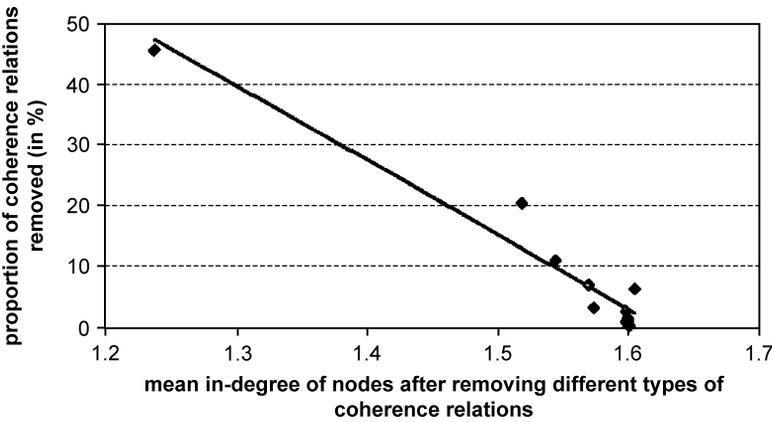


Figure 18

Correlation between percentage of removed coherence relations and mean in-degree of remaining nodes. Note that the data point for *elaboration + similarity* is not included in the figure. $R^2 = 0.9455, p < 10^{-4}$.

the percentage of coherence relations removed with the percentage of nodes with in-degree greater than one after removing different types of coherence relations. Figure 19 shows that the higher the percentage of removed coherence relations, the lower the percentage of nodes with in-degree greater than one. This correlation is also confirmed by the results of a linear regression ($R^2 = 0.9574, p < 10^{-4}$; after removing the *elaboration* data point: $R^2 = 0.8146, p < .0005$; note that these correlations do not include the data point *elaboration + similarity*).

Thus, although removing certain types of coherence relations (the same ones as for crossed dependencies, i.e., *elaboration* and *similarity*; cf. section 4.1.2) can reduce the mean in-degree of nodes and the proportion of nodes with in-degree greater than one, the result is a very impoverished coherence structure. For example, after removing both

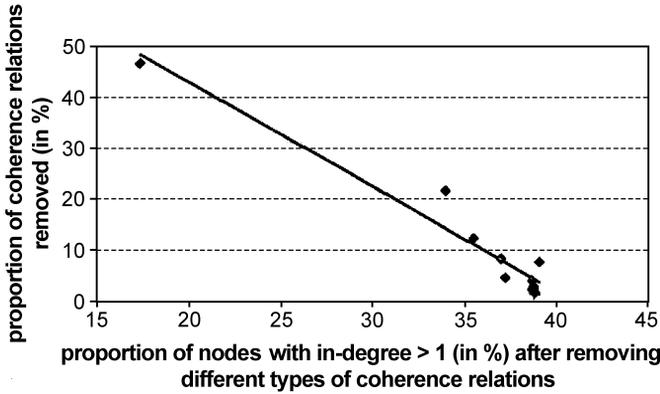


Figure 19

Correlation between percentage of removed coherence relations and percentage of nodes with in-degree > 1. Note that the data point for *elaboration + similarity* is not included in the figure. $R^2 = 0.9574$, $p < 10^{-4}$.

elaboration and *similarity* relations, only 52.13% of all coherence relations would still be represented (cf. Table 11). Furthermore, note that this pattern of results is not predicted by any literature we are aware of, including Knott (1996), although he predicts the results partially (he predicts that removing *elaboration* relations but not that removing *elaboration* as well as *similarity* relations is necessary in order to remove basically all nodes with multiple parents; cf. the discussion in the last paragraph of section 4.1.2). This issue will have to be investigated in future research.

4.2.3 Arc Lengths of Coherence Relations Ingoing to Nodes with Multiple Parents.

As for crossed dependencies, we also compared arc lengths. Here, we compared the length of arcs that are ingoing to nodes with multiple parents to the overall distribution of arc lengths. Again, we compared normalized arc lengths (see section 4.1.3 for the normalization procedure). By contrast to the comparison for crossed dependencies, we included in this comparison arcs of (absolute) length one, because such arcs can be ingoing to nodes with either single or multiple parents. Figure 20 shows that the distribution over arc lengths is practically identical for the overall database and for arcs ingoing to nodes with multiple parents (linear regression: $R^2 = 0.993$, $p < 10^{-4}$), suggesting a strong locality bias for coherence relations overall as well as for those participating in crossed dependencies.

4.2.4 Summary of Statistical Results on Nodes with Multiple Parents. In sum, the statistical results on nodes with multiple parents suggest that they are a frequent phenomenon and that they are not limited to certain kinds of coherence relations. However, as with crossed dependencies, removing certain kinds of coherence relations (*elaboration* and *similarity*) can reduce the mean in-degree of nodes and the proportion of nodes with in-degree greater than one. But also as with crossed dependencies, our data at present do not distinguish whether this reduction in nodes with multiple parents is due to a property of the coherence relations removed (*elaboration* and *similarity*) or whether it is just that removing more and more coherence relations simply reduces the chance for nodes to have multiple parents. We plan to address this question in future research. In addition to the results on frequency of nodes with multiple parents

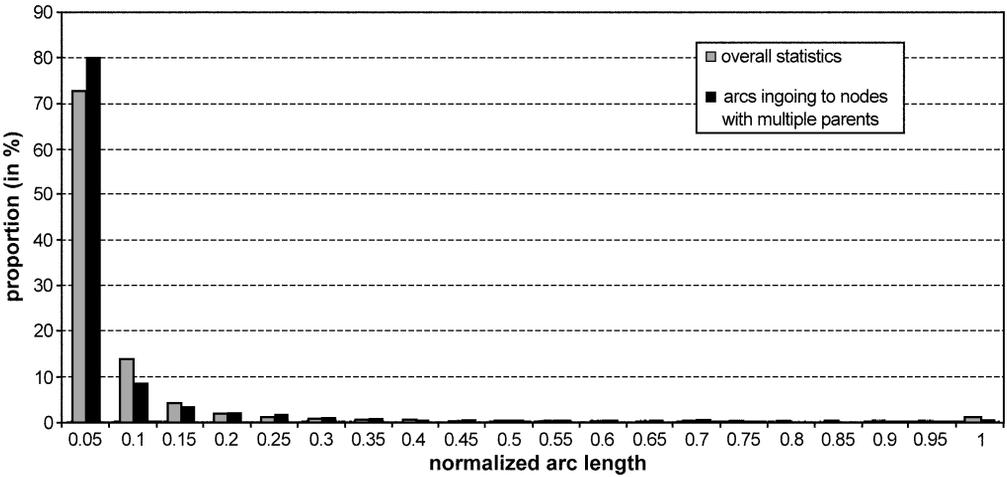


Figure 20
 Comparison of normalized arc length distributions. For each condition (“overall statistics” and “arcs ingoing to nodes with multiple parents”), the sum over all coherence relations is 100; each bar in each condition represents a fraction of the total of 100 in that condition.

and types of coherence relations ingoing to nodes with multiple parents, the statistical results reported here suggest that ingoing arcs to nodes with multiple parents are primarily local.

5. Conclusion

The goals of this article have been to present a set of coherence relations that are easy to code and to illustrate the inadequacy of trees as a data structure for representing discourse coherence structures. We have developed a coding scheme with high interannotator reliability and used that scheme to annotate 135 texts with coherence relations. An investigation of these annotations has shown that discourse structures of naturally occurring texts contain various kinds of crossed dependencies as well as nodes with multiple parents. Neither phenomenon can be represented using trees. This implies that existing databases of coherence structures that use trees are not descriptively adequate.

Our statistical results suggest that crossed dependencies and nodes with multiple parents are not restricted phenomena that could be ignored or accommodated with a few exception rules. Furthermore, even if one could find a way of augmenting tree structures to account for crossed dependencies and nodes with multiple parents, there would have to be a mechanism for unifying the tree structure with the augmentation features. Thus, in terms of derivational complexity, trees would just shift the burden from having to derive a less constrained data structure to having to derive a unification of trees and features or coindexation.

Because trees are neither a descriptively adequate data structure for representing coherence structures nor easier to derive, we argue for less constrained graphs as a data structure for representing coherence structures. In particular, we argue for a representation such as chain graphs (cf. final paragraph of section 3). Such less constrained graphs would have the advantage of being able to adequately represent coherence structures in one single data structure (cf. Brants et al. 2002; Skut et al. 1997; König and Lezius 2000).

Furthermore, they are at least not harder to derive than (augmented) tree structures. The greater descriptive adequacy might in fact make them easier to derive. However, this is still an open issue and will have to be addressed in future research.

In section 2.3 we briefly illustrated the possibility of more-fine-grained discourse segmentation than in the current project. Although such a detailed annotation of coherence relations was beyond the scope of the current project, future research should address this issue. More-fine-grained discourse segmentation could then also facilitate integration of discourse-level with sentence-level structural descriptions.

Another issue that should be addressed in future research is empirically viable constraints on inferences for building discourse structures. As pointed out in section 3, even though we have argued against trees as a data structure for representing discourse structures, that does not necessarily mean that discourse structures can be completely arbitrary. Future research should investigate questions such as whether there are structural constraints on coherence graphs (e.g., as proposed by Danlos [2004]) or whether there are systematic structural differences between the coherence graphs of texts that belong to different genres (e.g., as proposed by Bergler [1991]).

References

- Bergler, Sabine. 1991. The semantics of collocational patterns for reporting verbs. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, Germany.
- Birnbaum, Lawrence. 1982. Argument molecules: A functional representation of argument structures. In *Proceedings of the Third National Conference on Artificial Intelligence (AAAI-82)*, Pittsburgh, PA, pages 63–65.
- Brants, Sabine, Sabine Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The tiger treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.
- Britton, Bruce K. 1994. Understanding expository text. In Morton Ann Gernsbacher, editor, *Handbook of Psycholinguistics*. Academic Press, Madison, WI, pages 641–674.
- Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.
- Carlson, Lynn, Daniel Marcu, and Mary E. Okurowski. 2002. RST discourse treebank. Corpus number LDC 2002T07, Linguistic Data Consortium, Philadelphia.
- Chomsky, Noam, 1973. Conditions on transformations. In S. Anderson and P. Kiparsky, editors, *A Festschrift for Morris Halle*. Holt, Rinehart and Winston, New York, pages 232–286.
- Corston-Oliver, Simon. 1998. Computing representations of the structure of written discourse. Technical Report MSR-TR-98-15, Microsoft Research, Redmond, WA.
- Danlos, Laurence. 2004. Discourse dependency structures as dags. In *SigDIAL2004*, Cambridge, MA.
- Diestel, Reinhard. 2000. *Graph Theory*. Springer Verlag, New York.
- Frydenberg, Morten. 1989. The chain graph Markov property. *Scandinavian Journal of Statistics*, 17:333–353.
- Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
- Harman, Donna and Mark Liberman. 1993. Tipster complete. Corpus number LDC93T3A, Linguistic Data Consortium, Philadelphia.
- Hearst, Marti. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.
- Hirschberg, Julia and Christine H. Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, Santa Cruz, CA.
- Hobbs, Jerry R. 1985. On the coherence and structure of discourse. Technical Report 85-37, Center for the Study of Language and Information (CSLI), Stanford, CA.
- Hobbs, Jerry R., Martin E. Stickel, Douglas E. Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.
- Hovy, Eduard and Elisabeth Maier. 1995. Parsimonious or profligate: How many and which discourse relations? Technical report, University of Southern California.

- Kehler, Andrew. 2002. *Coherence, Reference, and the Theory of Grammar*. Stanford University Press, Stanford, CA.
- Knott, Alistair. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh.
- König, Esther and Wolfgang Lezius. 2000. A description language for syntactically annotated corpora. In *Proceedings of the Computational Linguistics Conference (COLING)*, pages 1056–1060, Saarbrücken, Germany.
- Lascarides, Alex and Nicholas Asher. 1991. Discourse relations and defeasible knowledge. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 55–63, Berkeley, CA.
- Lascarides, Alex and Nicholas Asher. 1993. Temporal interpretation, discourse relations and common sense entailment. *Linguistics and Philosophy*, 16(5): 437–493.
- Lauritzen, Steffen and Nanny Wermuth. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17:31–57.
- Longacre, Robert E. 1983. *The Grammar of Discourse*. Plenum, New York.
- Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.
- Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA.
- Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, NJ. San Francisco, CA. Morgan Kaufmann.
- McKeown, Kathleen R. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, Cambridge.
- Miltsakaki, Eleni, Rashmi Prasad, Aravind K. Joshi, and Bonnie L. Webber. 2004. The Penn discourse treebank. In *Proceedings of the Language and Resources and Evaluation Conference*, Lisbon.
- Moore, Johanna D. and Martha E. Pollack. 1992. A problem for rst: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.
- Moser, Megan and Johanna D. Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.
- Penstein Rose, Carolyn, Barbara Di Eugenio, Lori S. Levin, and Carol Van Ess-Dykema. 1995. Discourse processing of dialogues with multiple threads. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA.
- Polanyi, Livia. 1996. The linguistic structure of discourse. Technical Report 96-118, Center for the Study of Language and Information (CSLI), Stanford, CA.
- Polanyi, Livia, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004. A rule based approach to discourse parsing. In *SigDIAL 2004*, Cambridge, MA.
- Polanyi, Livia and Remko Scha. 1984. A syntactic approach to discourse semantics. In *Proceedings of the 10th International Conference on Computational Linguistics*, Stanford, CA.
- Reichman, Rachel. 1985. *Getting Computers to Talk Like You and Me*. MIT Press, Cambridge, MA.
- Shieber, Stuart M. 1986. *An introduction to unification-based approaches to grammar*. Lecture Notes 4, Center for the Study of Language and Information (CSLI), Stanford, CA.
- Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*, Washington, DC.
- van Dijk, Teun A. and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. Academic Press, New York.
- Walker, Marilyn A. 1998. Centering, anaphora resolution, and discourse structure. In E. Prince, A. K. Joshi, and M. A. Walker, editors, *Centering Theory in Discourse*. Oxford University Press, Oxford, pages 401–435.
- Webber, Bonnie L., Alistair Knott, Matthew Stone, and Aravind K. Joshi. 1999. Discourse relations: A structural and presuppositional account using lexicalized tag. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, College Park, MD, pages 41–48.

Webber, Bonnie L., Matthew Stone, Aravind K. Joshi, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Wolf, Florian, Edward Gibson, Amy Fisher, and Meredith Knight. 2003. A procedure for collecting a database of texts annotated with coherence relations. Technical report,

Massachusetts Institute of Technology, Cambridge, MA.
Zukerman, Ingrid and Richard McConachy. 1995. Generating discourse across several user modules: Maximizing belief while avoiding boredom and overload. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 1251–1257, Montreal.