The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach

Daniel Marcu[*]
Information Sciences Institute, USC

Coherent texts are not just simple sequences of clauses and sentences, but rather complex artifacts that have highly elaborate rhetorical structure. This paper explores the extent to which well-formed rhetorical structures can be automatically derived by means of surface-form-based algorithms. These algorithms identify discourse usages of cue phrases and break sentences into clauses, hypothesize rhetorical relations that hold among textual units, and produce valid rhetorical structure trees for unrestricted natural language texts. The algorithms are empirically grounded in a corpus analysis of cue phrases and rely on a first-order formalization of rhetorical structure trees.

The algorithms are evaluated both intrinsically and extrinsically. The intrinsic evaluation assesses the resemblance between automatically and manually constructed rhetorical structure trees. The extrinsic evaluation shows that automatically derived rhetorical structures can be successfully exploited in the context of text summarization.

1. Motivation

Consider the text given in (1), which was taken from *Scientific American*, November 1996.

(1) With its distant orbit—50 percent farther from the sun than Earth—and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about −60 degrees Celsius (−76 degrees Fahrenheit) at the equator and can dip to −123 degrees C near the poles. Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure.

Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian weather involves blowing dust or carbon dioxide. Each winter, for example, a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap. Yet even on the summer pole, where the sun remains in the sky all day long, temperatures never warm enough to melt frozen water.

A rhetorical structure representation (tree) of its first paragraph is shown in Figure 1. In the rhetorical representation, which employs the conventions proposed by Mann and Thompson (1988), each leaf of the tree is associated with a contiguous textual

[*] Information Sciences Institute, University of Southern California, 4676 Admiralty Way, Marina del Rey, CA 90292-6601

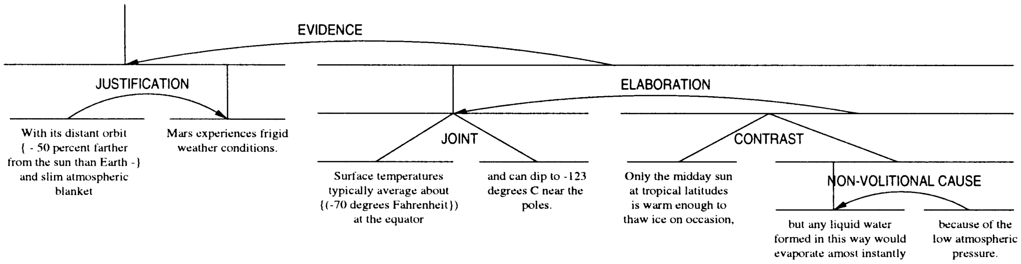
EVIDENCE
JUSTIFICATION
ELABORATION
With its distant orbit { - 50 percent farther from the sun than Earth -} and slim atmospheric blanket
Mars experiences frigid weather conditions.
JOINT
CONTRAST
Surface temperatures typically average about {(-70 degrees Fahrenheit}) at the equator
and can dip to -123 degrees C near the poles.
Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,
NON-VOLITIONAL CAUSE
but any liquid water formed in this way would evaporate amost instantly
because of the low atmospheric pressure.

Figure 1
A rhetorical structure representation of the first paragraph in text (1).

span. The internal nodes are labeled with the names of the rhetorical relations that hold between the textual spans that are subsumed by their child nodes. Each relation between two nodes is represented graphically by means of a combination of straight lines and arcs. The material subsumed by the text span that corresponds to the starting point of an arc is subsidiary to the material subsumed by the text span that corresponds to the end point of an arc. A relation represented only by straight lines corresponds to cases in which the subsumed text spans are equally important. Text spans that subsume subsidiary information, i.e., text spans that correspond to starting points of arcs, are called **satellites**. All other text spans are called **nuclei**. Text fragments surrounded by curly brackets denote parenthetical units: their deletion does not affect the understanding of the textual unit to which they belong.

For example, the textual unit *Mars experiences frigid weather conditions* is at the end of an arc that originates from the textual unit *With its distant orbit—50 percent farther from the sun than Earth—and slim atmospheric blanket* because the former represents something that is more essential to the writer's purpose than the latter and because the former can be understood even if the subsidiary span is deleted, but not vice versa. The satellite information JUSTIFIES in this case the writer's right to present the information in the nucleus. The text spans *Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,* and *but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure* are connected by straight lines because they are equally important with respect to the writer's purpose; they correspond to the elements of a CONTRAST relation. The text fragment—*50 percent farther from the sun than Earth*—is surrounded by curly brackets because it is parenthetical.

Traditionally, it has been assumed that rhetorical structures of the kind shown in Figure 1 can be built only if one understands fully the semantics of the text and the intentions of the writer. To understand, for example, that the information given in the last two sentences of the first paragraph of text (1) is EVIDENCE to the information given in the first sentence, one needs to understand that the last two sentences may increase the reader's belief of the first sentence. And to understand that it was the *low atmospheric pressure* that caused the *liquid water to evaporate*, one needs to understand that without the information presented in the satellite, the reader may not know the particular CAUSE of the situation presented in the nucleus.

In spite of the large number of discourse-related theories that have been proposed so far, there have emerged no algorithms capable of deriving the discourse structure of free, unrestricted texts. On one hand, the theories developed in the traditional, truth-based semantic perspective (Kamp 1981; Lascarides and Asher 1993; Asher 1993; Hobbs et al. 1993; Kamp and Reyle 1993; Asher and Lascarides 1994; Kameyama

1994; Polanyi and van den Berg 1996; van den Berg 1996; Gardent 1997; Schilder 1997; Cristea and Webber 1997; Webber et al. l999) take the position that discourse structures can be built only in conjunction with fully specified clause and sentence syntactic structures. These theories have a grammar as their backbone and rely on sophisticated logics of belief and default logics in order to intertwine and characterize the sentence- and discourse-based linguistic phenomena. Despite their formal elegance, implementations of these theories cannot yet handle naturally occurring texts, such as that shown in (1). On the other hand, the theories aimed at characterizing the constraints that pertain to the structure of unrestricted texts and the computational mechanisms that would enable the derivation of these structures (van Dijk 1972; Zock 1985; Grosz and Sidner 1986; Mann and Thompson 1988; Polanyi 1988, 1996; Hobbs 1990) are either too informal or incompletely specified to support a fully automatic approach to discourse analysis.

In this paper, I explore the ground found at the intersection of these two lines of research. More precisely, I explore the extent to which rhetorical structures of the kind shown in Figure 1 can be built automatically by relying only on cohesion and connectives, i.e., phrases such as *for example, and, although,* and *however* that are used "to link linguistic units at any level" (Crystal 1991, 74).[1] The results show that although cohesion and connectives are ambiguous indicators of rhetorical structure, when used in conjunction with a well-constrained mathematical model of valid rhetorical structures, they enable the implementation of surprisingly accurate rhetorical parsers.

2. Foundation

The hypothesis that underlies this work is that connectives, cohesion, shallow processing, and a well-constrained mathematical model of valid rhetorical structure trees (RS-trees) can be used to implement algorithms that determine

- the elementary units of a text, i.e., the units that constitute the leaves of the RS-tree of that text;
- the rhetorical relations that hold between elementary units and between spans of text;
- the relative importance (nucleus or satellite) and the size of the spans subsumed by these rhetorical relations.

In what follows, I examine each facet of this hypothesis intuitively and explain how it contributes to the derivation of a rhetorical parsing algorithm, i.e., an algorithm that takes as input free, unrestricted text and that determines its valid RS-trees. For each facet, I consider first the arguments that support the hypothesis and then discuss potential difficulties.

2.1 Determining the Elementary Units Using Connectives and Shallow Processing

2.1.1 Pro Arguments. Recent developments in the linguistics of punctuation (Nunberg 1990; Briscoe 1996; Pascual and Virbel 1996; Say and Akman 1996; Shiuan and Ann 1996) have emphasized the role that punctuation can have in solving a variety of natural language processing tasks ranging from syntactic parsing to information

[1] In this paper, I use the terms *connective* and *cue phrase* interchangeably. And I use the term *discourse marker* to refer to a connective that has a discourse function, i.e., a connective that signals a rhetorical relation that holds between two text spans.

packaging. For example, if a sentence consists of three arguments separated by semicolons, it is likely that one can determine the boundaries of these arguments without relying on sophisticated forms of syntactic analysis. Shallow processing is sufficient to recognize the occurrences of the semicolons and to break the sentence into three elementary units.

In a corpus study (described in Section 3), I noticed that in most of the cases in which a connective such as *Although* occurred at the beginning of a sentence, it marked the left boundary of an elementary unit whose right boundary was given by the first subsequent occurrence of a comma. Hence, it is likely that by using only shallow techniques and knowledge about connectives, one can determine, for example, that the elementary units of sentence (2) are those enclosed within square brackets.

(2) [*Although* Brooklyn College does not yet have a junior-year-abroad program,] [a good number of students spend summers in Europe.]

2.1.2 Difficulties. Obviously, by relying only on orthography, connectives, and shallow processing it is unlikely that one will be capable of correctly determining all elementary units of an RS-tree. It may very well be the case that knowledge about how *Although* is used in texts can be exploited to determine the elementary units of texts, but not all connectives are used as consistently as *Although* is. Just consider, for instance, the highly ambiguous connective *and*. In some cases, *and* plays a sentential, syntactic role, while in others, it plays a discourse role, i.e., it signals a rhetorical relation that holds between two textual units. For example, in sentence (3), the first *and* is sentential, i.e., it makes a semantic contribution to the interpretation of the complex noun phrase "John *and* Mary", while the second *and* has a discourse function, i.e., it signals a rhetorical relation of SEQUENCE that holds between the units enclosed within square brackets.

(3) [John *and* Mary went to the theatre] [*and* saw a nice play.]

If a system is to use connectives to determine elementary unit boundaries, it would need to figure out that a boundary is required before the second occurrence of *and* (the occurrence that has a discourse function), but not before the first occurrence. It seems clear that shallow processing is insufficient to properly solve this problem. It remains an open question, however, to what degree shallow processing and knowledge about connectives can be successfully used to determine the elementary units of texts. Our results show (see Section 4), that using only such lean knowledge resources, elementary unit boundaries can be determined with approximately 80% accuracy.

2.2 Determining Rhetorical Relations Using Connectives

2.2.1 Pro Arguments. The intuition behind this choice relies on the following facts:

- Linguistic and psycholinguistic research has shown that connectives are consistently used by humans both as cohesive ties between adjacent clauses and sentences (Halliday and Hasan 1976) and as "macroconnectors" that signal relations that hold between large textual units. For example, in stories, connectives such as *so*, *but*, and *and* mark boundaries between story parts (Kintsch 1977). In naturally occurring conversations, *so* marks the terminal point of a main discourse unit and a potential transition in a participant's turn, whereas *and* coordinates idea units and continues a speaker's action (Schiffrin 1987). In narratives, connectives signal structural relations between elements and are crucial for the understanding of the stories (Segal and Duchan 1997). In general, cue phrases are used consistently by both speakers and writers to highlight the most important shifts in their narratives, mark intermediate breaks, and signal areas of topical continuity (Bestgen and Costermans 1997; Schneuwly 1997). Therefore, it is likely that connectives can be used to determine rhetorical relations that hold both between elementary units and between large spans of text.

- The number of discourse markers in a typical text—approximately one marker for every two clauses (Redeker 1990)—is sufficiently large to enable the derivation of rich rhetorical structures for texts.[2] More importantly, the absence of markers correlates with a preference of readers to interpret the unmarked textual units as continuations of the topics of the units that precede them (Segal, Duchan, and Scott 1991). Hence, when there is no connective between two sentences, for example, it is likely that the second sentence elaborates on the first.

2.2.2 Difficulties. The above arguments tell us that connectives are used often and that they signal relations that hold both between elementary units and large spans of texts. Hence, previous research tells us only that connectives *are potentially* useful in determining the rhetorical structure of texts. Unfortunately, they cannot be used straightforwardly because they are ambiguous.

- In some cases, connectives have a sentential function, while in other cases, they have a discourse function. Unless we can determine when a connective has a discourse function, we cannot use connectives to hypothesize rhetorical relations.
- Connectives do not explicitly signal the size of the textual spans that they relate.
- Connectives can signal more than one rhetorical relation. That is, there is no one-to-one mapping between the use of connectives and the rhetorical relations that they signal.

I address each of these three problems in turn.

Sentential and Discourse Uses of Connectives. Empirical studies on the disambiguation of cue phrases (Hirschberg and Litman 1993) have shown that just by considering the orthographic environment in which they occur, one can distinguish between sentential and discourse uses in about 80% of cases and that these results can be improved with machine learning techniques (Litman 1996) or genetic algorithms (Siegel and McKeown 1994). I have taken Hirschberg and Litman's research one step further and designed a comprehensive corpus analysis of cue phrases that enabled me to design algorithms that improved their results and coverage. The corpus analysis is discussed in Section 3. The algorithm that determines elementary unit boundaries and identifies discourse uses of cue phrases is discussed in Section 4.

[2] A corpus of instructional texts that was studied by Moser and Moore (1997) and Di Eugenio, Moore, and Paolucci (1997) reflected approximately the same distribution of cue phrases: 181 of the 406 discourse relations that they analyzed were cued relations.

Discourse Markers are Ambiguous with Respect to the Size of the Spans They Connect. Assume, for example, that a computer is supposed to determine, using only surface-form algorithms and knowledge about connectives, the rhetorical relation that is signaled by the marker *In contrast*, in text (4).

(4) [John likes sweets.[1]] [Most of all, John likes ice cream and chocolate.[2]] [*In contrast*, Mary likes fruit.[3]] [*Especially* bananas and strawberries.[4]]

During the corpus study that I discuss in Section 3, I noticed that in all its occurrences in a sample of texts, the connective *In contrast* signaled a CONTRAST relation. Hence, it is likely that *In contrast* signals a CONTRAST relation in text (4) as well. Unfortunately, although we know what relation *In contrast* signals, we do not know which spans the CONTRAST relation holds between: does the relation hold between spans [1,2] and [3,4]; or between unit 2 and span [3,4]; or between span [1,2] and unit 3; or between units 1 and 3; or between other units and spans? The best that we can do in this case is to make an exclusively disjunctive hypothesis, i.e., to hypothesize that one and only one of these possible relations holds. However, it is still unclear what the elements of such an exclusively disjunctive hypothesis should be.

In my previous work (Marcu 1996, 1997a, 1997b, 1999b, 2000), I have argued that rhetorical relations that hold between large textual spans can be explained in terms of similar relations that hold between their most important elementary units. For example, the rhetorical relation of EVIDENCE that holds between the first sentence of the first paragraph in text (1) and the last two sentences of the same paragraph can be explained in terms of a similar relation that holds between the corresponding nuclei: an EVIDENCE relation also holds between the nucleus of the first sentence, *Mars experiences frigid weather conditions* and each of the most important nuclei of the last two sentences *Surface temperatures typically average about* -60 *degrees Celsius* (-76 *degrees Fahrenheit) at the equator* and *[Surface temperatures] can dip to* -123 *degrees C near the poles.* Similarly, the CONTRAST relation that holds between the two spans [1,2] and [3,4] in text (4) can be explained in terms of a CONTRAST relation that holds between units 1 and 3, the most important units in spans [1,2] and [3,4], respectively.

The fact that rhetorical relations that hold between large textual spans can be explained/determined in terms of rhetorical relations that hold between elementary textual units suggests that rhetorical structure trees can be constructed in a bottom-up fashion, from rhetorical relations that have been determined to hold between elementary textual units. Hence, to derive the rhetorical structure of text (4) it is sufficient to hypothesize with respect to the occurrence of the connective *In contrast*, the exclusively disjunctive hypothesis $\mathit{rhet\_rel}(\textsc{contrast}, 1, 3) \oplus \mathit{rhet\_rel}(\textsc{contrast}, 1, 4) \oplus \mathit{rhet\_rel}(\textsc{contrast}, 2, 3) \oplus \mathit{rhet\_rel}(\textsc{contrast}, 2, 4)$, because this hypothesis subsumes all the other possible rhetorical relations that may be signaled by the connective.[3] In Section 2.4, I will explain why exclusive-disjunctive hypotheses of this kind are sufficient for determining the rhetorical structure of texts.

The fact that rhetorical relations that hold between large spans can be explained in terms of rhetorical relations that hold between elementary units should not lead one to conclude that a computational system should only make rhetorical hypotheses whose arguments are elementary units. For example, a text fragment may consist of three paragraphs, clearly marked by the connectives *First, Second,* and *Third*. For such

[3] Throughout this paper, I use the convention that rhetorical relations are represented as sorted, first-order predicates having the form *rhet_rel*(NAME, SATELLITE, NUCLEUS) in the case of hypotactic relations and the form *rhet_rel*(NAME, NUCLEUS, NUCLEUS) in the case of paratactic relations.

a fragment, it is likely that the three paragraphs are in a LIST or SEQUENCE relation. If a computer program exploits the occurrence of these markers, it may be able to derive the high-level rhetorical structure of the text fragment without determining the important units and relations that underlie the three paragraphs. The work presented in this paper acknowledges the utility of dealing both with *simple* relations, i.e., rhetorical relations that hold between elementary textual units, and with *extended* rhetorical relations, i.e., relations that hold between large segments. Depending on the circumstances, a computational system will have to choose the types of relations it should hypothesize to determine the rhetorical structure of a text.

The observation that rhetorical relations that hold between large textual spans can be explained in terms of rhetorical relations that hold between elementary textual units and the need for dealing with extended rhetorical relations amount to providing a **compositionality criterion** for valid rhetorical structures. This criterion posits that a rhetorical structure tree is valid only if each rhetorical relation that holds between two spans is either an extended rhetorical relation or can be explained in terms of a simple rhetorical relation.

Discourse Markers are Ambiguous with Respect to the Rhetorical Relations They Signal. Discourse markers are also ambiguous with respect to the rhetorical relations they signal and the importance of the textual spans they relate. For example, the occurrence of the discourse marker *But* at the beginning of a sentence most often signals either a mononuclear relation of ANTITHESIS or CONCESSION between a satellite, a textual span that precedes the occurrence of *But*, and a nucleus, a textual span that starts with *But*; or a multinuclear relation of CONTRAST between two nuclei: a textual span that precedes the occurrence of *But* and a textual span that starts with *But*. An exclusive disjunction is again an adequate way to formalize this hypothesis. For example, the exclusive disjunction $\mathit{rhet\_rel}(\textsc{antithesis}, 1, 2) \oplus \mathit{rhet\_rel}(\textsc{concession}, 1, 2) \oplus \mathit{rhet\_rel}(\textsc{contrast}, 1, 2)$ expresses the best hypothesis that one can make on the basis of the occurrence of the marker *But* in text (5). As mentioned already, in Section 2.4 it will become apparent why such exclusively disjunctive hypotheses are sufficient for deriving the rhetorical structure of texts.

(5) [Bill had no parents.^{1}] [*But* he had seven brothers and sisters.^{1}]

2.2.3 Discussion. The more complex the texts one is trying to analyze and the more ambiguous the connectives a text employs, the more likely the rhetorical relations that hold between elementary units and spans cannot be hypothesized precisely by automatic means. Most often, connectives, tense, pronoun uses, etc. only suggest that some rhetorical relations hold between some textual units; rarely can hypotheses be made with 100% confidence.

When a computer program processes free texts and comes across a connective such as *But*, for example, unless it carries out a complete semantic analysis and understands the intentions of the writer, it won't be able to determine unambiguously what relation to use; and it won't be able to determine what units or spans are involved in the relation. What is certain, though, is that *But*, the hypothesis trigger in this example, can signal *at most one* such relation—in my empirical work (see Section 3), I have never come across a case in which a simple connective signaled more than one rhetorical relation. In general then, if *But* occurs in unit i of a text, we know that it can signal a rhetorical relation that holds between one unit in the interval $[i-k, i-1]$ and one unit in the interval $[i, i+k]$, where k is a sufficiently large constant; or a relation between two spans $[i-k_1, i-1]$ and $[i, i+k_2]$. Figure 2 provides a graphical representation of

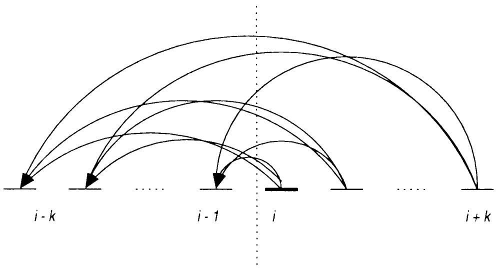

Figure 2
A graphical representation of the disjunctive hypothesis that is triggered by the occurrence of the marker *But* at the beginning of unit *i* of a text.

the simple rhetorical relations that can be hypothesized on the basis of the connective *But* in unit *i*.

In this paper, I will focus on dealing only with this sort of exclusively disjunctive hypotheses, i.e., hypotheses whose disjuncts subsume text spans that overlap. For example, in Figure 2, all disjuncts span over the segment $[i-1, i]$. From a linguistic perspective, only such hypotheses make sense. Although one can hypothesize on the basis of the occurrence of a discourse marker in unit *i* that a rhetorical relation R holds either between units $i-2$ and $i-1$ or between units i and $i+1$, for example, such a hypothesis will be ill-formed. In the discourse analyses I have carried out so far, I have never come across an example that would require one to deal with exclusively disjunctive hypotheses different from that shown graphically in Figure 2.

2.3 Determining Rhetorical Relations Using Cohesion

2.3.1 Pro Arguments. Youmans (1991), Hoey (1991), Morris and Hirst (1991), Salton et al. (1995), Salton and Allan (1995), and Hearst (1997) have shown that word co-occurrences and more sophisticated forms of lexical cohesion can be used to determine segments of topical and thematic continuity. And Morris and Hirst (1991) have also shown that there is a correlation between cohesion-defined textual segments and hierarchical, intentionally defined segments (Grosz and Sidner 1986). For example, if the first three paragraphs of a text talk about the moon and the subsequent two paragraphs talk about the Earth, it is possible that the rhetorical structure of the text is characterized by two spans that subsume these two sets of paragraphs and that a rhetorical relation of JOINT or LIST holds between the two spans. Also, studies by Harabagiu, Moldovan, and Maiorano (Harabagiu and Maiorano 1996; Harabagiu and Moldovan 1999) show that cohesion can be used to determine rhetorical relations that hold between smaller discourse constituents as well. For example, if one sentence talks about vegetables and another sentence talks about carrots and beets, it is possible that a rhetorical relation of ELABORATION holds between the two sentences because carrots and beets are kinds of vegetables.

2.3.2 Difficulties. For the purpose of this paper, I use a very coarse model of the relation between cohesion and rhetorical relations. More specifically, I assume that a mononuclear rhetorical relation of ELABORATION or BACKGROUND holds between two textual segments that talk about the same thing, i.e., they share some words, and that a multinuclear relation of JOINT holds between two segments that talk about different things. This assumption is consistent with the approaches discussed in Section 2.3.1, but does not follow from them. Section 5 empirically evaluates the impact that this assumption has on the problem of rhetorical structure derivation.

2.4 Determining Rhetorical Structure Using a Well-Constrained Mathematical Model

In my previous work (Marcu 1996, 1997b, 2000) I have formalized the constraints specific to valid rhetorical structures in the language of first-order logic. The axiomatization of valid rhetorical structures that I use throughout this paper relies on the following features and constraints.

- A valid rhetorical structure is a binary tree whose leaves denote elementary textual units.
- Rhetorical relations hold between textual units and spans of various sizes. These relations are paratactic or hypotactic. Paratactic relations are those that hold between units (spans) of equal importance. Hypotactic relations are those that hold between a unit (span) that is essential for the writer's purpose, i.e., a nucleus, and a unit (span) that increases the understanding of the nucleus but is not essential for the writer's purpose, i.e., a satellite.
- Each node of a rhetorical structure tree has associated a **status** (NUCLEUS or SATELLITE), a **type** (the rhetorical relation that holds between the text spans that the node spans over), and a set of **promotion units**. The set of promotion units of a textual span is determined recursively: it is given by the union of the promotion sets of the immediate subspans when the relation that holds between these subspans is paratactic, or by the promotion set of the nucleus subspan when the relation that holds between the immediate subspans is hypotactic. By convention, the type of a leaf is LEAF; and the promotion set of a leaf is a set that contains the leaf.
- The status and type associated with each node are unique. Hence, for example, a span cannot have both the status of NUCLEUS and the status of SATELLITE.
- The rhetorical relations of a valid rhetorical structure hold only between adjacent spans.
- There exists a span, which corresponds to the root node of the structure, that spans over the entire text.
- The status, type, and promotion set associated with each node reflect the compositionality criterion discussed in Section 2.2: if a rhetorical relation holds between two textual spans of the tree structure of a text, either that relation is extended or it can be explained in terms of a simple relation that holds between the promotion units of the constituent subspans.

Let us focus our attention again on text (4). We have seen that a computer program may be able to hypothesize the first exclusive disjunction in (6) using only knowledge about the discourse function of the connective *In contrast*. Similarly, a computer may be able to hypothesize that a rhetorical relation of ELABORATION holds between sentences 2 and 1 because both of them talk about John. A computer may also be able to hypothesize that a rhetorical relation of ELABORATION holds between sentence 4,

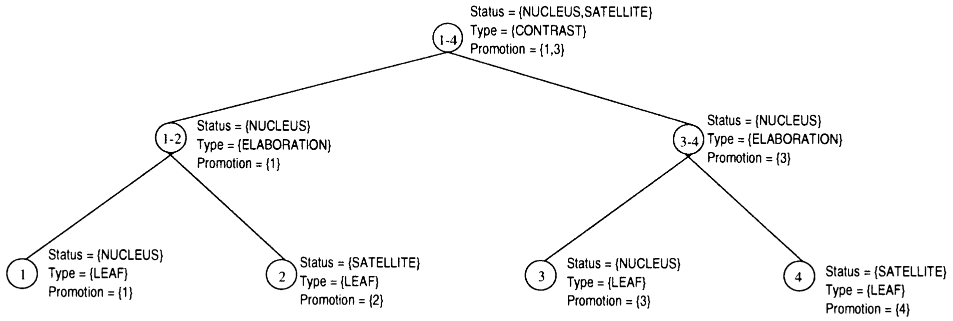

Figure 3
A valid rhetorical structure representation of text (4), which makes explicit the status, type, and promotion units that characterize each node.

which starts with the marker *Especially*, and a sentence that precedes it.

$$(6) \qquad RR = \begin{cases} \mathit{rhet\_rel}(\textsc{contrast}, 1, 3) \oplus \mathit{rhet\_rel}(\textsc{contrast}, 1, 4) \oplus \\ \quad \mathit{rhet\_rel}(\textsc{contrast}, 2, 3) \oplus \mathit{rhet\_rel}(\textsc{contrast}, 2, 4) \\ \mathit{rhet\_rel}(\textsc{elaboration}, 2, 1) \\ \mathit{rhet\_rel}(\textsc{elaboration}, 4, 1) \oplus \mathit{rhet\_rel}(\textsc{elaboration}, 4, 2) \oplus \\ \quad \mathit{rhet\_rel}(\textsc{elaboration}, 4, 3) \end{cases}$$

When these hypotheses are evaluated against the constraints of valid rhetorical structure trees, they yield only one valid rhetorical structure representation, which is shown in Figure 3. This representation makes explicit the status, type, and promotion set of each of the nodes in the tree. Note, for example, that the CONTRAST relation that holds between spans [1,2] and [3,4] is explained/determined by the simple rhetorical relation $\mathit{rhet\_rel}(\textsc{contrast}, 1, 3)$, which is one of the exclusive disjuncts shown in (6); hence, the rhetorical structure in Figure 3 is consistent with the compositionality criterion. Note also that the hypothesis $\mathit{rhet\_rel}(\textsc{elaboration}, 4, 2)$, for example, cannot be used instead of the CONTRAST relation to link spans [1,2] and [3,4], because the relation $\mathit{rhet\_rel}(\textsc{elaboration}, 4, 3)$ was used to link units 3 and 4 and because relations $\mathit{rhet\_rel}(\textsc{elaboration}, 4, 2)$ and $\mathit{rhet\_rel}(\textsc{elaboration}, 4, 3)$ are exclusively disjunctive.

In fact, even though one could have hypothesized a different relation R to hold, say, between the satellite 4 and the nucleus 2, such a hypothesis would not yield other valid trees because such trees would violate the compositionality criterion for two reasons:

- Relation R cannot be used to link spans [1,2] and [3,4], for example, because units 2 and 4 are not in the promotion sets of spans [1,2] and [3,4], respectively.
- There is no combination of rhetorical relations that would promote units 2 and 4 as salient in spans [1,2] and [3,4], respectively.

Hence, although we were not able to hypothesize precisely the spans and units between which the CONTRAST relation signaled by *In contrast* and the ELABORATION relation signaled by *Especially* hold, we were able to derive only one valid structure because the mathematical model that underlies our approach is well-constrained.

2.5 Discussion

Throughout Section 2, I have argued that connectives, cohesion, shallow processing, and a well-constrained model of discourse can be used to automatically derive the rhetorical structure of free, unrestricted texts. In order to substantiate this claim, I need to solve two problems:

1. First, I need to show how starting from free, unrestricted text, connectives and cohesion can be used to automatically determine the elementary units of text and hypothesize simple, extended, and exclusively disjunctive rhetorical relations that hold between these units and spans of units. I refer to this problem as **the problem of rhetorical grounding**.
2. Second, I need to show how starting from a sequence of textual units $U = 1, 2, \ldots, n$ and a set RR of simple, extended, and exclusively disjunctive rhetorical relations that hold among these units and among contiguous textual spans that are defined over U, the valid rhetorical structures of U can be determined, i.e., the rhetorical structures that are consistent with the constraints given in Section 2.4. I refer to this as **the problem of rhetorical structure derivation**.

The keen reader may have noted that in this formulation, the problem of determining the rhetorical structure of text is not modeled as an incremental process in which elementary units are determined and attached to an increasingly complex RS-tree. Rather, it is assumed that all elementary units of a text are determined first; that knowledge of connectives and cohesion is then used to (over-)hypothesize simple, extended, and exclusively disjunctive rhetorical relations that hold between units and spans of units; and that these hypotheses and the well-constrained model of valid RS-trees are used to determine the set of valid rhetorical interpretations that are consistent with both the mathematical model and the hypotheses.

In the rest of the paper, I provide solutions to the rhetorical grounding (Sections 3, 4.2, 4.3, and 4.4) and rhetorical structure derivation problems (Section 4.5). The problems are solved in the context of presenting a rhetorical parsing algorithm (see Figure 5), an algorithm that takes as input free text and determines the RS-tree of that text.

3. A Corpus Analysis of Cue Phrases

When I began this research, no empirical data existed which could answer the question of the extent to which connectives could be used to identify elementary units and hypothesize rhetorical relations. To better understand this problem, I carried out a corpus study. The corpus study was designed to investigate how cue phrases can be used to identify the elementary units of texts, as well as to determine what rhetorical relations hold between units and spans of text, the nuclearity of the units, and the sizes of the related spans. In this section, I describe the annotation schema that I used in the study. In Section 4, I explain how the annotated data was used to derive algorithms that identify connective occurrences (Section 4.2), determine elementary units of discourse and determine which connectives have a discourse function (Section 4.3), and hypothesize rhetorical relations that hold between elementary units and spans of texts (Section 4.4).

3.1 Materials

Many researchers have published lists of potential discourse markers and cue phrases (Halliday and Hasan 1976; Grosz and Sidner 1986; Martin 1992; Hirschberg and Litman 1993; Knott 1995; Fraser 1996). I took the union of their lists and created an initial set of more than 450 potential discourse markers. For each potential discourse marker, I then used an automatic procedure that extracted from the Brown corpus a set of text fragments. Each text fragment contained a "window" of approximately 300 words and an emphasized occurrence of a cue phrase. My initial goal was to select for each cue phrase 10 texts in which the phrase was used at the beginning of a sentence and 20 texts in which the phrase was used in the middle of a sentence. (In a prestudy, I had noticed that phrases occurring in the middle of sentences were more ambiguous and difficult to handle than those at the beginning of sentences.) However, since some of the phrases occurred in the corpus very seldom, I ended up with an average of 17 text fragments per cue phrase. Overall, I randomly selected more than 7,600 texts.

All the text fragments associated with a cue phrase were paired with a set of fields/slots in which I described two types of information.

Discourse-related Information. This information concerned the cue phrase under scrutiny and was described in the following fields:

Marker The field Marker encodes the orthographic environment that characterizes the use of the cue phrase. This included occurrences of periods, commas, colons, semicolons, etc. For example, when the cue phrase *besides* occurred within a sentence and was preceded by a comma, the Marker field was set to ", besides". When it occurred at the beginning of a paragraph and was immediately followed by a comma, the Marker field was set to "# Besides, ", where # denotes a paragraph break.

Usage The field Usage encodes the functional role of the cue phrase. The role can be one or more of the following: SENTENTIAL, DISCOURSE, and PRAGMATIC. A cue phrase has a sentential role if it makes a semantic contribution to the interpretation of text (Hirschberg and Litman 1993). A cue phrase has a discourse role if it signals a rhetorical relation that holds between two text spans. A cue phrase has a pragmatic role if it signals a relation between the unit to which the cue phrase belongs and the beliefs, plans, intentions, and/or communicative goals of the speaker/hearer (Fraser 1996).

Position The field Position specifies the position of the marker under scrutiny in the textual unit to which it belongs. The possible values for this field are: BEGINNING, MEDIAL, and END.

Right boundary The Right boundary of the textual unit associated with the marker under scrutiny contains the last cue phrase, orthographic marker, or word of that textual unit.

Where to link The field Where to link describes whether the textual unit that contains the discourse marker under scrutiny is related to a textual unit found BEFORE or AFTER it.

Rhetorical relation The field Rhetorical relation specifies one or more names of rhetorical relations that are signaled by the cue phrase under scrutiny. To encode the information specific to this field, I used a set of 54 rhetorical relations (see Section 3.2).

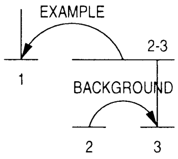

Figure 4
The discourse tree of text (7).

Types of textual units The field Types of textual units describes the types of textual units connected through a rhetorical relation that was signaled by the cue phrase under scrutiny. It takes values from CLAUSE to MULTIPLE_PARAGRAPH. I distinguished between these types of spans because I intended to use the corpus study to implement a rhetorical parser that hypothesizes both simple and extended rhetorical relations.

Statuses The field Statuses specifies the rhetorical statuses (separated by a semicolon) of the two textual units involved in the relation. The status of a textual unit can be NUCLEUS or SATELLITE.

Clause distance The field Clause distance contains a count of the clause-like units that separate the units related by the marker. The count is 0 when the related units are adjacent.

Sentence distance The field Sentence distance contains a count of the sentences that are found between the units that are related by the marker. The count is −1 when the related units belong to the same sentence.

Distance to salient unit The field Distance to salient unit contains a count of the clause-like units that separate the textual unit that contains the marker under scrutiny and the textual unit that is the most salient unit of the span that is rhetorically related to a unit that is before or after that under scrutiny. In most cases, this distance is −1, i.e., the unit that contains a marker is directly related to a unit that went before or to a unit that comes after. However, in some cases, this is not so. Consider, for example, the text given in (7) below, with respect to the cue phrase *for example*.

(7) [There are many things I do not like about fast food.^{1}] [Let's assume, *for example,* that you want to go out with someone^2.] [There is no way you can take them to a fast food restaurant!^{3}]

A rhetorical analysis of text (7) is shown in Figure 4. It is easy to see that although *for example* signals a rhetorical relation of EXAMPLE, the relation does not hold between units 2 and 1, but rather, between span 2–3 and unit 1. More precisely, the relation holds between unit 3, which is the most salient unit of span 2–3, and unit 1. The field Distance to salient unit reflects this state of affairs. For text (7) and marker *for example*, its value is 0.

When a discourse marker had more than one function or signaled more than one discourse relation, I enumerated all functions and relations.

Algorithmic Information. In contrast to the discourse-related information, which has a general linguistic interpretation, the algorithmic information was specifically tailored

Table 1
A corpus analysis of the cue phrase *Although* from text (8).

Field	Content
Marker	#⊔Although⊔
Usage	DISCOURSE
Right boundary	,
Where to link$_1$	AFTER
Types of textual units$_1$	CLAUSE;CLAUSE
Clause distance$_1$	0
Sentence distance$_1$	-1
Distance to salient unit$_1$	-1
Position$_1$	BEGINNING
Statuses$_1$	SATELLITE;NUCLEUS
Rhetorical relation$_1$	CONCESSION
Where to link$_2$	BEFORE
Types of textual units$_2$	SENTENCE;SENTENCE
Clause distance$_2$	6
Sentence distance$_2$	4
Distance to salient unit$_2$	-1
Position$_2$	BEGINNING
Statuses$_2$	NUCLEUS;SATELLITE
Rhetorical relation$_2$	ELABORATION
Break action	COMMA

to the surface analysis aimed at determining the elementary textual units of a text. It concerned only one field, Break action, which specified the action that a left-to-right surface-based elementary unit identifier will need to take to determine the boundaries of elementary textual units found in the vicinity of the cue phrase. For example, an action of type NORMAL associated with the occurrence of the connective *but* encoded the fact that an elementary unit boundary had to be inserted immediately before the connective. Since a discussion of the actions and their semantics is meaningless in isolation, I will provide it below in Section 4.3.3, in conjunction with the clause-like unit boundary and discourse marker identification algorithm.

One can argue that encoding algorithmic information in a corpus study is not necessary. After all, one can use the annotated data to derive such information automatically. However, during my prestudy of cue phrases, I noticed that there is a finite number of ways in which cue phrases can be used to identify the elementary units of text. By encoding algorithmic specific information in the corpus, I only bootstrap the step that can take one from annotated data to algorithmic information. This encoding does not preclude the employment of more sophisticated methods that derive algorithmic information automatically.

3.2 Methods and Results

Once the database had been created, I analyzed its records and updated the fields according to the requirements described above. For example, Table 1 shows the information that I associated with the fields when I analyzed the text fragment shown in (8), with respect to the cue phrase *Although*. The square brackets in (8) enclose the elementary units of interest.

(8) [How well do faculty members govern themselves?] [There is little evidence that they are giving any systematic thought to a general theory

> of the optimum scope and nature of their part in government.] [They sometimes pay more attention to their rights] [than to their own internal problems of government.] [They, too, need to learn to delegate.] [Letting the administration take details off their hands would give them more time to inform themselves about education as a whole,] [an area that would benefit by more faculty attention.]
>
> [*Although* faculties insist on governing themselves,] [they grant little prestige to a member who actively participates in college or university government.]

The information encoded in Table 1 specifies that the marker *Although*, which occurs in the beginning of a paragraph (#⊔Although⊔), has a DISCOURSE role and that the right boundary of the elementary unit to which it belongs is a comma. *Although* signals a relation of CONCESSION between the clause to which it belongs, which has a rhetorical status of SATELLITE, and the clause that comes immediately AFTER it, which has a rhetorical status of NUCLEUS. In addition to the discourse relation signaled by a marker such as *Although*, which introduces expectations (Cristea and Webber 1997), I also found it useful to annotate the rhetorical relation that held between the sentence to which an expectation-based marker belonged and the text span that went before. For example, with respect to the connective *Although* in text (8), I also represented explicitly in the corpus the fact that an ELABORATION relation holds between the sentence that contains the connective, which has the status of SATELLITE, and a sentence found six clauses (four sentences) BEFORE it, which has the status of NUCLEUS. It turned out that in most of the cases in which a phrase such as *Although* was used at the beginning of a sentence/paragraph, it not only signaled a CONCESSION relation between two clauses, but its use also correlated with an ELABORATION relation that held between two sentences or paragraphs.

Overall, I have manually analyzed 2,100 of the text fragments in the corpus. I annotated only 2,100 fragments because the task was too time-consuming to complete. Of the 2,100 instances of cue phrases that I considered, 1,197 had a discourse function, 773 were sentential, and 244 were pragmatic.[4]

The taxonomy of relations that I used to label the 1,197 discourse uses in the corpus contained 54 relations. Marcu (1997b) lists their names and the number of instances in which each rhetorical relation was used. The number of relations is much larger than 24, which is the size of the taxonomy proposed initially by Mann and Thompson (1988), because during the corpus analysis, it often happened that the relations proposed by Mann and Thompson seemed inadequate to capture the semantics of the relationship between the units under consideration. Because the study described here was exploratory, I considered it appropriate to introduce relations that would better capture the meaning of these relationships. The rhetorical relation names were chosen so as to reflect the intended semantics of the relations. To manage the new relations, I did not provide for them definitions similar to those proposed by Mann and Thompson (1988); instead, I kept a list of text examples that I considered to reflect the meaning of each new rhetorical relation that I introduced.

In Section 4, I will explain how the annotated data was used in order to implement algorithms that solve the problem of rhetorical grounding defined in Section 2.5.

[4] The three numbers add up to more than 2,100 because some cue phrases had multiple roles in some text fragments.

3.3 Discussion

The elementary textual units that I considered, such as those enclosed within square brackets in examples (4) and (8) were not necessarily clauses in the traditional, grammatical sense. Rather, they were contiguous spans of text that could be smaller than a clause and that could provide grounds for deriving rhetorical inferences. For example, although the text in italics in the sentence "Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly *because of the low atmospheric pressure.*" does not represent a full-fledged clause, I decided to label it as an elementary unit because it provides the grounds for inferring a causal relation.

Hence, in the texts that I analyzed, I did not use an objective definition of elementary unit. Rather, I relied on a more intuitive one: whenever I found that a rhetorical relation held between two spans of text of significant sizes (the relation could be signaled or not by a cue phrase, or not), I assigned those spans an elementary unit status, although in some cases they were not full-fledged clauses. In the rest of the paper I refer to such elementary units with the term **clause-like unit**.

The main advantage of the empirical work described here is the empirical grounding that it provides for a set of algorithms that derive the rhetorical structures of unrestricted texts. These algorithms are grounded partly in the empirical data derived from the corpus and partly in the intuitions that I developed during the discourse analysis of the 2,100 fragments of text.

Since I was the only analyst of 2,100 of the 7,600 of the text fragments in the corpus and since I wanted to avoid evaluating the algorithms that I developed against my own subjective standard, I used the corpus analysis only for algorithm development. The testing of the algorithms was done against data that did not occur in the corpus and that was analyzed independently by other judges.

4. The Rhetorical Parsing Algorithm

The rhetorical parsing algorithm takes as input a free, unrestricted text and determines its rhetorical structure. The algorithm presented in this paper assumes that the rhetorical structure of a text correlates with the orthographic layout of that text. That is, it assumes that sentences, paragraphs, and sections correspond to hierarchical spans in the rhetorical representation of the text that they subsume.

Obviously, this assumption is controversial because there is no clear-cut evidence that the rhetorical structure of a text correlates with its paragraph structure, for example. In fact, some psycholinguistic and empirical research of Heurley (1997) and Hearst (1997) indicates that paragraph breaks do not always occur at the same locations as the thematic boundaries. In contrast, experiments of Bruder and Wiebe (1990) and Wiebe (1994) show that paragraph breaks help readers to interpret private-state sentences in narratives, i.e., sentences about psychological states such as wanting and perceptual states such as seeing. Hence, paragraph breaks play an important role in story comprehension. In my own experiments (see Section 5), I observed that, in nine out of ten cases, human judges manually built rhetorical structures that correlated with the underlying paragraph boundaries.

The main reason for assuming that the orthographic layout of text correlates with its rhetorical structure is primarily one of efficiency. In the same way sentences are ambiguous and syntactic parsers can derive thousands of syntactic trees, so texts are ambiguous and rhetorical parsers can derive thousands of rhetorical trees. Assuming that the rhetorical structure of a text correlates with sentence, paragraph, and section

Input: A text *T*.
Output: The valid rhetorical structures of *T*.

```
1.  I. Determine the set $D$ of all cue phrase (potential discourse marker) instances in $T$.
2.  II. Use information derived from the corpus analysis in order to determine
3.      recursively all the sections, paragraphs, sentences, and clause-like units of the
4.      text and the set $D_d \in D$ of cue phrases that have a discourse function.
5.  III. For each of the three highest levels of granularity (sentences, paragraphs,
6.       and sections)
7.          III.1 Use information derived from the corpus analysis about the
8.                discourse markers $D_d$ in order to hypothesize rhetorical relations
9.                among the elementary units that correspond to that level.
10.         III.2 Use cohesion in order to hypothesize rhetorical relations among
11.               the units for which no hypotheses were made in step III.1.
12.         III.3 Apply the proof theory discussed in Section 4.5 in order to
13.               determine all the valid text trees that correspond to that level.
14.         III.4 Assign a weight to each of the text trees and determine the tree
15.               with maximal weight.
16. IV. Merge the best trees that correspond to each level into a discourse tree that
17.     spans the whole text and that has clause-like units as its elementary units.
```

Figure 5
Outline of the rhetorical parsing algorithm.

boundaries significantly reduces the search space of possible rhetorical interpretations and increases the speed of a rhetorical parser.

4.1 A Bird's-Eye View

The rhetorical parsing algorithm, which was implemented C++, is outlined in Figure 5. The rhetorical parser first determines the set of all instances of cue phrases that occur in the text; this set includes punctuation marks such as commas, periods, and semicolons. In the second step (lines 2–4 in Figure 5), the rhetorical parser retraverses the input and by using information derived from the corpus study discussed in Section 3, it determines the elementary units and the cue phrases that have a discourse function in structuring the text. In the third step, the rhetorical parser builds the valid text structures for each of the three highest levels of granularity, which are the sentence, paragraph, and section levels (see lines 5–15 in Figure 5). Tree construction is carried out in four substeps.

III.1 First, the rhetorical parser uses the cue phrases that were assigned a discourse function in step II to hypothesize rhetorical relations between clause-like units, sentences, and paragraphs (see lines 7–9). Most of the discourse markers yield exclusively disjunctive hypotheses.

III.2 When the textual units under consideration are characterized by no discourse markers, rhetorical relations are hypothesized on the basis of a simple cohesive device, which is similar to that used by Hearst (1997) (see lines 10–11).

III.3 Once the set of textual units and the set of rhetorical relations that hold among the units have been determined, the algorithm derives discourse trees at each of the three levels that are assumed to be in correlation with the discourse structure: sentence, paragraph, and section levels (see lines 12–13). The derivation is accomplished by a chart-based implementation of a proof theory that solves the rhetorical structure derivation problem (see Section 4.5).

III.4 Since the rhetorical parsing process is ambiguous, more than one discourse tree is usually obtained at each of these levels. To deal with this ambiguity, a "best" tree is selected according to a metric to be discussed in Section 4.6 (see lines 14–15).

In the final step, the algorithm assembles the trees built at each level of granularity, thus obtaining a discourse tree that spans over the whole text (lines 16–17 in Figure 5).

In the rest of the paper, I discuss in detail the steps that the rhetorical parser follows when it derives the valid structures of a text and the algorithms that implement them. In the cases in which the algorithms rely on data derived from the corpus study in Section 3, I also discuss the relationship between the predominantly linguistic information that characterizes the corpus and the procedural information that can be exploited at the algorithmic level. Throughout the discussion, I will use the text in (1) as an example.

4.2 Determining the Potential Discourse Markers of a Text

4.2.1 From the Corpus Analysis to the Potential Discourse Markers of a Text. The corpus analysis discussed in Section 3 provides information about the orthographic environment of cue phrases and the function they have in the text (sentential, discourse, or pragmatic). Different orthographic environments often correlate with different discourse functions and different ways of breaking the surrounding text into elementary units. For example, if the cue phrase *Besides* occurs at the beginning of a sentence and is not followed by a comma, as in text (9), it usually signals a rhetorical relation that holds between the clause-like unit that contains it and the following clause(s). However, if the same cue phrase occurs at the beginning of a sentence and is immediately followed by a comma, as in text (10), it usually signals a rhetorical relation that holds between the sentence to which *Besides* belongs and a textual unit that precedes it.

(9) [*Besides* the lack of an adequate ethical dimension to the Governor's case,] [one can ask seriously whether our lead over the Russians in quality and quantity of nuclear weapons is so slight as to make the tests absolutely necessary.]

(10) [For pride's sake, I will not say that the coy and leering vade mecum of those verses insinuated itself into my soul.] [*Besides,* that particular message does no more than weakly echo the roar in all fresh blood.]

I have taken each cue phrase in the corpus and evaluated its potential contribution in determining the elementary textual units and in hypothesizing the rhetorical relations that hold among the units for each orthographic environment that characterized its usage. I used the cue phrases that had a discourse role in most of the text fragments and the orthographic environments that characterized them to manually develop a set of regular expressions that can be used to recognize potential discourse markers in naturally occurring texts. If a cue phrase had different discourse functions in different orthographic environments and could be used in different ways in identifying the elementary units of the surrounding text, as was the case with *Besides*, I created one regular expression for each function. I ignored both cue phrases that had a sentential role in a majority of the instances in the corpus and those that were too ambiguous to be exploited in the context of a surface-based approach. In general, I preferred to be

Table 2
A list of regular expressions that correspond to occurrences of some of the potential discourse markers and punctuation marks.

Marker	Regular Expression
Although	[⊔\t\n]Although(⊔ | \t | \n)
because	[,][⊔\t\n]+because(⊔ | \t | \n)
but	[⊔\t\n]+but(⊔ | \t | \n)
for example	[,][⊔\t\n]+for[⊔\t \n]+example(⊔ |, | \t | \n)
where	,[⊔\t\n]+where(⊔ | \t | \n)
With	[⊔\t\n]With(⊔ | \t | \n)
Yet	[⊔\t\n]Yet(⊔ | \t | \n)
COMMA	,(⊔ | \t | \n)
OPEN_PAREN	[,][⊔\t\n]+(
CLOSE_PAREN	)(⊔ | \t | \n)
DASH	[,][⊔\t\n]+—(⊔ | \t | \n)
END_SENTENCE	(".")|("?")|("!")|(".″")|("?″")|("!″"))
BEGIN_PARAGRAPH	⊔⋆((\n\t[⊔\t]⋆)|(\n[⊔\t\n]{2,}))

conservative and to consider only potential cue phrases whose discourse role could be determined with a relatively high level of confidence. Table 2 shows a set of regular expressions that correspond to some of the cue phrases in the corpus. Because orthographic markers, such as commas, periods, dashes, paragraph breaks, etc., play an important role in our surface-based approach to discourse processing, I included them in the list of potential discourse markers as well.

By considering only cue phrases having a discourse function in most of the cases, I deliberately chose to focus more on precision than on recall with respect to the task of identifying the elementary units of text. That is, I chose to determine fewer units than humans do, hoping that, in this way, most of the identified units would be correct.

4.2.2 An Algorithm for Determining the Potential Discourse Markers of a Text. Once the regular expressions that match potential discourse markers were derived, it was trivial to implement the first step of the rhetorical parser (line 1 in Figure 5). A program that uses the Unix tool *lex* traverses the text given as input and determines the locations at which potential discourse markers occur. For example, when the regular expressions are matched against text (1), the algorithm recognizes all punctuation marks and the cue phrases shown in italics in text (11) below.

(11) *With* its distant orbit—50 percent farther from the sun than Earth—*and* slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -60 degrees Celsius (-76 degrees Fahrenheit) at the equator *and* can dip to -123 degrees C near the poles. Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, *but* any liquid water formed in this way would evaporate almost instantly *because* of the low atmospheric pressure.

Although the atmosphere holds a small amount of water, *and* water-ice clouds sometimes develop, most Martian weather involves blowing dust *or* carbon dioxide. Each winter, *for example*, a blizzard of frozen carbon dioxide rages over one pole, *and* a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap. *Yet* even on the summer pole, *where* the sun

> remains in the sky all day long, temperatures never warm enough to melt frozen water.

4.3 Determining the Elementary Units of a Text

4.3.1 From the Corpus Analysis to the Elementary Units of a Text. As I discussed in Section 3, the corpus study encoded not only linguistic information but also algorithmic information, in the field Break action. During the corpus analysis, I generated a set of 11 actions that constitutes the foundation of an algorithm to automatically determine the elementary units of a text. The algorithm processes each sentence in the text given as input in a left-to-right fashion and "executes" the actions that are associated with each potential discourse marker and each punctuation mark that occurs in that sentence. Because the algorithm does not use any traditional parsing and tagging techniques, I call it a **shallow analyzer**.

The names and the intended semantics of the actions used by the shallow analyzer are:

- Action NOTHING instructs the shallow analyzer to treat the cue phrase under consideration as a simple word. That is, no textual unit boundary is normally set when a cue phrase associated with such an action is processed. For example, the action associated with the cue phrase *accordingly* is NOTHING.
- Action NORMAL instructs the analyzer to insert a textual boundary immediately before the occurrence of the marker. Textual boundaries correspond to elementary unit breaks.
- Action COMMA instructs the analyzer to insert a textual boundary immediately after the occurrence of the first comma in the input stream. If the first comma is followed by an *and* or an *or*, the textual boundary is set after the occurrence of the next comma instead. If no comma is found before the end of the sentence, a textual boundary is created at the end of the sentence.
- Action NORMAL_THEN_COMMA instructs the analyzer to insert a textual boundary immediately before the occurrence of the marker and to insert another textual boundary immediately after the occurrence of the first comma in the input stream. As in the case of the action COMMA, if the first comma is followed by an *and* or an *or*, the textual boundary is set after the occurrence of the next comma. If no comma is found before the end of the sentence, a textual boundary is created at the end of the sentence.
- Action END instructs the analyzer to insert a textual boundary immediately after the cue phrase.
- Action MATCH_PAREN instructs the analyzer to insert textual boundaries both before the occurrence of the open parenthesis that is normally characterized by such an action, and after the closed parenthesis that follows it.
- Action COMMA_PAREN instructs the analyzer to insert textual boundaries both before the cue phrase and after the occurrence of the next comma in the input stream.

- Action MATCH_DASH instructs the analyzer to insert a textual boundary before the occurrence of the cue phrase. The cue phrase is usually a dash. The action also instructs the analyzer to insert a textual boundary after the next dash in the text. If such a dash does not exist, the textual boundary is inserted at the end of the sentence.

  The preceding three actions, MATCH_PAREN, COMMA_PAREN, and MATCH_DASH, are used for determining the boundaries of parenthetical units.
- Action SET_AND/SET_OR instructs the analyzer to store the information that the input stream contains the lexeme *and*/*or*.
- Action DUAL instructs the analyzer to insert a textual boundary immediately before the cue phrase under consideration if there is no other cue phrase that immediately precedes it. If there exists such a cue phrase, the analyzer will behave as in the case of the action COMMA. The action DUAL is usually associated with cue phrases that can introduce some expectations about the discourse (Cristea and Webber 1997). For example, the cue phrase *although* in text (12) signals a rhetorical relation of CONCESSION between the clause to which it belongs and the previous clause. However, in text (13), where *although* is preceded by an *and*, it signals a rhetorical relation of CONCESSION between the clause to which it belongs and the next clause in the text.

(12) [I went to the theatre] [*although* I had a terrible headache.]

(13) [The trip was fun,] [*and although* we were badly bitten by blackflies,] [I do not regret it.]

In addition to the algorithmic information that is explicitly encoded in the field Break action, the shallow analyzer uses information about the position of cue phrases in the elementary textual units to which they belong. The position information is extracted directly from the corpus, from the field Position. Hence, each regular expression that has a corresponding instantion in the texts in the corpus that could play a discourse function is assigned a structure with two features:

- the action that the shallow analyzer should perform in order to determine the boundaries of the textual units found in its vicinity;
- the relative position of the marker in the textual unit to which it belongs (beginning, middle, or end).

Table 3 lists the actions and the positions in the elementary units of the cue phrases and orthographic markers shown in Table 2.

4.3.2 The Section, Paragraph, and Sentence Identification Algorithm. As discussed in Section 4.1, the rhetorical parser assumes that sentences, paragraphs, and sections correspond to hierarchical spans in the rhetorical representation of the text that they subsume.

The algorithm that determines the section, paragraph, and sentence boundaries is a very simple one, which uses the set of regular expressions that are associated with the potential discourse markers END_SENTENCE and BEGIN_PARAGRPH found in Table 2 and a list of abbreviations, such as *Mr.*, *Mrs.*, and *Inc.*, that prevent the setting of

Table 3
The list of actions that correspond to the potential discourse markers and punctuation marks shown in Table 2; B = beginning, M = middle, and E = end.

Marker	Position	Action
Although	B	COMMA
because	B	DUAL
but	B	NORMAL
for example	M	NOTHING
where	B	COMMA_PAREN
With	B	COMMA
Yet	B	NOTHING
COMMA	E	NOTHING
OPEN_PAREN	B	MATCH_PAREN
CLOSE_PAREN	E	NOTHING
DASH	B	MATCH_DASH
END_SENTENCE	E	NOTHING
BEGIN_PARAGRAPH	B	NOTHING

sentence and paragraph boundaries at places that are inappropriate. This simple algorithm correctly located all of the paragraph boundaries and all but one of the sentence boundaries found in the texts that I used to evaluate the clause-like unit and discourse marker identification algorithm that I will present in Section 4.3.3. Other texts and semistructured HTML/SGML documents may need more sophisticated algorithms to solve this segmentation problem, such as those described by Palmer and Hearst (1997).

4.3.3 The Clause-Like Unit and Discourse Marker Identification Algorithm. On the basis of the information derived from the corpus, I have designed an algorithm that identifies elementary textual unit boundaries in sentences and cue phrases that have a discourse function. Figure 6 shows only its skeleton and focuses on the variables and steps that are used to determine the elementary units. The steps that assert the discourse function of a marker are not shown; however, these steps are mentioned in the discussion of the algorithm given below. Marcu (1997b) provides a full description of the algorithm.

The algorithm takes as input a sentence S and the array markers[n] of cue phrases (potential discourse markers) that occur in that sentence; the array is produced by a trivial algorithm that recognizes regular expressions (see Section 4.2.2). Each element in markers[n] is characterized by a feature structure with the following entries:

- the action associated with the cue phrase;
- the position in the elementary unit of the cue phrase;
- a flag *has_discourse_function* that is initially set to "no."

The clause-like unit and discourse marker identification algorithm traverses the array of cue phrases left-to-right (see the loop between lines 2 and 20) and identifies the elementary textual units in the sentence on the basis of the types of the markers that it processes. Crucial to the algorithm is the variable "status," which records the set of markers that have been processed earlier and that may still influence the identification of clause and parenthetical unit boundaries.

```
Input:    A sentence S.
          The array of $n$ potential discourse markers markers[$n$] that occur in S.
Output:   The clause-like units, parenthetical units, and discourse markers of S.

1.  status := NIL; . . .;
2.  for $i$ from 1 to $n$
3.      if MATCH_PAREN $\in$ status $\vee$ MATCH_DASH $\in$ status $\vee$ COMMA_PAREN $\in$ status
4.        ⟨deal with parenthetical information⟩
5.      if COMMA $\in$ status $\wedge$ markerTextEqual($i$,“,”) $\wedge$
6.            NextAdjacentMarkerIsNotAnd() $\wedge$ NextAdjacentMarkerIsNotOr()
7.        ⟨insert textual boundary after comma⟩
8.      if (SET_AND $\in$ status $\vee$ SET_OR $\in$ status) $\wedge$ markerAdjacent($i - 1, i$)
9.        ⟨deal with adjacent markers⟩
10.     switch(getActionType($i$)){
11.         case DUAL: ⟨deal with DUAL markers⟩
12.         case NORMAL: ⟨insert textual boundary before marker⟩
13.         case COMMA: status := status $\cup$ {COMMA};
14.         case NORMAL_THEN_COMMA: ⟨insert textual boundary before marker⟩
15                                  status := status $\cup$ {COMMA};
16.         case NOTHING: ⟨assign discourse usage⟩$\star$
17.         case MATCH_PAREN, COMMA_PAREN, MATCH_DASH: status := status $\cup$
                  {getActionType($i$)};
18.         case SET_AND, SET_OR: status := status $\cup$ {getActionType($i$)};
19.     }
20. end for
21. finishUpParentheticalsAndClauses();
```

Figure 6
The skeleton of the clause-like unit and discourse marker identification algorithm.

The clause-like unit identification algorithm has two main parts: lines 10–20 concern actions that are executed when the status variable is NIL. These actions can insert textual unit boundaries or modify the value of the status variable, thus influencing the processing of further markers. Lines 3–9 concern actions that are executed when the status variable is not NIL. We discuss each of these actions in turn.

Lines 3–4 of the algorithm treat parenthetical information. Once an open parenthesis, a dash, or a discourse marker whose associated action is COMMA_PAREN has been identified, the algorithm ignores all other potential discourse markers until the element that closes the parenthetical unit is processed. Hence, the algorithm searches for the first closed parenthesis, dash, or comma, ignoring all other markers on the way. Obviously, this implementation does not assign a discourse usage to discourse markers that are used *within* a span that is parenthetic. However, this choice is consistent with the decision, discussed in Section 4.3.1, to assign parenthetical information no elementary textual unit status. Because of this, the text shown in italics in text (14), for example, is treated as a single parenthetical unit, which is subordinated to “Yet, even on the summer pole, temperatures never warm enough to melt frozen water.” In dealing with parenthetical units, the algorithm avoids setting boundaries in cases in which the first comma that comes after a COMMA_PAREN marker is immediately followed by an *or* or an *and*. As example (14) shows, taking the first comma as the boundary of the parenthetical unit would be inappropriate.

(14) [Yet, even on the summer pole, {*where the sun remains in the sky all day long, and where winds are not as strong as at the Equator,*} temperatures never warm enough to melt frozen water.]

Obviously, one can easily find counterexamples to this rule (and to other rules that are employed by the algorithm). For example, the clause-like unit and discourse marker identification algorithm will produce erroneous results when it processes the sentence shown in (15) below.

(15) [I gave John a boat,] [which he liked, and a duck,] [which he didn't.]

Nevertheless, the evaluation results discussed in Section 4.3.4 show that the algorithm produces correct results in the majority of the cases.

If the status variable contains the action COMMA, the occurrence of the first comma that is not adjacent to an *and* or an *or* marker determines the identification of a new elementary unit (see lines 5–7 in Figure 6).

Usually, the discourse role of the cue phrases *and* and *or* is ignored because the surface-form algorithm that we propose is unable to distinguish accurately enough between their discourse and sentential usages. However, lines 8–9 of the algorithm concern cases in which their discourse function can be unambiguously determined. For example, in our corpus, whenever *and* and *or* immediately preceded the occurrence of other discourse markers (function markerAdjacent($i-1, i$) returns "true"), they had a discourse function. For example, in sentence (16), *and* acts as an indicator of a JOINT relation between the first two clauses of the text.

(16) [Although the weather on Mars is cold] [*and although* it is very unlikely that water exists,] [scientists have not dismissed yet the possibility of life on the Red Planet.]

If a discourse marker is found that immediately follows the occurrence of an *and* (or an *or*) and if the left boundary of the elementary unit under consideration is found to the left of the *and* (or the *or*), a new elementary unit is identified whose right boundary is just before the *and* (or the *or*). In such a case, the *and* (or the *or*) is considered to have a discourse function as well, so the flag *has_discourse_function* is set to "yes."

If any of the complex conditions in lines 3, 5, or 8 in Figure 6 is satisfied, the algorithm not only inserts textual boundaries as discussed above, but also resets the status variable to NIL.

Lines 10–19 of the algorithm concern the cases in which the status variable is NIL. If the type of the marker is DUAL, the determination of the textual unit boundaries depends on the marker under scrutiny being adjacent to the marker that precedes it. If it is, the status variable is set such that the algorithm will act as in the case of a marker of type COMMA. If the marker under scrutiny is not adjacent to the marker that immediately preceded it, a textual unit boundary is identified. This implementation will modify, for example, the status variable to COMMA when processing the marker *although* in example (17), but only insert a textual unit boundary when processing the same marker in example (18). The final textual unit boundaries that are assigned by the algorithm are shown using square brackets.

(17) [John is a nice guy,] [*but although* his colleagues do not pick on him,] [they do not invite him to go camping with them.]

(18) [John is a nice guy,] [*although* he made a couple of nasty remarks last night.]

Line 12 of the algorithm concerns the most frequent marker type. The type NORMAL determines the identification of a new clause-like unit boundary just before the marker under scrutiny. Line 13 concerns the case in which the type of the marker is COMMA. If the marker under scrutiny is adjacent to the previous one, the previous marker is considered to have a discourse function as well. In either case, the status variable is updated such that a textual unit boundary will be identified at the first occurrence of a comma. When a marker of type NORMAL_THEN_COMMA is processed, the algorithm identifies a new clause-like unit as in the case of a marker of type NORMAL, and then updates the status variable such that a textual unit boundary will be identified at the first occurrence of a comma. In the case in which a marker of type NOTHING is processed, the only action that might be executed is that of assigning that marker a discourse usage.

Lines 17–18 of the algorithm concern the treatment of markers that introduce expectations with respect to the occurrence of parenthetical units: the effect of processing such markers is that of updating the status variable according to the type of the action associated with the marker under scrutiny. The same effect is observed in the cases in which the marker under scrutiny is an *and* or an *or*.

After processing all the markers, it is possible that some text will remain unaccounted for: this text usually occurs between the last marker and the end of the sentence. The procedure finishUpParentheticalsAndClauses() in line 21 of Figure 6 puts this text into the last clause-like unit that is under consideration.

The clause-like unit boundary and discourse marker identification algorithm has been implemented in C++. When it processes text (11), it determines that the text has 10 elementary units and that six cue phrases have a discourse function. Text (19) shows the elementary units within square brackets. The instances of parenthetical information are shown within curly brackets. The cue phrases that are assigned by the algorithm as having a discourse function are shown in italics.

(19) [*With* its distant orbit {— 50 percent farther from the sun than Earth —} and slim atmospheric blanket,^{1}] [Mars experiences frigid weather conditions.^{2}] [Surface temperatures typically average about −60 degrees Celsius {(−76 degrees Fahrenheit)} at the equator and can dip to −123 degrees C near the poles.^{3}] [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,^{4}] [*but* any liquid water formed in this way would evaporate almost instantly^{5}] [*because* of the low atmospheric pressure.^{6}]

[*Although* the atmosphere holds a small amount of water, and water-ice clouds sometimes develop,^{7}] [most Martian weather involves blowing dust or carbon dioxide.^{8}] [Each winter, *for example,* a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap.^{9}] [*Yet* even on the summer pole, {where the sun remains in the sky all day long,} temperatures never warm enough to melt frozen water.^{10}]

4.3.4 Evaluation of the Clause-Like Unit and Discourse Marker Identification Algorithm. The algorithm shown in Figure 6 determines clause-like unit boundaries and identifies discourse uses of cue phrases using methods based on surface form. The algorithm relies heavily on the corpus study discussed in Section 3.

The most important criterion for using a cue phrase in the clause-like unit and discourse marker identification algorithm is that the cue phrase (together with its orthographic neighborhood) functions as a discourse marker in the majority of the examples in the corpus. On the one hand, the enforcement of this criterion reduces the recall of the discourse markers that can be detected, but on the other hand, it significantly increases the precision. I chose to ignore the ambiguous markers deliberately because, during the corpus analysis, I noticed that many of the markers that connect large textual units *can* be identified by a shallow analyzer. In fact, the discourse marker responsible for most of the algorithm recall failures is *and*. Since a shallow analyzer cannot identify with sufficient precision whether an occurrence of *and* has a discourse or a sentential usage, most of its occurrences are therefore ignored. It is true that, in this way, the discourse structures that the rhetorical parser eventually builds lose some potentially finer granularity, but fortunately, from a rhetorical analysis perspective, the loss has insignificant global repercussions: the majority of the relations that the algorithm misses due to recall failures of *and* are JOINT and SEQUENCE relations that hold between adjacent clause-like units.

To evaluate the clause-like unit and discourse marker identification algorithm, I randomly selected three texts, each belonging to a different genre:

1. an expository text of 5,036 words from *Scientific American*;
2. a magazine article of 1,588 words from *Time*;
3. a narration of 583 words from the Brown corpus (segment P25:1250–1710).

No fragment of any of the three texts was used during the corpus analysis. Three independent judges, graduate students in computational linguistics, broke the texts into elementary units. The judges were given no detailed instructions about the criteria that they were to apply in determining the clause-like unit boundaries. Rather, they were supposed to rely on their intuition and preferred definition of clause and to insert a boundary between two clause-like units when they believed that a rhetorical relation held between those units. The locations in texts that were labeled as clause-like unit boundaries by at least two of the three judges were considered to be valid elementary unit boundaries.

I used the valid elementary unit boundaries assigned by judges as indicators of discourse usages of cue phrases and I manually determined the cue phrases that signaled a discourse relation. For example, if an *and* was used in a sentence and if the judges agreed that a textual unit boundary existed just before the *and*, I assigned that *and* a discourse use. Otherwise, I assigned it a sentential usage. I applied this procedure to instances of all 450 cue phrases in the corpus, not only to the subset of phrases that were used by the rhetorical parser. Hence, I manually determined all discourse usages of cue phrases and all discourse boundaries between elementary units.

I then applied the clause-like unit and discourse marker identification algorithm to the same texts. The algorithm found 80.8% of the discourse markers with a precision of 89.5% (see Table 4), a result that seems to outperform Hirschberg and Litman's (1993) algorithm.[5] The large difference in recall between the first and the third texts is due to the different text genres. In the third text, which is a narration, the discourse marker *and* occurs frequently. As discussed above, the clause-like unit and discourse marker identification algorithm correctly labels only a small percentage of these occurrences.

The algorithm correctly identified 81.3% of the clause-like unit boundaries, with a precision of 90.3% (see Table 5).

[5] Since the algorithm proposed here and Hirschberg and Litman's algorithm were evaluated on different corpora, it is impossible to carry out a fair comparison. Also, the discourse markers in my three texts were not identified using an independent definition, as Hirschberg and Litman were.

Table 4
Evaluation of the marker identification procedure.

Text	Number of Discourse Markers Identified Manually	Number of Discourse Markers Identified by the Algorithm	Number of Discourse Markers Identified Correctly by the Algorithm	Recall	Precision
1.	174	169	150	86.2%	88.8%
2.	63	55	49	77.8%	89.1%
3.	38	24	23	63.2%	95.6%
Total	275	248	222	80.8%	89.5%

Table 5
Evaluation of the clause-like unit boundary identification procedure.

Text	Number of Sentence Boundaries	Number of Clause-like Unit Boundaries Identified Manually	Number of Clause-like Unit Boundaries Identified by the Algorithm	Number of Clause-like Unit Boundaries Identified Correctly by the Algorithm	Recall	Precision
1.	242	428	416	371	86.7%	89.2%
2.	80	151	123	113	74.8%	91.8%
3.	19	61	37	36	59.0%	97.3%
Total	341	640	576	520	81.3%	90.3%

4.4 Hypothesizing Rhetorical Relations between Textual Units of Various Granularities

4.4.1 From Discourse Markers to Rhetorical Relations. To hypothesize rhetorical relations, I manually associated with each of the regular expressions that can be used to recognize potential discourse markers in naturally occurring texts (see Section 4.2.1) a set of features for each of the discourse roles that a discourse marker can play. Each set had six distinct features:

- The feature Statuses specifies the rhetorical status of the units that are linked by the discourse marker. Its value is given by the content of the instances of the database field **Statuses** that were consistent with the discourse usage being considered. Hence, the accepted values are SATELLITE_NUCLEUS, NUCLEUS_SATELLITE, and NUCLEUS_NUCLEUS.
- The feature Where to link specifies whether the rhetorical relations signaled by the discourse marker concern a textual unit that goes BEFORE or AFTER the unit that contains the marker. Its value is given by the content of the instances of the database field **Where to link** that were consistent with the discourse usage being considered.
- The feature Types of textual units specifies the nature of the textual units that are involved in the rhetorical relations. Its value is given by the

content of the instances of the database field **Types of textual units** that were consistent with the discourse usage being considered. The accepted values are CLAUSE, SENTENCE, and PARAGRAPH.

- The feature Rhetorical relation specifies the names of rhetorical relations that may be signaled by the cue phrase under consideration. Its value is given by the names listed in the instances of the database field **Rhetorical relation** that were consistent with the discourse usage being considered.
- The feature Maximal distance specifies the maximal number of units of the same kind found between the textual units that are involved in the rhetorical relation. Its value is given by the maximal value of the database field **Clause distance** of the instances that were consistent with the discourse usage being considered when the related units are clause-like units, and by the maximal value of the field **Sentence distance** when the related units are sentences. The value is 0 when the related units were adjacent in all the instances in the corpus.
- The feature Distance to salient unit is given by the maximum of the values of the database field **Distance to salient unit** of the instances that were consistent with the discourse usage being considered.

Table 6 lists the feature sets associated with the cue phrases that were initially listed in Table 2.

For example, the cue phrase *Although* has two sets of features. The first set, {SATELLITE_NUCLEUS, AFTER, CLAUSE, CONCESSION, 1, −1}, specifies that the marker signals a rhetorical relation of CONCESSION that holds between two clause-like units. The first unit has the status SATELLITE and the second has the status NUCLEUS. The clause-like unit to which the textual unit that contains the cue phrase is to be linked comes AFTER the one that contains the marker. The maximum number of clause-like units that separated two clauses related by *Although* in the corpus was one. And there were no cases in the corpus in which *Although* signaled a CONCESSION relation between a clause that preceded it and one that came after (Distance to salient unit = −1). The second set, {NUCLEUS_SATELLITE, BEFORE, SENTENCE ∨ PARAGRAPH, ELABORATION, 5, 0} specifies that the occurrence of the marker correlates with an ELABORATION relation holding between two sentences or two paragraphs. The first sentence or paragraph has the status NUCLEUS, and the second sentence or paragraph has the status SATELLITE. The sentence or paragraph to which the textual unit that contains the marker is to be linked comes BEFORE the one that contains it. The maximum number of sentences that separated two units related by *Although* in the corpus was five. And in at least one example in the corpus, *Although* marked an ELABORATION relation between some unit that preceded it and a sentence that came immediately after the one that contained the marker (Distance to salient unit = 0).

4.4.2 A Discourse-marker-based Algorithm for Hypothesizing Rhetorical Relations. At the end of step II of the rhetorical parsing algorithm (see Figure 5), the text given as input has been broken into sections, paragraphs, sentences, and clause-like units; and the cue phrases that have a discourse function have been explicitly marked. In step III.1, a set of rhetorical relations that hold between the clause-like units of each sentence, the sentences of each paragraph, and the paragraphs of each section is hypothesized, on the basis of information extracted from the corpus.

At each level of granularity (sentence, paragraph, and section levels), a discourse-marker-based hypothesizing algorithm iterates over all textual units of that level and

Table 6
The list of features sets that are used to hypothesize rhetorical relations for the discourse markers and punctuation marks shown in Table 2; N_S = NUCLEUS_SATELLITE, N_N = NUCLEUS_NUCLEUS, S_N = SATELLITE_NUCLEUS, B = BEFORE, A = AFTER, C = CLAUSE-LIKE UNIT, S = SENTENCE, and P = PARAGRAPH.

Marker	Statuses	Where to Link	Types of Textual Units	Rhetorical Relations	Maximal Distance	Distance to Salient Unit
Although	S_N	A	C	CONCESSION	1	−1
	N_S	B	S ∨ P	ELABORATION	5	0
because	S_N	A	C	CAUSE EVIDENCE	1	0
	N_S	B	C	CAUSE EVIDENCE	1	0
but	N_N	B	C	CONTRAST	1	0
for example	N_S	B	S ∨ P	EXAMPLE	2	1
where	NULL	NULL	NULL	NULL		
With	N_S	B	S ∨ P	ELABORATION	5	−1
	S_N	A	C	BACKGROUND JUSTIFICATION	0	1
Yet	S_N	B	S ∨ P	ANTITHESIS	4	1
COMMA	NULL	NULL	NULL	NULL		
OPEN_PAREN	NULL	NULL	NULL	NULL		
CLOSE_PAREN	NULL	NULL	NULL	NULL		
DASH	NULL	NULL	NULL	NULL		
END_SENTENCE	NULL	NULL	NULL	NULL		
BEGIN_PARAGRAPH	NULL	NULL	NULL	NULL		

over all discourse markers that are relevant to them. For each discourse marker, the algorithm constructs an exclusively disjunctive hypothesis concerning the rhetorical relations that the marker under scrutiny may signal. Hence, the algorithm assumes that the rhetorical structure at each level can be derived by hypothesizing rhetorical relations that hold between the units at that level. When it hypothesizes rhetorical relations that hold between clause-like units at the sentence level, it hypothesizes simple relations. When it hypothesizes rhetorical relations that hold between sentences and paragraphs (at the paragraph and section levels), it hypothesizes extended rhetorical relations. In all cases, it overgenerates exclusively disjunctive relations and subsequently uses the discourse model to determine the combinations of hypotheses that are consistent with the constraints specific to well-formed RS-trees.

Assume that the algorithm is processing the ith unit of the sequence of n units and assume that unit i contains a discourse marker that signals a rhetorical relation NAME that links the unit under scrutiny with one that went before, and whose satellite goes after the nucleus. An appropriate disjunctive hypothesis in this case is then the one that corresponds to the graphical representation in Figure 2. Such an exclusively disjunctive hypothesis enumerates all possible relations that could hold over members of the Cartesian product $\{i, i+1, \ldots, i+\text{Dist\_sal}(m)+1\} \times \{i-\text{Max}(m), i-\text{Max}(m)+1, \ldots, i-1\}$, where $\text{Max}(m)$ is the maximum number of units that separated the satellite and the nucleus of such a relation in all the examples found in the corpus, and $\text{Dist\_sal}(m)$ is the maximum distance to the salient unit found in the rightmost position. The discourse-marker-based hypothesizer iterates over all units at the sentence, paragraph, and section levels, and constructs exclusively disjunctive hypotheses such as those described here.

Let us consider, as an example, text (1). Given the textual units and the discourse markers that were identified by the clause-like unit and discourse-marker identification algorithm (see text (19)), we now examine the relations that are hypothesized by the discourse-marker-based algorithm at each level of granularity. Text (19) has three sentences that have more than one elementary unit. For the sentence shown in (20), the discourse-marker-based algorithm hypothesizes the disjunction shown in (21). This hypothesis is consistent with the information given in Table 6, which shows that, in the corpus, the marker *With* consistently signaled BACKGROUND and JUSTIFICATION relations between a satellite, the unit that contained the marker, and a nucleus, the unit that followed it.

(20) [*With* its distant orbit {— 50 percent farther from the sun than Earth —}
and slim atmospheric blanket,[1]] [Mars experiences frigid weather
conditions.[2]]

(21) $\textit{rhet\_rel}(\textsc{background}, 1, 2) \oplus \textit{rhet\_rel}(\textsc{justification}, 1, 2)$

For the sentence shown in (22), the discourse-marker-based algorithm hypothesizes the two disjunctions shown in (23).

(22) [Only the midday sun at tropical latitudes is warm enough to thaw ice
on occasion,[4]] [*but* any liquid water formed in this way would evaporate
almost instantly[5]] [*because* of the low atmospheric pressure.[6]]

(23) $$\begin{cases} \textit{rhet\_rel}(\textsc{contrast}, 4, 5) \oplus \textit{rhet\_rel}(\textsc{contrast}, 4, 6) \\ \textit{rhet\_rel}(\textsc{cause}, 6, 4) \oplus \textit{rhet\_rel}(\textsc{evidence}, 6, 4) \oplus \\ \quad \textit{rhet\_rel}(\textsc{cause}, 6, 5) \oplus \textit{rhet\_rel}(\textsc{evidence}, 6, 5) \end{cases}$$

This hypothesis is consistent with the information given in Table 6 as well: *but* signals a CONTRAST between the clause-like unit that contains the marker and a unit that went before; however, it is also possible that this relation involves the clause-like unit that comes after the one that contains the marker *but* (the **Distance to salient unit** feature has value 0), so $\mathit{rhet\_rel}(\textsc{contrast}, 4, 6)$ is hypothesized as well. The second disjunct concerns the marker *because*, which can signal either a CAUSE or an EVIDENCE relation.

For sentence (24), which is the first sentence in the second paragraph of text (1), there is only one rhetorical relation that is hypothesized, that shown in (25).

(24) [*Although* the atmosphere holds a small amount of water, and water-ice clouds sometimes develop,[7]] [most Martian weather involves blowing dust or carbon dioxide.[8]]

(25) $\mathit{rhet\_rel}(\textsc{concession}, 7, 8)$

Text (19) has two paragraphs, each of three sentences. The first paragraph contains no discourse markers that could signal relations between sentences. Hence, the discourse-marker-based algorithm does not make any hypotheses of rhetorical relations that hold among the sentences of the first paragraph. In contrast, when the discourse-marker-based algorithm examines the markers of the second paragraph, it hypothesizes that a rhetorical relation of type EXAMPLE holds either between sentences 9 and $[7, 8]$ or between sentences 10 and $[7, 8]$, because the discourse marker *for example* is used in sentence 9. This is consistent with the information presented in Table 6, which specifies that a rhetorical relation of EXAMPLE holds between a satellite, the sentence that contains the marker, and a nucleus, the sentence that went before. However, the satellite of the relation could also be the sentence that follows the sentence that contains the discourse marker (the value of the **Distance to salient unit** feature is 0). Given the marker *Yet*, the discourse-marker-based algorithm hypothesizes that an ANTITHESIS relation holds between a sentence that preceded the one that contains the marker, and the sentence that contains it. The set of disjuncts shown in (26) represents all the hypotheses that are made by the algorithm. Note that these hypotheses concern extended rhetorical relations.

(26)
$$\begin{cases} \mathit{rhet\_rel}(\textsc{example}, 9, [7, 8]) \oplus \mathit{rhet\_rel}(\textsc{example}, 10, [7, 8]) \\ \mathit{rhet\_rel}(\textsc{antithesis}, 9, 10) \oplus \mathit{rhet\_rel}(\textsc{antithesis}, [7, 8], 10) \end{cases}$$

During the corpus analysis, I was not able to make a connection between discourse markers that signal sentence-level rhetorical relations and relations that hold between sequences of sentences, paragraphs, and multiparagraphs. However, I noticed that a discourse marker signals a paragraph-level rhetorical relation when the marker under scrutiny is located either at the end of the first paragraph or at the beginning of the second paragraph. The rhetorical parser implements this observation by assuming that rhetorical relations between paragraphs can be signaled only by markers that occur in the first sentence of the paragraph, when the marker signals a relation whose other unit precedes the marker, or in the last sentence of the paragraph, when the marker signals a relation whose other unit follows the marker. According to the results derived from the corpus analysis, the use of the discourse marker *Although* at the beginning of a sentence or paragraph correlates with a rhetorical relation of ELABORATION that holds between a satellite, the sentence or paragraph that contains the marker, and a nucleus, the sentence or paragraph that precedes it. The discourse-marker-based algorithm hypothesizes only one rhetorical relation that holds between the two paragraphs of text (19), that shown in (27), below.

(27) $\mathit{rhet\_rel}(\textsc{elaboration}, [7, 10], [1, 6])$

When a section has more than two paragraphs, the rhetorical parser generates exclusively disjunctive hypotheses at the paragraph level as well. The current implementation of the rhetorical parser does not hypothesize any relations among the sections of a text.

4.4.3 A Word-co-occurrence-based Algorithm for Hypothesizing Rhetorical Relations. The rhetorical relations hypothesized by the discourse-marker-based algorithm rely entirely on occurrences of discourse markers. In building the valid rhetorical structures of sentences, the set of rhetorical relations that are hypothesized on the basis of discourse marker occurrences provides sufficient information. After all, the clause-like units of a sentence are determined on the basis of discourse marker occurrences as well; so every unit of a sentence is related to at least one other unit of the same sentence. This might not be the case when we consider the paragraph and section levels, however, because discourse markers might not provide sufficient information for hypothesizing rhetorical relations among all sentences of a paragraph and among all paragraphs of a text. In fact, it is even possible that there are full paragraphs that use no discourse markers at all; or that use only markers that link clause-like units within sentences.

In step III.2, the rhetorical parser uses cohesion (Halliday and Hasan 1976; Hearst 1997; Hoey 1991; Salton et al. 1995) to hypothesize rhetorical relations. The algorithm that hypothesizes such rhetorical relations assumes that if two sentences or paragraphs talk about the same thing, it is either the case that the sentence or paragraph that comes later ELABORATES on the topic of the sentence or paragraph that went before; or that the sentence or paragraph that comes before provides the BACKGROUND for interpreting the sentence or paragraph that comes later. If two sentences or paragraphs talk about different things, it is assumed that a multinuclear JOINT relation holds between the two units. The decision as to whether two sentences/paragraphs talk about the same thing is made by measuring the similarity between the sentences/paragraphs. If this similarity is above a certain threshold, the textual units are considered to be related. Otherwise, a JOINT relation is assumed to hold between the two units.

Once the discourse-marker-based algorithm has hypothesized all relations it could, a word-co-occurrence-based algorithm examines every sentence/paragraph boundary for which a marker-based rhetorical relation has not been hypothesized and uses cohesion to produce such a hypothesis. As in the case of the discourse-marker-based algorithm, each hypothesis is an exclusive disjunction over the members of the Cartesian product $\{i - LD, \ldots, i\} \times \{i + 1, \ldots, i + RD\}$, which contains the units found to the left and to the right of the boundary between units i and $i + 1$. Variables LD and RD represent arbitrarily set sizes of the spans that are considered to be relevant from a cohesion-based perspective. The current implementation of the rhetorical parser sets LD to 3 and RD to 2.

To assess the similarity between two units $l \in \{i - LD, \ldots, i\}$ and $r \in \{i + 1, \ldots, i + RD\}$, stopwords such as *the*, *a*, and *and* are initially eliminated from the texts that correspond to these units. The suffixes of the remaining words are removed as well, so that words that have the same root can be accounted for by the similarity measurement even if they are used in different cases, moods, tenses, etc. If the similarity is above a certain threshold, an ELABORATION or a BACKGROUND relation is hypothesized to hold between two units; otherwise, a JOINT relation is hypothesized. The value of the threshold is computed for each type of textual unit on the basis of the average similarity of all textual units at that level.

As we have already discussed, the first paragraph in text (19) contains no discourse markers that could signal relations between sentences. When the word-co-occurrence-based algorithm examines the boundary between the first two sentences, no stemmed words are found to co-occur in the first two sentences, but the stem *sun* is found to co-occur in the first and third sentences. Therefore, the algorithm hypothesizes the first disjunct in (28). When the boundary between the last two sentences is examined, a disjunct having the same form is hypothesized (the last two sentences of the first paragraph have no words in common). To distinguish between the two different sources that generated the disjuncts, I assign different subscripts to the rhetorical relations shown in (28).

$$\tag{28} \begin{cases} \mathit{rhet\_rel}(\textsc{joint}_1, [1,2], 3) \oplus \mathit{rhet\_rel}(\textsc{elaboration}_1, [4,6], [1,2]) \oplus \\ \qquad \mathit{rhet\_rel}(\textsc{background}_1, [1,2], [4,6]) \\ \mathit{rhet\_rel}(\textsc{elaboration}_2, [4,6], [1,2]) \oplus \mathit{rhet\_rel}(\textsc{background}_2, [1,2], [4,6]) \oplus \\ \qquad \mathit{rhet\_rel}(\textsc{joint}_2, 3, [4,6]) \end{cases}$$

During my corpus study, I noticed that in most of the cases in which the number of sentences in a paragraph or the number of paragraphs in a section was small and no discourse markers were used, the relation that held between the sentences/paragraphs was ELABORATION. The rhetorical parser implements this empirical observation as well. Since the first paragraph in text (1) has only three sentences and no discourse marker can be used to hypothesize rhetorical relations that hold between these sentences, the word-co-occurrence-based algorithm hypothesizes the relations shown in (29).

$$\tag{29} \begin{cases} \mathit{rhet\_rel}(\textsc{elaboration}, 3, [1,2]) \\ \mathit{rhet\_rel}(\textsc{elaboration}, [4,6], 3) \end{cases}$$

4.5 A Proof-Theoretic Account of the Problem of Rhetorical Structure Derivation

Once the elementary units of a text have been determined and the rhetorical relations between them have been hypothesized at sentence, paragraph, and section levels, we need to determine the rhetorical structures that are consistent with these hypotheses and with the constraints specific to valid RS-trees. That is, we need to solve the problem of rhetorical structure derivation.

One way to formalize the problem of rhetorical structure derivation is to assume that given as input a set of units $U = 1, 2, \ldots, n$ and a set RR of simple, extended, and exclusively disjunctive hypotheses that hold between these units, we are interested in deriving objects of the form $\mathit{tree}(\mathit{status}, \mathit{type}, \mathit{promotion}, \mathit{left}, \mathit{right})$, where *status* can be either NUCLEUS or SATELLITE; *type* can be a name of a rhetorical relation; *promotion* can be a set of natural numbers from 1 to N; and *left* and *right* can be either NULL or recursively defined objects of type *tree*.

The objects having the form $\mathit{tree}(\mathit{status}, \mathit{type}, \mathit{promotion}, \mathit{left}, \mathit{right})$ provide a functional representation of valid rhetorical structures. For example, with respect to the elementary units of text (4) and the rhetorical relations that hold between the units of this text (see (6)), the subtree in Figure 3 that subsumes units 1 and 2 can be represented functionally using an object of type *tree* as shown in (30).

$$\begin{aligned} &\mathit{tree}(\textsc{nucleus}, \textsc{elaboration}, \{1\},\\ &\quad \mathit{tree}(\textsc{nucleus}, \textsc{leaf}, \{1\}, \textsc{null}, \textsc{null}),\\ &\quad \mathit{tree}(\textsc{satellite}, \textsc{leaf}, \{2\}, \textsc{null}, \textsc{null})) \end{aligned} \tag{30}$$

Using objects of type *tree*, I devised a proof theory that can be used to determine all valid rhetorical structures of a text. The theory consists of a set of axioms and rewriting rules that encode all possible ways in which one can derive the valid RS-trees of a text. In this paper, I present the proof theory only at the intuitive level. The interested reader can find further detail in Marcu (2000).

The proof theory that I outline here assumes that the problem of rhetorical structure derivation can be encoded as a rewriting problem in which valid RS-trees are constructed bottom-up. Initially, each elementary unit i in the input is associated with an elementary tree that has either status NUCLEUS or SATELLITE, type LEAF, and promotion set $\{i\}$. In the beginning, any of the hypothesized relations RR can be used to join these elementary trees into more complex trees. Once the elementary trees have been built, the rhetorical structure is constructed by joining adjacent trees into larger trees and by making sure that at every step, the resulting structure is valid. The set of rhetorical relations associated with each tree keeps track of the rhetorical relations that can still be used to extend that tree. In the beginning, an elementary tree can be extended using any of the hypothesized relations RR, but as soon as a relation is used, it becomes unavailable for subsequent extensions.

We encode the derivation of the elementary trees using axioms (31) and (32). Axiom (31), for example, specifies that if i is an elementary unit in U and if relations RR have been hypothesized to hold between the units in U, then one can build an elementary tree across text span $[i, i]$, having the status NUCLEUS, the type LEAF, and promotion set $\{i\}$; and that this tree can be rewritten into a larger tree by using relations from the set RR. Hence, the last argument RR enumerates the hypotheses that can be used to expand the *tree* that characterizes the text span under consideration.

$$[\mathit{unit}(i) \land \mathit{hold}(RR)] \rightarrow S(i, i, \mathit{tree}(\textsc{nucleus}, \textsc{leaf}, \{i\}, \textsc{null}, \textsc{null}), RR) \tag{31}$$

$$[\mathit{unit}(i) \land \mathit{hold}(RR)] \rightarrow S(i, i, \mathit{tree}(\textsc{satellite}, \textsc{leaf}, \{i\}, \textsc{null}, \textsc{null}), RR) \tag{32}$$

A set of 12 axioms (rewriting rules) explains how trees can be assembled into larger trees in a bottom-up fashion. Let us focus for the moment on the pair of axioms (33) and (34), which are given below.

$$\begin{aligned} &[S(l, b, \mathit{tree}_1(\textsc{nucleus}, \mathit{type}_1, p_1, \mathit{left}_1, \mathit{right}_1), rr_1) \land\\ &\ S(b+1, h, \mathit{tree}_2(\textsc{satellite}, \mathit{type}_2, p_2, \mathit{left}_2, \mathit{right}_2), rr_2) \land\\ &\ \mathit{rhet\_rel}(\mathit{name}, s, n) \in_{\oplus} rr_1 \land \mathit{rhet\_rel}(\mathit{name}, s, n) \in_{\oplus} rr_2 \land\\ &\ s \in p_2 \land n \in p_1 \land \mathit{hypotactic}(\mathit{name})] \rightarrow\\ &S(l, h, \mathit{tree}(\textsc{nucleus}, \mathit{name}, p_1, \mathit{tree}_1(\ldots), \mathit{tree}_2(\ldots)),\\ &\qquad\qquad\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{\mathit{rhet\_rel}(\mathit{name}, s, n)\}) \end{aligned} \tag{33}$$

$$\begin{aligned} &[S(l, b, \mathit{tree}_1(\textsc{nucleus}, \mathit{type}_1, p_1, \mathit{left}_1, \mathit{right}_1), rr_1) \land\\ &\ S(b+1, h, \mathit{tree}_2(\textsc{satellite}, \mathit{type}_2, p_2, \mathit{left}_2, \mathit{right}_2), rr_2) \land\\ &\ \mathit{rhet\_rel}(\mathit{name}, s, n) \in_{\oplus} rr_1 \land\ \mathit{rhet\_rel}(\mathit{name}, s, n) \in_{\oplus} rr_2 \land\\ &\ s \in p_2 \land n \in p_1 \land \mathit{hypotactic}(\mathit{name})] \rightarrow\\ &S(l, h, \mathit{tree}(\textsc{satellite}, \mathit{name}, p_1, \mathit{tree}_1(\ldots), \mathit{tree}_2(\ldots)),\\ &\qquad\qquad\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{\mathit{rhet\_rel}(\mathit{name}, s, n)\}) \end{aligned} \tag{34}$$

Assume that there exist two spans: one from unit l to unit b that is characterized by valid rhetorical structure $\mathit{tree}_1(\ldots)$ and rhetorical relations rr_1, and the other from unit $b+1$ to unit h that is characterized by valid rhetorical structure $\mathit{tree}_2(\ldots)$ and rhetorical relations rr_2. Assume also that rhetorical relation $\mathit{rhet\_rel}(\mathit{name}, s, n)$ holds between a unit s that is in the promotion set of span $[b+1, h]$ and a unit n that is in the promotion set of span $[l, b]$, that $\mathit{rhet\_rel}(\mathit{name}, s, n)$ can be used to extend both spans $[l, b]$ and $[b+1, h]$ ($\mathit{rhet\_rel}(\mathit{name}, s, n) \in_{\oplus} rr_1$ and $\mathit{rhet\_rel}(\mathit{name}, s, n) \in_{\oplus} rr_2$), and that the relation is hypotactic. In such a case, one can combine spans $[l, b]$ and $[b+1, h]$ into a larger span $[l, h]$ that has a valid structure whose status is either NUCLEUS (see axiom (33)) or SATELLITE (see axiom (34)), type *name*, promotion set p_1, and whose children are given by the valid structures of the immediate subspans. The set of rhetorical relations that can be used to further extend this structure is given by $rr_1 \cap rr_2 \setminus_{\oplus} \{\mathit{rhet\_rel}(\mathit{name}, s, n)\}$.

We use operators $\in_{\oplus}$ and $\setminus_{\oplus}$ instead of $\in$ and $\setminus$ because we treat each exclusive disjunction as a whole because each exclusive disjunction was hypothesized using one and only one hypothesis trigger (cue phrase). That is, we say that a rhetorical relation $r \in_{\oplus} r_i \oplus r_{i+1} \oplus \cdots \oplus r_{i+j}$ if r matches any of the rhetorical relations $r_i \cdots r_{i+j}$. We consider that the result of the difference $RR_i \setminus_{\oplus} r$ is a subset of RR_i that contains all the members of RR_i except the exclusive disjunction that uses relation r. Because axioms (33) and (34) treat each exclusive disjunction as a whole, they ensure that no rhetorical relation occurs more than once in a discourse structure.

Similarly, we can define rules of inference for the cases in which an extended rhetorical relation holds across spans $[l, b]$ and $[b+1, h]$; for the cases in which the satellite precedes the nucleus; and for the cases in which the relation under scrutiny is paratactic. (See Marcu [2000] for a complete list of these axioms.) Rule (35), for examples, corresponds to the case in which the relation under scrutiny is a simple, paratactic relation.

(35)
$$\begin{aligned}
&[S(l, b, \mathit{tree}_1(\textsc{nucleus}, \mathit{type}_1, p_1, \mathit{left}_1, \mathit{right}_1), rr_1) \wedge \\
&S(b+1, h, \mathit{tree}_2(\textsc{nucleus}, \mathit{type}_2, p_2, \mathit{left}_2, \mathit{right}_2), rr_2) \wedge \\
&\mathit{rhet\_rel}(\mathit{name}, n_1, n_2) \in_{\oplus} rr_1 \wedge \mathit{rhet\_rel}(\mathit{name}, n_1, n_2) \in_{\oplus} rr_2 \wedge \\
&n_1 \in p_1 \wedge n_2 \in p_2 \wedge \mathit{paratactic}(\mathit{name})] \rightarrow \\
&S(l, h, \mathit{tree}(\textsc{nucleus}, \mathit{name}, p_1 \cup p_2, \mathit{tree}_1(\ldots), \mathit{tree}_2(\ldots)), \\
&\quad rr_1 \cap rr_2 \setminus_{\oplus} \{\mathit{rhet\_rel}(\mathit{name}, n_1, n_2)\})
\end{aligned}$$

Example of a Derivation of a Valid Rhetorical Structure. If we take any text of N units that is characterized by a set RR of rhetorical relations, the proof-theoretic account provides all the necessary support for deriving the valid rhetorical structures of that text. Assume, for example, that we are given text (4), among which rhetorical relations RR given in (6), hold. In Figure 7, we sketch the derivation of the theorem that corresponds to the valid rhetorical structure shown in Figure 3. The relations RR_1 and RR_2 that the derivation refers to are shown below.

(36)
$$RR_1 = \begin{cases} \mathit{rhet\_rel}(\textsc{contrast}, 1, 3) \oplus \mathit{rhet\_rel}(\textsc{contrast}, 1, 4) \oplus \\ \quad \mathit{rhet\_rel}(\textsc{contrast}, 2, 3) \oplus \mathit{rhet\_rel}(\textsc{contrast}, 2, 4) \\ \mathit{rhet\_rel}(\textsc{elaboration}, 4, 1) \oplus \mathit{rhet\_rel}(\textsc{elaboration}, 4, 2) \oplus \\ \quad \mathit{rhet\_rel}(\textsc{elaboration}, 4, 3) \end{cases}$$

(37)
$$RR_2 = \begin{cases} \mathit{rhet\_rel}(\textsc{contrast}, 1, 3) \oplus \mathit{rhet\_rel}(\textsc{contrast}, 1, 4) \oplus \\ \quad \mathit{rhet\_rel}(\textsc{contrast}, 2, 3) \oplus \mathit{rhet\_rel}(\textsc{contrast}, 2, 4) \\ \mathit{rhet\_rel}(\textsc{elaboration}, 2, 1) \end{cases}$$

1. $\mathit{hold}(RR)$	Input
2. $\mathit{unit}(1)$	Input
3. $\mathit{unit}(2)$	Input
4. $\mathit{unit}(3)$	Input
5. $\mathit{unit}(4)$	Input
6. $S(1,1,\mathit{tree}(\textsc{nucleus},\textsc{leaf},\{1\},\textsc{null},\textsc{null}),RR)$	1, 2, Axiom (31), MP
7. $S(2,2,\mathit{tree}(\textsc{satellite},\textsc{leaf},\{2\},\textsc{null},\textsc{null}),RR)$	1, 3, Axiom (32), MP
8. $S(1,2,\mathit{tree}(\textsc{nucleus},\textsc{elaboration},\{1\},$ $\mathit{tree}(\textsc{nucleus},\textsc{leaf},\{1\},\textsc{null},\textsc{null}),$ $\mathit{tree}(\textsc{satellite},\textsc{leaf},\{2\},\textsc{null},\textsc{null}),$ $RR_1))$	6, 7, Axiom (33), MP
9. $S(3,3,\mathit{tree}(\textsc{nucleus},\textsc{leaf},\{3\},\textsc{null},\textsc{null}),RR)$	1, 4, Axiom (31) , MP
10. $S(4,4,\mathit{tree}(\textsc{satellite},\textsc{leaf},\{4\},\textsc{null},\textsc{null}),RR)$	1, 5, Axiom (32) , MP
11. $S(3,4,\mathit{tree}(\textsc{nucleus},\textsc{elaboration},\{3\},$ $\mathit{tree}(\textsc{nucleus},\textsc{leaf},\{3\},\textsc{null},\textsc{null}),$ $\mathit{tree}(\textsc{satellite},\textsc{leaf},\{4\},\textsc{null},\textsc{null}),$ $RR_2))$	9, 10, Axiom (33), MP
12. $S(1,4,\mathit{tree}(\textsc{nucleus},\textsc{contrast},\{1,3\},$ $\mathit{tree}(\textsc{nucleus},\textsc{elaboration},\{1\},$ $\mathit{tree}(\textsc{nucleus},\textsc{leaf},\{1\},\textsc{null},\textsc{null}),$ $\mathit{tree}(\textsc{satellite},\textsc{leaf},\{2\},\textsc{null},\textsc{null})),$ $\mathit{tree}(\textsc{nucleus},\textsc{elaboration},\{3\},$ $\mathit{tree}(\textsc{nucleus},\textsc{leaf},\{3\},\textsc{null},\textsc{null}),$ $\mathit{tree}(\textsc{satellite},\textsc{leaf},\{4\},\text{sc null},\textsc{null}))),$ $\emptyset)$	8, 11, Axiom (35), MP

Figure 7
A derivation of the theorem that corresponds to the valid rhetorical structure shown in Figure 3.

The derivation starts with five axioms that are straightforwardly derived from the input of the problem. Using the axioms in lines 1 and 2, axiom (31), and the modus ponens rule, we derive the theorem in line 6. Using the axioms in lines 1 and 3, axiom (32), and modus ponens, we derive the theorem in line 7. Similarly, we derive the theorems in lines 9 and 10. These four theorems all correspond to valid rhetorical structures that can be built on top of elementary units. Using the theorems in lines 6 and 7, axiom (33), and modus ponens, we derive the theorem in line 8. It corresponds to a valid rhetorical structure that can be built across span $[1,2]$. Since this structure uses rhetorical relation $\mathit{rhet\_rel}(\textsc{elaboration},2,1)$, the set of rhetorical relations that can be used to further expand the rhetorical structure will be given by the set RR_1, shown in (36). Line 11 corresponds to a valid rhetorical structure that can be built on top of elementary span [3,4]. Since this structure uses rhetorical relation $\mathit{rhet\_rel}(\textsc{elaboration},4,3)$, the set of rhetorical relations that can be used to further expand the rhetorical structure will be given by the set RR_2, shown in (37). Using the theorems derived in lines 8 and 11, axiom (35), and modus ponens gives us the theorem in line 12 that corresponds to a valid structure for the entire text, the structure shown in Figure 3.

As I have shown in Marcu (2000), the proof-theoretic account outlined here is both sound and complete with respect to the constraints that characterize the valid rhetorical structures enumerated in Section 2.4. That is, all theorems that are derived using the disjunctive proof-theoretic account correspond to valid text structures; and

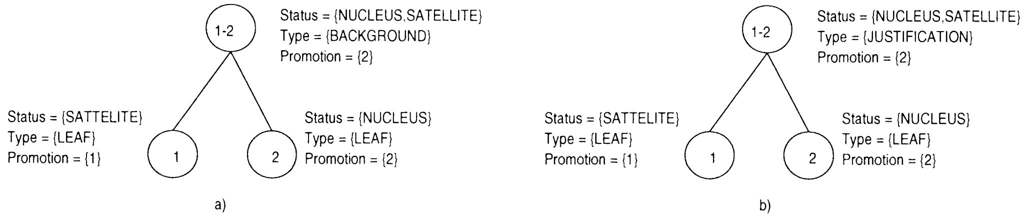

Figure 8
All valid rhetorical structures of sentence (20).

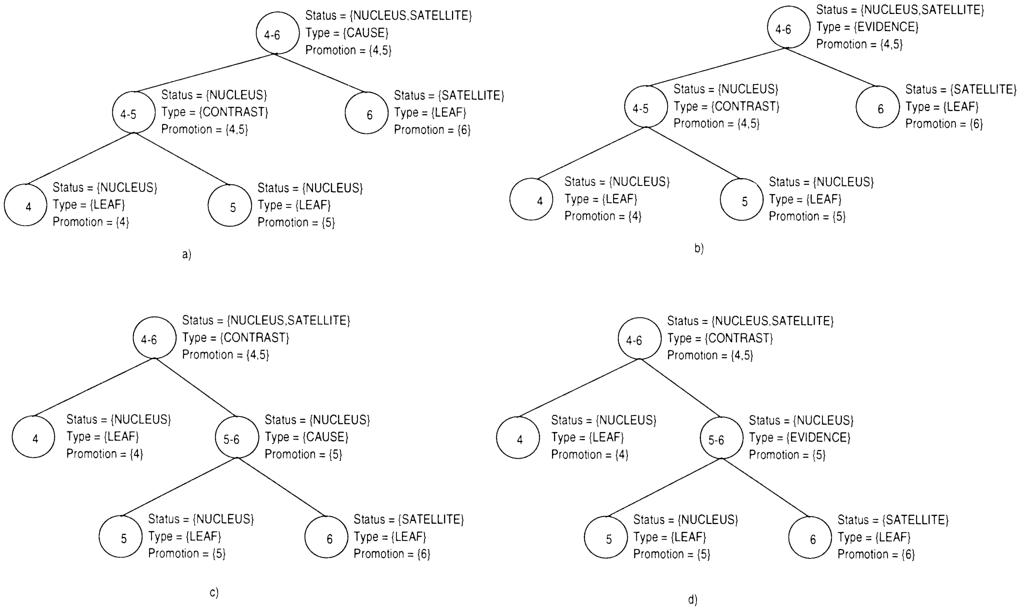

Figure 9
All valid rhetorical structures of sentence (22).

any valid rhetorical structure can be derived through the successive application of modus ponens and the axioms of the disjunctive proof-theoretic account.

Implementing the Proof-Theoretic Account. There are many ways in which one can implement the proof theory described in this section. Since all axioms of the theory are Horn clauses, they can be immediately translated into a Prolog program. Equally trivial is to implement the proof-theoretic account using traditional parsing techniques that combine terminal and nonterminal symbols only when the constraints enumerated in the axioms of the proof-theoretic account are satisfied. The rhetorical parser implements the proof-theoretic account as a chart-parsing algorithm (see Marcu [2000] for details). When a chart-parsing implementation uses as input the rhetorical relations that were hypothesized by the discourse-marker- and word-co-occurrence-based algorithms at the sentence, paragraph, and section levels of text (19), it derives the valid rhetorical structures shown in Figures 8–13.

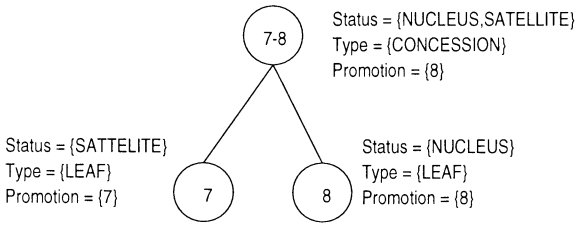

Figure 10
The valid rhetorical structure of sentence (24).

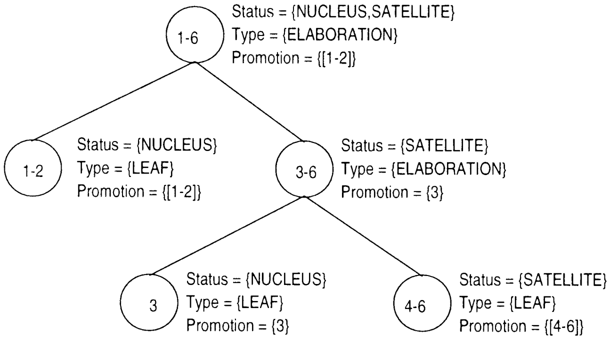

Figure 11
The valid rhetorical structure of the first paragraph of text (19); see the relations in (29).

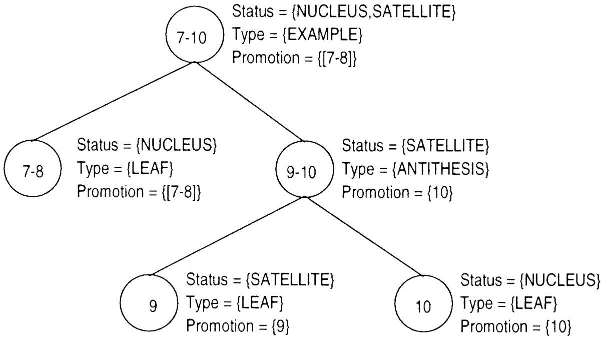

Figure 12
The valid rhetorical structure of the second paragraph of text (19); see the relations in (26).

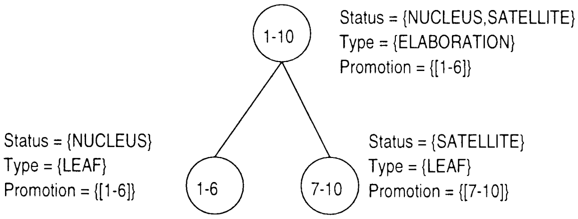

Figure 13
The valid rhetorical structure of text (19); see the relation in (27).

4.6 The Ambiguity of Discourse

4.6.1 A Weight Function for Rhetorical Structures. Discourse is ambiguous the same way sentences are: usually, more than one discourse structure is produced for any given text. For example, we have seen that the rhetorical parser finds four different valid rhetorical structures for sentence (22) (see Figure 9). In my experiments, I noticed that the "best" discourse trees are usually those that are skewed to the right. I believe that the explanation for this observation is that text processing is essentially a left-to-right process. Usually, people write texts so that the most important ideas go first, both at the paragraph and at the text level. In fact, journalists are trained to consciously employ this "pyramid" approach to writing (Cumming and McKercher 1994). The more text writers add, the more they elaborate on the text that went before: as a consequence, incremental discourse building consists mostly of expansion of the right branches. A preference for trees that are skewed to the right is also consistent with research in psycholinguistics that shows that readers have a preference for interpreting unmarked textual units as continuations of the topics of the units that precede them (Segal, Duchan, and Scott 1991). At the structural level, this corresponds to textual units that elaborate on the information that has been presented before.

In order to disambiguate the discourse, the rhetorical parser computes a weight for each valid discourse tree and retains only the trees that are maximal. The weight function w, which is shown in (38), is computed recursively by summing up the weights of the left and right branches of a rhetorical structure and the difference between the depth of the right and left branches of the structure. Hence, the more skewed to the right a tree is, the greater its weight w is.

(38) $$w(tree) = \begin{cases} 0 & \text{if } isLeaf(tree), \\ w(leftOf(tree)) + w(rightOf(tree)) + & \text{otherwise.} \\ \qquad depth(rightOf(tree)) - depth(leftOf(tree)) & \end{cases}$$

For example, when applied to the valid rhetorical structures of sentence (22), the weight function will assign the value -1 to the trees shown in Figures 9(a) and 9(b), and the value $+1$ to the trees shown in Figures 9(c) and 9(d).

4.6.2 The Ambiguity of Discourse—An Implementation Perspective. There are two ways one can disambiguate discourse. One way is to consider, during the parsing process, all of the valid rhetorical structures of a text. When the parsing is complete, the structures of maximal weight can be then assigned to the text given as input. The other way is to consider, during the parsing process, only the partial structures that could lead to a structure of maximal weight. For example, if a chart parsing algorithm is used, we can keep in the chart only the partial structures that could lead to a final structure of maximal weight.

In step III.4, the rhetorical parser shown in Figure 5 implements the second approach. Hence, instead of keeping all the partial structures that characterize sentence (22), it will keep only the partial structures of maximal weight, i.e., the structures shown in Figures 9(c) and 9(d). In this way, the overall efficiency of the system is increased.

When the rhetorical parser selects the trees of maximal weight for text (19), at each of the three levels of abstraction, it selects the trees shown in Figures 8(a), 9(c), 10, 11, 12, and 13. If no weight function were used, the rhetorical parser would generate eight distinct valid rhetorical structures for the whole text.

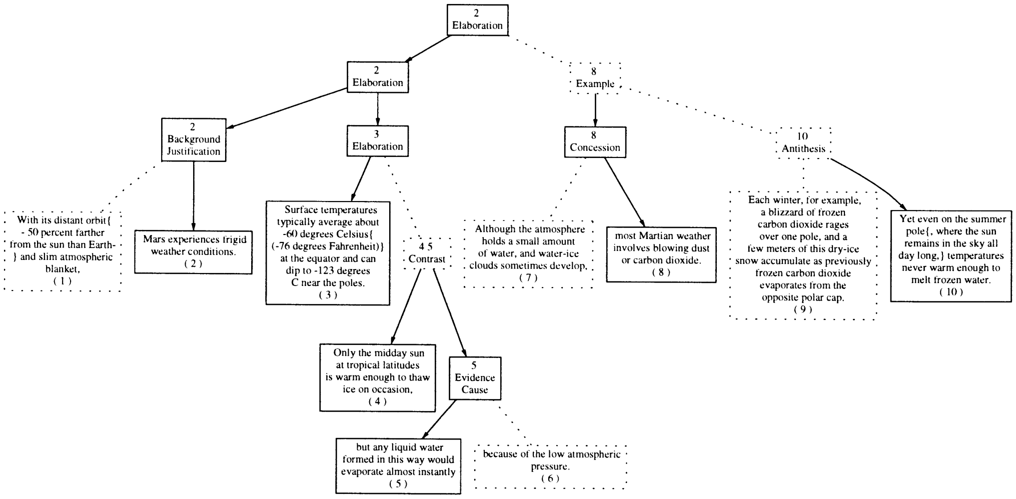
2
Elaboration
2
Elaboration
8
Example
2
Background
Justification
3
Elaboration
8
Concession
10
Antithesis
With its distant orbit{ - 50 percent farther from the sun than Earth- } and slim atmospheric blanket,
(1)
Mars experiences frigid weather conditions.
(2)
Surface temperatures typically average about -60 degrees Celsius{ (-76 degrees Fahrenheit)} at the equator and can dip to -123 degrees C near the poles.
(3)
4 5
Contrast
Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop,
(7)
most Martian weather involves blowing dust or carbon dioxide.
(8)
Each winter, for example, a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap.
(9)
Yet even on the summer pole{, where the sun remains in the sky all day long,} temperatures never warm enough to melt frozen water.
(10)
Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,
(4)
5
Evidence
Cause
but any liquid water formed in this way would evaporate almost instantly
(5)
because of the low atmospheric pressure.
(6)

Figure 14
The discourse tree of maximal weight that is built by the rhetorical parsing algorithm for text (1). Nuclei are surrounded by solid boxes and satellites by dotted boxes; links between a node and the subordinate nucleus or nuclei are represented by solid arrows; links between a node and the subordinate satellites by dotted lines. Occurrences of parenthetical information are enclosed in the text by curly brackets; the leaves of the discourse structure are numbered from 1 to N, where N represents the number of elementary units in the whole text. The numbers associated with each node denote the units that are members of its promotion set.

4.7 Deriving the Final Rhetorical Structure

In the last step (lines 16–17 in Figure 5), after the trees of maximal weight have been obtained at the sentence, paragraph, and section levels, the rhetorical parser merges the valid structures into a structure that spans the whole text of a section. In this way, the rhetorical parser builds one tree for each of the sections of a given document. The merging process is a trivial procedure that assembles the trees obtained at each level of granularity. That is, the trees that correspond to the sentence level are substituted for the leaves of the structures built at the paragraph level, and the trees that correspond to the paragraph levels are substituted for the leaves of the structures built at the section level. The promotion units associated with each span are recomputed in a bottom-up fashion so that they correspond to elementary units and not to sentence and paragraph labels. The rhetorical parser has a back-end process that uses "dot," a preprocessor for drawing oriented graphs, to automatically generate PostScript representations of the rhetorical structures of maximal weight. When applied to text (1), the rhetorical parser builds the rhetorical structure shown in Figure 14.

5. Evaluation

There are two ways to evaluate the correctness of the discourse trees that an automatic process builds. One is to compare the automatically derived trees with trees that have been built manually. The other is to evaluate the impact that they have on the accuracy of other natural language processing tasks, such as anaphora resolution, intention recognition, or text summarization. The rhetorical parser presented here was evaluated by following both of these avenues.

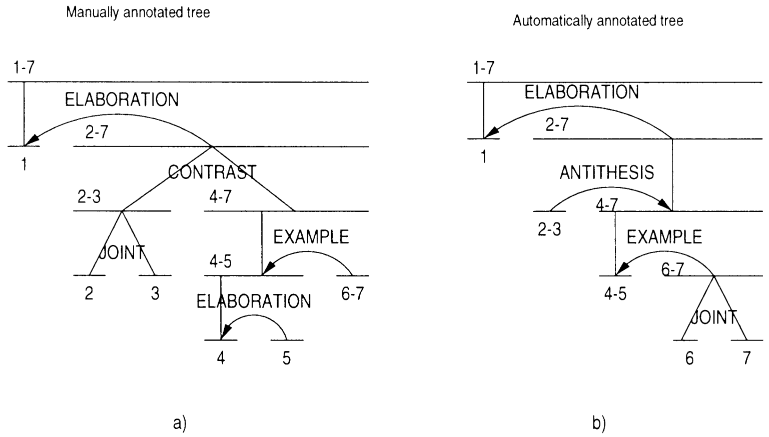

Figure 15
Example of discourse trees: (a) represents a manually built tree; (b) represents an automatically built tree.

5.1 Evaluating the Correctness of the Trees

5.1.1 Labeled Recall and Precision Figures. To evaluate the correctness of the trees built by the rhetorical parser, two analysts have manually built the rhetorical structure of five texts from *Scientific American*, which ranged in size from 161 to 725 words. The analysts were computational linguists who were familiar with Rhetorical Structure Theory (Mann and Thompson 1988). They did not agree beforehand on any annotation style or protocol and were not given any specific instructions besides being asked to build trees that were consistent with the requirements put forth by Mann and Thompson. The analysts were supposed to use only the set of relations proposed by RST and the relation TEXTUAL to link the subtrees subsuming the title and body of a text. Analysts were not asked to build binary structures (similar to those derived by the rhetorical parser), although we knew that this could negatively affect the performance of our system.

The performance of the rhetorical parser was estimated by applying labeled recall and precision measures, which are extensively used to study the performance of syntactic parsers. Labeled recall reflects the number of correctly labeled constituents identified by the rhetorical parser with respect to the number of labeled constituents in the corresponding manually built tree. Labeled precision reflects the number of correctly labeled constituents identified by the rhetorical parser with respect to the total number of labeled constituents identified by the parser. Labeled recall and precision figures were computed with respect to the ability of the rhetorical parser to identify elementary units, hierarchical text spans, text span nuclei and satellites, and rhetorical relations.

To understand how these figures were computed, assume for example that an analyst identified six elementary units in a text and built the discourse structure in Figure 15(a) and that the program identified five elementary units and built the discourse structure in Figure 15(b). When we align the two structures, we obtain the labels in Table 7, which show that the program did not identify the breaks between units 2 and 3, and 4 and 5 in the analyst's annotation; and that it considered the unit labeled 6-7 in the analyst's annotation to be made of two units. Table 7 lists all constituents in the two structures, the associated labels at the elementary unit, span, nuclei, and rhetorical levels, and the corresponding recall and precision figures. As Table 7 shows,

Table 7
Computing the performance of a rhetorical parser (P = Program; A = Analyst).

Constituent	Units		Spans		Nuclearity		Relations	
	A	P	A	P	A	P	A	P
1-1	*	*	*	*	N	N	Span	Span
2-2	*		*		N		Joint	
3-3	*		*		N		Joint	
4-4	*		*		N		Span	
5-5	*		*		S		elaboration	
6-6		*		*		N		Joint
7-7		*		*		N		Joint
2-3		*	*	*	N	S	Contrast	antithesis
4-5		*	*	*	N	N	Span	Span
6-7	*		*	*	S	S	example	example
4-7			*	*	N	N	Contrast	Span
2-7			*	*	S	S	elaboration	elaboration
	R = 1/6 P = 1/5		R = 6/10 P = 6/8		R = 5/10 P = 5/8		R = 4/10 P = 4/8	

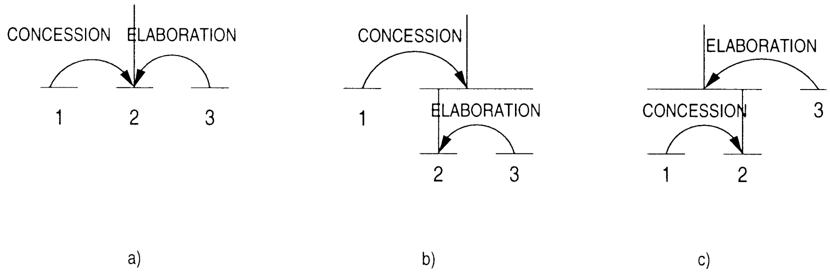

Figure 16
Discourse trees (b) and (c) represent alternative binary representations of the nonbinary discourse tree in (a).

the program in this example identified only one of the six elementary units identified by the analyst (unit 1), for a recall of 1/6. Since the program identified a total of five units, the precision is 1/5. Similarly, recall and precision figures can be computed for span, nuclearity, and rhetorical relation assignments.

This evaluation assumes that rhetorical labels are associated with the children nodes, and not with the father nodes, as in the formalization. For example, the example relation that holds between spans [4,5] and [6,7] in the tree in Figure 15(a), is not associated with span [4,7], but rather, with the span [6,7], which is the satellite of the relation; and by convention, the rhetorical relation of the span [4,5] is set to Span. The rationale for this choice is the fact that the analysts did not construct only binary trees; some of the nodes in their manually built representations had multiple children. Representing in binary form a tree such as that shown in Figure 16(a), for example, would require an additional hierarchical level on the spans, as shown in Figures 16(b) and 16(c), that was not part of the original analysis. To avoid introducing in the annotation choices that were not part of what the analysts did, I decided for the purpose of evaluation to follow the procedure outlined above.

Table 8 shows average recall and precision figures that reflect the performance of the rhetorical parser on the five *Scientific American* texts. In addition to the recall

Table 8
Performance of the rhetorical parser.

	Analysts		Program	
	Recall	Precision	Recall	Precision
Elementary units	87.9	87.9	51.2	95.9
Spans	89.6	89.6	63.5	87.7
Nuclearity	79.4	88.2	50.6	85.1
Relations	83.4	83.4	47.0	78.4

and precision figures specific to the program, Table 8 also displays average recall and precision figures obtained for the trees built only by the analysts. These figures reflect how similar the annotations of the two analysts were and provides an upper bound for the performance of the rhetorical parser: if the recall and precision figures of the parser were the same as the figures for the analysts, the discourse trees built by the rhetorical parser would be indistinguishable from those built by a human.

As the results in Table 8 show, the rhetorical parser fails to identify a fair number of elementary units (51.2% recall); but the units it identifies tend to be correct (95.9% precision). As a consequence, performance at all other levels is affected. With respect to identifying hierarchical spans, recall is about 25% lower than the average human performance; with respect to labeling the nuclear status of spans, recall is about 30% below human performance; and with respect to labeling the rhetorical relations that hold between spans, recall is about 40% below human performance. In general, the precision of the rhetorical parser comes close to the human performance level. However, since the level of granularity at which the rhetorical parser works is much coarser than that used by human judges, many sentences are assigned a much simpler structure than the structure built by humans. For example, whenever an analyst used a JOINT relation to connect two clause-like units separated by an *and*, the rhetorical parser failed to identify the two units; it often treated them as a single elementary unit. As a consequence, the recall figures at all levels were significantly lower than those specific to the humans.

5.1.2 Confusion Matrices. Another way to evaluate the performance of the rhetorical parser is to build a confusion matrix over the most frequently used relations. To enable the reader to distinguish between rhetorical and nuclearity errors, I follow the same strategy as in the case of computing labeled recall and precision figures. That is, I consider by convention that the nuclei nodes of a rhetorical representation are labeled with the relation SPAN.

Table 9 shows the distributions of rhetorical relation labels used by one of the analysts and the program. The most frequently used relations were JOINT, ELABORATION, and CONTRAST. (Label SPAN denotes the nucleus of any mononuclear relation.) Overall, the 15 most frequently used relations account for more than 92% of the relations in the corpus. The distribution of relations inferred by the program is somewhat similar, with the most frequently used relations being JOINT, ELABORATION, and CONTRAST as well. The program, though, shows a stronger preference for ELABORATION relations over JOINTS.

Table 10 shows a confusion matrix that reflects the ability of the rhetorical parser to derive rhetorical structure trees. The confusion matrix compares cumulatively, over the entire corpus, the rhetorical relations inferred by the parser with the rhetorical relations

Table 9
Distribution of the most frequently used 15 relations.

Relation	Judge (%)	Program (%)
SPAN	32.62	35.65
JOINT	14.53	14.78
ELABORATION	12.76	20.43
CONTRAST	7.80	7.82
TEXTUAL	3.54	3.47
CONDITION	3.19	1.73
EXAMPLE	2.83	3.04
SEQUENCE	2.83	–
EVIDENCE	2.12	–
OTHERWISE	2.12	0.86
PURPOSE	1.77	0.86
CONCESSION	1.77	1.73
CIRCUMSTANCE	1.77	1.30
BACKGROUND	1.77	1.73
CAUSE	1.41	2.60

Table 10
Confusion matrix.

Relation	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)
SPAN (a)	**51**	8	9												3	
JOINT (b)	4	**4**	5				1						1			
ELABORATION (c)	2	5	**18**	1												
CONTRAST (d)		1	1	**14**												1
TEXTUAL (e)					**10**											
CONDITION(f)						**5**								1		
EXAMPLE (g)		1					**6**						1			
SEQUENCE (h)	3	4	1													
EVIDENCE (i)	1	1	4													
OTHERWISE(j)										**2**						
PURPOSE (k)											**2**					
CONCESSION (l)			2									**3**				
CIRCUMSTANCE (m)	1	1												1		
BACKGROUND (n)	2	1														
CAUSE (o)	1		1												**5**	
OTHER (p)	2		1	1										1		**2**
NO SPAN (r)	14	8	5	2									1	1	1	2

chosen by one analyst. For any relation R, a column in the confusion matrix reflects the number of relations of type R that were (in)correctly identified by the rhetorical parser with respect to the relations identified by the analyst. For example, the column labeled (d) shows that out of 18 textual spans labeled with the relation CONTRAST by the parser, 14 were labeled as CONTRASTs, one as the satellite of an ELABORATION relation, and one as the satellite of an OTHER relation by the analyst. In addition, two of the 18 spans had no corresponding span in one analyst's representation.

The confusion matrix shows that the rhetorical parser does a fairly good job at recognizing rhetorical relations that are usually marked by cue phrases, i.e., CONTRAST, CONDITION, EXAMPLE, OTHERWISE, PURPOSE, CONCESSION, and CAUSE relations. The confusion matrix also shows that the simple model of cohesion that our parser employs is not adequate for distinguishing between rhetorical relations of ELABORATION and

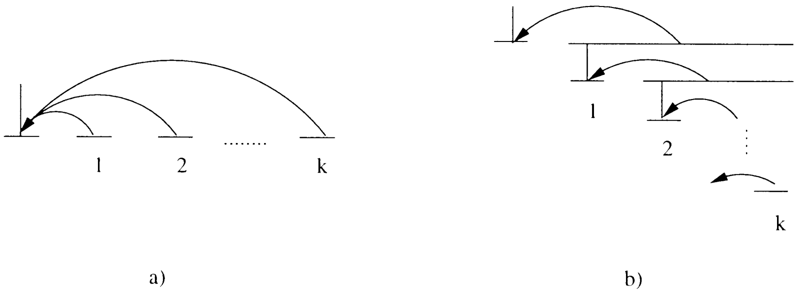

Figure 17
A flat (a) and a binary (b) representation of the discourse structure of a text.

JOINT; the submatrix that subsumes the labels SPAN, JOINT, and ELABORATION shows the highest levels of confusion. Clearly, semantic similarity is not sufficient if one is to decide whether a rhetorical relation of JOINT, ELABORATION, or BACKGROUND holds between two textual segments.

The rhetorical parser is unable to recognize "purely" intentional relations, such as EVIDENCE, which are seldomly marked. As the confusion matrix shows, the parser labels no relation as EVIDENCE; rather, it chooses instead ELABORATION, four times, JOINT, once, and a relation with a different nuclearity, once. Since the parser draws no temporal inferences, it labels most of the SEQUENCE relations as JOINTs, which in many cases is an adequate approximation.

The relatively large number of spans that have no correspondence in one analyst's representations can be explained by two factors. The first factor concerns our representation choice. Since the discourse parser builds binary trees, it also derives text spans that cannot be matched against a nonbinary representation. For example, although the tree in Figure 17(b) is a binary reformulation of the tree in Figure 17(a), the intermediate spans [1,k], [2,k], . . . , [k-1,k] in Figure 17(b) cannot be matched against any span in the tree in Figure 17(a). Humans built structures such as those shown in Figure 17(a); the discourse parser did not. The second factor concerns the difficulty of the task. Some texts have very sophisticated text structures, which are difficult to infer only on the basis of cue phrase occurrences and word-based similarity measures. We discuss some of the difficult cases in Section 5.1.3, below.

5.1.3 Qualitative Evaluation.

As the quantitative evaluation results in the preceding section show, the rhetorical structures that can be derived by relying only on discourse markers and cohesion are adequate in some cases and not in others. I discuss some of these cases below.

Good Discourse Structures at the Paragraph Level. In most of the cases that I inspected visually on a variety of texts, the partial structures built above sentences and within paragraphs appeared to be adequate. The explanation is simple: Paragraphs that use few discourse markers tend to express the most important information at their beginning, which corresponds to the first sentence being a nucleus and subsequent sentences elaborating on it. Such paragraphs usually have discourse structures that are similar to those preferred by the rhetorical parser, which favors structures that are skewed to the right and which hypothesizes that ELABORATION relations hold between unmarked sentences. If a paragraph has a more complex discourse structure, it usually employs discourse markers, which are detected and exploited by the rhetorical parser.

Good Discourse Structures at the Text Level, for Short Texts. The same argument applies for short texts as well. Texts that consist of only a few paragraphs also tend to have structures that are skewed to the right. When they do not, these texts usually rely on discourse markers to signal to the reader that the content of a paragraph should not be interpreted as a simple ELABORATION of the material that was presented before.

Good Discourse Structures for Sentences, Paragraphs, and Texts that Use Unambiguous Discourse Markers. Discourse markers such as *although, in contrast,* and *however* signal in most cases CONCESSION, CONTRAST, and ANTITHESIS relations, respectively. Since the use of these markers is consistent, most of the rhetorical relations that are signaled by such markers are correctly identified. For example, more than 75% of the CONTRAST relations that hold across clauses, sentences, and paragraphs in our corpus of *Scientific American* texts were correctly identified.

Good Discourse Structures for Sentences that Use Markers Other than And. The structures that the discourse parser derives for sentences that use discourse markers such as *because, if,* and *when* closely match those built by humans. Although the discourse parser overhypothesizes relations, the constrained mathematical model it relies upon considerably reduces the space of valid discourse interpretations. The nuclearity preferences associated with the discourse markers eliminate many of the invalid interpretations. As a consequence, the discourse structures built for sentences that have clause-like units as leaves are correct in most of these cases.

Bad Discourse Structures for Sentences that Use the Discourse Marker And. Problems from this category are readily observed in the trees in Figure 1 and Figure 14. For example, the rhetorical parser is not able to identify that a discourse boundary should be inserted before the occurrence of *and* in the sentence "[Surface temperatures typically average about -60 degrees Celsius (-76 degrees Fahrenheit) at the equator] [*and* can dip to -123 degrees C near the poles.]." As a consequence, the recall figure with respect to identifying the elementary units of this sentence is 0. The recall figure with respect to identifying the hierarchical spans of this sentence is 1/3 (the parser correctly identifies only the span that subsumes the entire sentence but not the two subspans that subsume the elementary units). The recall figures with respect to identifying the nuclearity of the spans and the rhetorical relations that hold between them are also negatively affected. Hence, it seems clear that surface-based methods are not sufficient if we are to approach human performance levels at the task of identifying elementary discourse units.

By examining the failures of the elementary unit boundary identification algorithm, I have come to believe that some of the problematic cases could be solved by using part-of-speech tags and other syntactic information. For example, in many of the sentences in which *and* is followed by a verb, an elementary unit boundary needs to be inserted before its occurrence. Such a rule would be sufficient for breaking into two units the example sentence considered above.[6] It remains to be seen whether a rhetorical parser can approach human performance levels without building full syntactic trees for the sentences under consideration.

[6] For research that uses part-of-speech tags in order to identify elementary unit boundaries, see Marcu (1999a).

Incorrectly Labeled Intentional Relations. We can see from the trees in Figure 1 and Figure 14 that although the rhetorical parser correctly identified the hierarchical segments and the nuclearity statuses in the first paragraph, it was unable to determine that a rhetorical relation of EVIDENCE holds between the last two sentences and the first sentence of the first paragraph. Instead, the parser used an ELABORATION relation. In general, the discourse parser is unable to correctly identify intentional relations, in particular, relations of EVIDENCE that hold between sentences and paragraphs. Such relations are usually not marked; to derive them one needs to "understand" what a text is about. For example, our rhetorical parser mislabeled as ELABORATION and JOINT all six EVIDENCE relations that hold between sentences and paragraphs in the texts in our *Scientific American* corpus.

It seems that to build RS-trees as accurately as humans, relying only on cue phrases and cohesion is not sufficient. In some cases, a deeper analysis of the relation between connectives and rhetorical relations, such as that proposed by Grote et al. (1997) in the context of natural language generation, may help hypothesize better relations. In general, though, it is unclear what forms of reasoning to use to derive, for unrestricted texts, relations that are as difficult to infer as the EVIDENCE relation in Figure 1.

Bad Discourse Structures for Very Large Texts. When the discourse parser attempts to derive the structure of very large texts, the preference for structures that are skewed to the right and the modeling of discourse as binary trees do not always work. For example, some newspaper articles are written so that *k* facets of the most important idea are presented in the first paragraph. And then, each of these facets is elaborated in turn in subsequent paragraphs. An adequate discourse structure for such a text is one that has the first paragraph as nucleus and *k* satellites directly linked to it at the same level of embedding (see Figure 17(a)). The choice of modeling the discourse structure of texts using binary representations appears to be infelicitous in such cases because binary trees induce an unjustified number of additional levels of embedding (see Figure 17(b)). Since the rhetorical parser derives binary trees only, it cannot represent discourse structures that would closely match the structure of newspaper articles of this kind.

The preference for discourse trees that are skewed to the right is also problematic when handling texts that start by providing some background information or by motivating the reader before presenting the main idea. For example, the text in italics in (39) should be the satellite of a MOTIVATION relation whose nucleus subsumes the rest of the text.

(39) *Running nose. Raging fever. Aching joints. Splitting headache. Are there any poor souls suffering from the flu this winter who haven't longed for a pill to make it all go away? Relief may be in sight.* Researchers at Gilead Sciences, a pharmaceutical company in Foster City, California, reported last week in the Journal of the American Chemical Society that they have discovered a compound that can stop the influenza virus from spreading in animals. Tests on humans are set for later this year.

Unfortunately, cohesion is not enough for determining this relation. Consequently, the discourse structure built by the rhetorical parser for this text is erroneous.

5.2 Evaluating the Usefulness of the Trees for Text Summarization

From a salience perspective, the elementary units in the promotion set of a node of a tree structure denote the most important units of the textual span that is dominated by that node. For example, according to the rhetorical structure in Figure 14, unit 3 is the most important unit of span [3,6], units 4 and 5 are the most important units of span [4,6], and unit 2 is the most important unit of the whole text. If we apply the concept of salience over all elementary units in a text, we can use the rhetorical structure to induce a partial ordering on the importance of these units. The intuition behind this approach is that the textual units in the promotion sets of the top nodes of a discourse tree are more important than the units that are salient in the nodes found at the bottom. When applied to the rhetorical structure in Figure 14, such an approach induces the partial ordering in (40), because unit 2 is the only promotion unit of the root; unit 8 is the only unit found in the promotion set of a node immediately below the root (unit 2 has been already accounted for); units 3 and 10 are the only units that belong to promotion sets of nodes that are two levels below the root; and so on. (See Marcu [1999b] for a mathematical formulation of this method that uses rhetorical structures for deriving a partial ordering of the important units in texts.)

(40) $$2 > 8 > 3, 10 > 1, 4, 5, 7, 9 > 6$$

If we are interested in generating a very short summary of text 19, for example, we can then produce an extract containing only unit 2, because this is the most important unit given by the partial ordering derived from the corresponding rhetorical representation. A longer summary will contain units 2 and 8; a longer one, units 2, 8, 3, and 10; and so on.

Using this idea, I have implemented a rhetorical-based summarization algorithm. The algorithm uses the rhetorical parser described in this paper to determine the discourse structure of a text given as input, it uses the discourse structure to induce a partial ordering on the elementary units in the text, and then, depending on the desired compression rate, it selects the p most important units in the text.

To evaluate this summarization program, I used two corpora: the five *Scientific American* texts that I have mentioned above, and a collection of 40 short newspaper articles from the TREC collection (Jing et al. 1998). Both corpora were labeled for textual salience by a panel of independent judges: 13 judges labeled clause-like units as being important, somewhat important, and nonimportant in the texts of the *Scientific American* corpus; and 5 judges labeled sentences as worthy to be included in 10% and 20% summaries of the texts in the TREC corpus. The clauses/sentences which the human judges agreed were important were taken as the gold standard for summarization.

The rhetorical parser derived the RS-tree of each of the 45 texts in the two corpora, and used the RS-tree to induce a partial ordering of the importance of the elementary units in the corresponding text. The rhetorical summarizer then selected the most important k units in a text, where k was chosen so as to match as closely as possible the number of units in the gold standard. The number of units selected for summarization was determined similarly for the other summarization programs that I used in the evaluation.

To assess the performance of the rhetorical-based summarizer (and of the other summarizers that I discuss below), I use recall, precision, and F-value figures. The recall figure is given by the number of units that were correctly identified by the summarizer as being important, over the total number of important units in the gold standard. The precision figure is given by the number of units that were correctly identified by the summarizer as being important, over the total number of units identified by the summarizer. The F-value is a combined Recall-Precision value, given by the formula $2 \times \mathit{Recall} \times \mathit{Precision}/(\mathit{Recall} + \mathit{Precision})$.

Table 11
The performance of the rhetorical-based summarizer.

Corpus	Method	Recall	Precision	F-value
Scientific American	Judges	72.66	69.63	71.11
(Clause-level	Rhetorical-based summarizer with learning	67.57	73.53	70.42
summarization)	Rhetorical-based summarizer	51.35	63.33	56.71
	Microsoft Office97 summarizer	27.77	25.44	26.55
	Lead baseline	39.68	39.68	39.68
	Random baseline	25.70	25.70	25.70
TREC	Judges	82.83	64.93	72.80
Sentence-level	Rhetorical-based summarizer with learning	61.79	60.83	61.31
summarization	Rhetorical-based summarizer	46.54	49.73	48.08
(20% compression	Microsoft Office97 summarizer	39.00	32.00	35.15
rate)	Lead baseline	70.91	46.96	56.50
	Random baseline	15.80	15.80	15.80

In order to compare the performance of the rhetorical-based summarizer with that of humans, I have also determined the performance of the human judges, by averaging the performance of each judge with respect to the gold standard. As Table 11 shows, the human-level F-value for the task of identifying important clauses in the *Scientific American* corpus was 71.11%; the human-level F-value for the task of identifying the most important 20% of the sentences in the TREC texts was 72.80%. To better assess the performance of the rhetorical-based summarizer, I also determined the performance of two baseline summarizers. The lead-based summarizer assumes that the most important k units in a text are the first k units in that text. The random-based summarizer assumes that the most important k units in a text can be selected stochastically.

As Table 11 shows, for both corpora, the rhetorical-based summarizer performs better than the random baseline summarizer and better than a commercial system, the Microsoft Office97 summarizer. The rhetorical-based summarizer outperforms the lead-based summarizer only for texts in the *Scientific American* corpus. Most of the newspaper articles in the TREC collection employ the pyramid journalistic style and have the most important sentences at the beginning of the articles. As a consequence, the performance of the lead-based summarizer on TREC texts is quite high. However, an implementation of the rhetorical parser that uses learning techniques to choose rhetorical interpretations that are likely to increase the performance of the rhetorical-based summarizer yields a program that identifies important units at levels of performance that are close to human performance for *Scientific American* texts and that are about 10% below human performance for TREC newspaper articles and about 5% above the lead baseline. The rhetorical-based summarizer that employs learning techniques to improve its performance is discussed in detail in Marcu (2000).

The data in Table 11 shows that although the rhetorical parser does not produce perfect rhetorical structure trees, it can be used successfully to determine the important units of texts.

6. Related Work

When this research was carried out, there was no rhetorical parser for English. However, very recently, Corston-Oliver (1998) has explored a different facet of the work described here and investigated the possibility of using syntactic information to hypothesize relations. His system uses 13 rhetorical relations and builds discourse trees for articles in Microsoft's *Encarta 96 Encyclopedia*. I believe that the research that comes closest to that described in this chapter is that of Sumita et al. (1992) and Kurohashi and Nagao (1994).

Sumita et al. (1992) report on a discourse analyzer for Japanese, which differs from mine in a number of ways. Particularly important is the fact that the theoretical foundations of Sumita et al.'s analyzer do not seem to be able to accommodate the ambiguity of discourse markers; in their system, discourse markers are considered unambiguous with respect to the relations that they signal. In contrast, my rhetorical parser uses a mathematical model in which this ambiguity is acknowledged and appropriately treated. Furthermore, the discourse trees that the rhetorical parser builds are more constrained structures (Marcu 2000): as a consequence, the rhetorical parser does not overgenerate invalid trees as Sumita et al.'s does. Finally, my rhetorical parser uses only surface-form methods for determining the markers and textual units and uses clause-like units as the minimal units of the discourse trees. In contrast, Sumita et al. use deep syntactic and semantic processing techniques for determining the markers and the textual units and use sentences as minimal units in the discourse structures that they build.

Kurohashi and Nagao (1994) describe a discourse structure generator that builds discourse trees in an incremental fashion. The algorithm proposed by Kurohashi and Nagao starts with an empty discourse tree and then incrementally attaches sentences to its right frontier, in the style of Polanyi (1988). The node of attachment is determined on the basis of a ranking score that is computed using three different sources: cue phrases, chains of identical and similar words, and similarities in the syntactic structure of sentences. As in the case of Sumita's system, Kurohashi and Nagao's system takes as input a sequence of parse trees; hence, in order to work, it must be preceded by a full syntactic analysis of the text. The elementary units of the discourse trees built by Kurohashi and Nagao are sentences.

Since the systems developed by Corston-Oliver (1998), Sumita et al. (1992), and Kurohashi and Nagao (1994) were not evaluated intrinsically, it is difficult to compare the performance of their systems to ours.

A parallel line of research has been investigated recently by Strube and Hahn (1999). They have extended the centering model proposed by Grosz, Joshi, and Weinstein (1995) by devising algorithms that build hierarchies of referential discourse segments. These hierarchies induce a discourse structure on text, which constrains the reachability of potential anaphoric antecedents. The referential segments are constructed through an incremental process that compares the centers of each sentence with those of the structure that has been built up to that point.

The referential structures that are built by Hahn and Strube exploit a language facet different from that exploited by the rhetorical parser: their algorithms rely primarily on cohesion and not on coherence. Because of this, the referential structures are not as constrained as the discourse structures that the rhetorical parser builds. In fact, the discourse relations between the referential segments are not even labeled. Still, I believe that studying the commonalities and differences between the referential and rhetorical segments could provide new insights into the nature of discourse.

7. Discussion and Conclusion

Automatically deriving the discourse structure of texts is not trivial. This paper discusses extensively the strengths and weaknesses of an approach to discourse parsing that relies on cue phrases, cohesion, and a formal model of discourse. Quantitative and qualitative analyses of the results show that many relations can be identified correctly within this framework. However, this approach is not sufficient for identifying intentional relations, such as EVIDENCE, or for choosing between ELABORATION, BACKGROUND, SEQUENCE, and JOINT relations.

The brightest side of the story is that the results in this paper show that the rhetorical structures derived by my parser can be used successfully in the context of text summarization. Hence, although the rhetorical parser does not get the RS-trees perfectly right, it still manages to determine the important units of text at levels of performance that are not far from those of humans. One possible explanation may be that the rhetorical-based summarizer described here exploits only the difference between satellites and nuclei and the hierarchical structure of text to determine text units that are important. The reader should not infer from this that correctly identifying the rhetorical relations that hold between spans cannot be useful in a summarization setting. It is likely, for instance, that one may want to systematically exclude from an abstract information that is subsumed by the satellite of an EXAMPLE relation; to do so, it is necessary to identify correctly the relation.

The rhetorical summarizer is a niche application that shows how an understanding of the hierarchical organization of text can make solving difficult natural language problems easier. Recent research has shown that by exploiting the structure of discourse, one can decrease storage space in information retrieval applications (Corston-Oliver and Dolan 1999) and address discourse-specific problems in machine translation (Marcu, Carlson, and Watanabe, 2000). It is possible that discourse structures of the kinds derived by this parser can have a positive impact on other problems as well. For example, Cristea et al. (1999) have shown that a hierarchical model of discourse has a higher potential for improving the performance of a coreference resolution system than a linear model of discourse. And Hirschman et al. (1999) have suggested that certain types of questions can be better answered if one has access to rhetorical structure representations of the texts that contain the answers to the questions. How much of an impact the rhetorical parser presented here can have on solving these problems, of course, remains an empirical question.

Acknowledgments

I am grateful to Graeme Hirst for the help and advice he gave me during every stage of this work; and to Marilyn Walker for suggestions that led to significant improvements, especially in the Evaluation section of the paper. I am also grateful to four anonymous reviewers who provided thoughtful feedback on an earlier draft. Most of this research was conducted while I was at the University of Toronto and was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Asher, Nicholas. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Dordrecht.

Asher, Nicholas and Alex Lascarides. 1994. Intentions and information in discourse. In *Proceedings of the 32nd Annual Meeting*, pages 34–41, New Mexico State University, Las Cruces, NM, June. Association for Computational Linguistics.

Bestgen, Yves and Jean Costermans. 1997. Temporal markers of narrative structure: Studies in production. In Jean Costermans and Michel Fayol, editors, *Processing Interclausal Relationships. Studies in the Production and Comprehension of Text*. Lawrence Erlbaum Associates, Hillsdale, NJ, pages 201–218.

Briscoe, Ted. 1996. The syntax and semantics of punctuation and its use in interpretation. In *Proceedings of the Association for Computational Linguistics Workshop on Punctuation*, pages 1–7, Santa Cruz, CA, June.

Bruder, Gail A. and Janice M. Wiebe. 1990. Psychological test of an algorithm for recognizing subjectivity in narrative text. In *Proceedings of the Twelfth Annual Conference on the Cognitive Science Society*, pages 947–953, Cambridge, MA, July.

Corston-Oliver, Simon H. 1998. Beyond string matching and cue phrases: Improving efficiency and coverage in discourse analysis. In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, pages 9–15, Stanford, March.

Corston-Oliver, Simon H. and William B. Dolan. 1999. Less is more: Eliminating index terms from subordinate clauses. In *Proceedings of the 37th Annual Meeting*, pages 349–356, University of Maryland, June. Association for Computational Linguistics.

Cristea, Dan, Nancy Ide, Daniel Marcu, and Valentin Tablan. 1999. Discourse structure and coreference: An empirical study. In *Proceedings of the ACL'99 Workshop on the Relationship Between Discourse/Dialogue Structure and Reference*, pages 46–53, University of Maryland, June.

Cristea, Dan and Bonnie L. Webber. 1997. Expectations in incremental discourse processing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL–97)*, pages 88–95, Madrid, Spain, July.

Crystal, David. 1991. *A Dictionary of Linguistics and Phonetics*. Third edition. Basil Blackwell, Oxford.

Cumming, Carmen and Catherine McKercher. 1994. *The Canadian Reporter: News Writing and Reporting*. Harcourt Brace.

Di Eugenio, Barbara, Johanna D. Moore, and Massimo Paolucci. 1997. Learning features that predict cue usage. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL–97)*, pages 80–87, Madrid, Spain, July.

Fraser, Bruce. 1996. Pragmatic markers. *Pragmatics*, 6(2):167–190.

Gardent, Claire. 1997. Discourse TAG. Technical Report CLAUS-Report Nr. 89, Universität des Saarlandes, Saarbrücken, April.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grote, Brigitte, Nils Lenke, and Manfred Stede. 1997. Ma(r)king concessions in English and German. *Discourse Processes*, 24:87–117.

Halliday, Michael A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Harabagiu, Sanda and Steven Maiorano. 1999. Knowledge-lean coreference resolution and its relation to textual cohesion and coreference. In *Proceedings of the ACL'99 Workshop on Discourse/Dialogue Structure and Reference*, pages 29–38, University of Maryland, June.

Harabagiu, Sanda M. and Dan I. Moldovan. 1996. Textnet—A text-based intelligent system. In *Working Notes of the AAAI Fall Symposium on Knowledge Representation Systems Based on Natural Language*, pages 32–43, Cambridge, MA.

Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Heurley, Laurent. 1997. Processing units in written texts: Paragraphs or information blocks. In Jean Costermans and Michel Fayol, editors, *Processing Interclausal Relationships. Studies in the Production and Comprehension of Text*. Lawrence Erlbaum Associates, pages 179–200.

Hirschberg, Julia and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Hirschman, Lynette, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting*, pages 325–332, University of Maryland, June. Association for Computational Linguistics.

Hobbs, Jerry R. 1990. *Literature and Cognition*. CSLI Lecture Notes Number 21.

Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

Hoey, Michael. 1991. *Patterns of Lexis in Text*. Oxford University Press.

Jing, Hongyan, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *Proceedings of the AAAI–98 Spring Symposium on Intelligent Text Summarization*, pages 60–68, Stanford, March.

Kameyama, Megumi. 1994. Indefeasible semantics and defeasible pragmatics. Technical Note 544, SRI International. A shorter version to appear in Kanazawa Makoto, Christopher Pinon, and Henriette de Swart, editors, *Quantifiers, Deduction, and Context*. CSLI, Stanford.

Kamp, Hans. 1981. A theory of truth and semantic interpretation. In J. A. G. Groenendijk, T. M. V. Janssen, and M. B. J. Stokhof, editors, *Formal Methods in the*

Study of Language, Mathematical Centre Tracts 135. Mathematisch Centrum, Amsterdam, pages 277–322.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to ModelTheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, London, Boston, Dordrecht. Studies in Linguistics and Philosophy, Volume 42.

Kintsch, Walter. 1977. On comprehending stories. In Marcel Just and Patricia Carpenter, editors, *Cognitive Processes in Comprehension*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Knott, Alistair. 1995. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh.

Kurohashi, Sadao and Makoto Nagao. 1994. Automatic detection of discourse structure by checking surface information in sentences. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING–94)*, volume 2, pages 1,123–1,127, Kyoto, Japan, August.

Lascarides, Alex and Nicholas Asher. 1993. Temporal interpretation, discourse relations, and common sense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Litman, Diane J. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5:53–94.

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. 1996. Building up rhetorical structure trees. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI–96)*, volume 2, pages 1,069–1,074, Portland, OR, August.

Marcu, Daniel. 1997a. The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL–97)*, pages 96–103, Madrid, Spain, July.

Marcu, Daniel. 1997b. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. thesis, Department of Computer Science, University of Toronto, December.

Marcu, Daniel. 1999a. A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting*, pages 365–372, University of Maryland, June. Association for Computational Linguistics.

Marcu, Daniel. 1999b. Discourse trees are good indicators of importance in text. In Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, pages 123–136.

Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA. To appear.

Marcu, Daniel, Lynn Carlson, and Maki Watanabe. 2000. The automatic translation of discourse structures. In *Proceedings of the Language Technology Joint Conference ANLP-NAACL2000*, Seattle, WA.

Martin, James R. 1992. *English Text. System and Structure*. John Benjamin Publishing Company, Philadelphia, Amsterdam.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Moser, Megan and Johanna D. Moore. 1997. On the correlation of cues with discourse structure: Results from a corpus study. Forthcoming.

Nunberg, G. 1990. *The Linguistics of Punctuation*. CSLI Lecture Notes 18, Stanford. University of Chicago Press.

Palmer, David D. and Marti A. Hearst. 1997. Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics*, 23(2):241–269.

Pascual, Elsa and Jacques Virbel. 1996. Semantic and layout properties of text punctuation. In *Proceedings of the Association for Computational Linguistics Workshop on Punctuation*, pages 41–48, Santa Cruz, CA, June.

Polanyi, Livia. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638.

Polanyi, Livia. 1996. The linguistic structure of discourse. Technical Report CSLI–96–200, Center for the Study of Language and Information.

Polanyi, Livia and Martin H. van den Berg. 1996. Discourse structure and discourse interpretation. In P. Dekker and M. Stokhof, editors, *Proceedings of the Tenth Amsterdam Colloquium*, pages 113–131. Department of Philosophy, University of Amsterdam.

Redeker, Gisela. 1990. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics*, 14:367–381.

Salton, Gerard and James Allan. 1995. Selective text utilization and text traversal. *International Journal of Human-Computer Studies*, 43:483–497.

Salton, Gerard, Amit Singhal, Chris Buckley, and Mandar Mitra. 1995. Automatic text decomposition using text segments and text themes. Technical Report TR-95-1555, Department of Computer Science, Cornell University.

Say, Bilge and Varol Akman. 1996. Information-based aspects of punctuation. In *Proceedings of the Association for Computational Linguistics Workshop on Punctuation*, pages 49–56, Santa Cruz, CA, June.

Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge University Press.

Schilder, Frank. 1997. Tree discourse grammar, or how to get attached a discourse. In *Proceedings of the Second International Workshop on Computational Semantics (IWCS-II)*, pages 261–273, Tilburg, The Netherlands, January.

Schneuwly, Bernard. 1997. Textual organizers and text types: Ontogenetic aspects in writing. In Jean Costermans and Michel Fayol, editors, *Processing Interclausal Relationships. Studies in the Production and Comprehension of Text*. Lawrence Erlbaum Associates, Hillsdale, NJ, pages 245–263.

Segal, Erwin M. and Judith F. Duchan. 1997. Interclausal connectives as indicators of structuring in narrative. In Jean Costermans and Michel Fayol, editors, *Processing Interclausal Relationships. Studies in the Production and Comprehension of Text*. Lawrence Erlbaum Associates, Hillsdale, NJ, pages 95–119.

Segal, Erwin M., Judith F. Duchan, and Paula J. Scott. 1991. The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse Processes*, 14:27–54.

Shiuan, Peh Li and Christopher Ting Hian Ann. 1996. A divide-and-conquer strategy for parsing. In *Proceedings of the Association for Computational Linguistics Workshop on Punctuation*, pages 57–66, Santa Cruz, CA, June.

Siegel, Eric V. and Kathleen R. McKeown. 1994. Emergent linguistic rules from inducing decision trees: Disambiguating discourse clue words. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI–94)*, volume 1, pages 820–826, Seattle, WA.

Strube, Michael and Udo Hahn. 1999. Functional centering—Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

Sumita, K., K. Ono, T. Chino, T. Ukita, and S. Amano. 1992. A discourse structure analyzer for Japanese text. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, volume 2, pages 1,133–1,140.

van den Berg, Martin H. 1996. Discourse grammar and dynamic logic. In P. Dekker and M. Stokhof, editors, *Proceedings of the Tenth Amsterdam Colloquium*, pages 93–112. Department of Philosophy, University of Amsterdam.

van Dijk, Teun A. 1972. *Some Aspects of Text Grammars; A Study in Theoretical Linguistics and Poetics*. Mouton, The Hague.

Webber, Bonnie, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse relations: A structural and presuppositional account using lexicalized TAG. In *Proceedings of the 37th Annual Meeting*, pages 41–48, University of Maryland, June. Association for Computational Linguistics.

Wiebe, Janice M. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–288.

Youmans, Gilbert. 1991. A new tool for discourse analysis: The vocabulary-management profile. *Language*, 67(4):763–789.

Zock, Michael. 1985. Le fil d'ariane ou les grammaires de texte comme guide dans l'organisation et l'expression de la pensée en langue maternelle et/ou étrangère. Technical Report, Rapport pour l'Unesco, Juin.