THE THEORY AND PRACTICE OF DISCOURSE PARSING AND SUMMARIZATION

DANIEL MARCU

THE THEORY AND PRACTICE OF DISCOURSE PARSING AND SUMMARIZATION

DANIEL MARCU

Until now, most discourse researchers have assumed that full semantic understanding is necessary to derive the discourse structure of texts. This book documents the first serious attempt to construct automatically and use nonsemantic computational structures for text summarization. Daniel Marcu develops a semantics-free theoretical framework that is both general enough to be applicable to naturally occurring texts and concise enough to facilitate an algorithmic approach to discourse analysis. He presents and evaluates two discourse parsing methods: one uses manually written rules that reflect common patterns of usage of cue phrases such as "however" and "in addition to"; the other uses rules that are learned automatically from a corpus of discourse structures. By means of a psycholinguistic experiment, Marcu demonstrates how a discourse-based summarizer identifies the most important parts of texts at levels of performance that are close to those of humans.

Marcu also discusses how the automatic derivation of discourse structures may be used to improve the performance of current natural language generation, machine translation,

The Theory and Practice of Discourse Parsing and Summarization

The Theory and Practice of Discourse Parsing and Summarization

Daniel Marcu

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times Roman by Windfall Software using ZzT$_E$X and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Marcu, Daniel.
The theory and practice of discourse parsing and summarization / Daniel Marcu.
p. cm.
"A Bradford book."
Includes bibliographical references and index.
ISBN 0-262-13372-5 (hc.: alk. paper)
I. Discourse analysis—Data processing. 2. Parsing (Computer grammar) 3. Abstracts—Data processing. I. Title.

P302.3 .M37 2000
401.410285—dc21 00-038690

To my parents and to my friends who have helped and influenced me the most, Mica, Tatatel, Marin, Cornel, Vasile, Cuțu, More, Adi, Rareș, Vivi, Călin, Doina, Bilă, Ion, Juvete, Horace, Pelicanii, Țicrea, Grir, Cașu, Almi, Băsă, E6, Reli, Ciupe, Brîndu, Oana, Monica, Cipi, Ed, Gelu, Melanie, Jin, Bil, Laura, Alex, Attila, Suflețel.

Contents

Figures

Tables

Preface

Many researchers of discourse agree that coherent texts have internal structure and that this structure is conveniently characterized by discourse/rhetorical relations, i.e., relations that reflect semantic and functional judgments about the text spans they connect. Yet, despite significant progress in understanding the linguistic phenomena above the sentence boundary, the discourse parsing of free, unrestricted text remains an elusive goal. To date, most researchers have assumed that in order to derive the discourse structure of texts, one needs full semantics. In this book, I explore an alternative approach to discourse processing that need not be grounded in a full semantic account of sentence processing.

Instead of focusing on the semantics of discourse relations and on the relationship between the semantics of discourse and that of the individual sentences and clauses, I provide a completely specified axiomatization of the most widely accepted mathematical properties of discourse structures, which are amenable to straightforward formalization. The axiomatization is strong enough to reduce significantly the space of discourse interpretations. Also, it is strong enough to enable one to derive well-formed discourse structures for unrestricted texts with surprisingly good results, although the rhetorical relations that hold between textual units and spans cannot themselves be determined unambiguously.

The reason one can derive the discourse structure of texts despite their inherent rhetorical ambiguity may be found in the fact that the axiomatization proposed here enables an explicit enumeration of all valid interpretations. In the same way a syntactic theory enables all valid syntactic interpretations of a sentence to be derived, the axiomatization proposed in this book enables all valid discourse interpretations of a text to be derived. But in the same way a syntactic theory may produce interpretations that are incorrect from a semantic perspective, this axiomatization may produce interpretations that are incorrect when additional discourse-specific phenomena, such as focus, cohesion, and intentions, are factored in.

Since the formalism and algorithms described in this book can be applied to any text, the strengths and weaknesses of the approach and the generality of the principles it is based on can be immediately and properly evaluated. The evaluations carried out are both intrinsic and extrinsic:

- For the intrinsic evaluation, I assess how closely the discourse structures derived automatically for a set of texts matched the discourse structures that were constructed by humans.
- For the extrinsic evaluation, I estimate the utility of automatically derived discourse structures to produce summaries of texts. To this end, I first show by means of a psycholinguistic experiment that discourse structures can be used effectively in order to determine which portions of texts humans perceive as being important. I then use the lessons learned from the experiment in order to implement a discourse-based summarization algorithm, which

identifies important clauses and sentences in text at levels of performance that exceed those of current commercial systems and are close to those of humans.

Automatically deriving the discourse structure of text is a difficult problem. This book does not solve it. Importantly, though, the book shows how one can estimate quantitatively the validity of the theoretical assumptions that it relies upon and the success of the discourse parsing and discourse-based summarization algorithms that it proposes. That is, the book allows one to make not only qualitative statements, such as "discourse processing is hard", but also quantitative ones, such as the following:

- "By using cue phrases and cohesion, I can implement a discourse parser that is 30% below human performance."
- "If I determine the elementary units of discourse correctly and use knowledge about cue phrases, cohesion, part of speech tags, and Wordnet lexical relations, I can build discourse structures whose hierarchical scaffold is as good as the scaffold of the structures built by humans."
- "If I use machine learning techniques, I can train a discourse-based summarizer to identify important units in short scientific articles as well as humans do. Using the same techniques, I can train a discourse-based summarizer to identify important units in newspaper articles at levels of performance that are 10% below the level of humans."

Such quantitative estimates of the effects of the hypotheses and choices one makes in developing theories and algorithms are crucial for furthering progress in the field.

Being able to derive automatically the structure of text can have a significant impact on solving a variety of problems in syntactic processing, natural language generation, machine translation, summarization, question answering, and information retrieval. Some of these problems may be addressed using only the theory and algorithms presented in this book. Some of them may need more elaborate theories and algorithms. I hope this book will provide a starting point to those who want to address these problems and inspire those who believe that automatic discourse processing is feasible.

Acknowledgments

This book is based on research that I carried out as a Ph.D. student at the University of Toronto and on subsequent work that I conducted at the Information Sciences Institute, University of Southern California. Many people have contributed directly and indirectly to its current form. I am grateful to all of them.

Most of all, I am grateful to Graeme Hirst who taught me to write, love language, and see in it more than a string of characters that is subject to immediate formalization and processing. I am grateful to Hector Levesque and Raymond Reiter who taught me logics and how to lean on the chair and ask: "what's the scientific problem that you solve?" To Eduard Hovy for creating at ISI a supportive environment in which research ideas can flourish and for reading patiently several versions of this book and providing invaluable feedback.

Besides them, many other people contributed comments and suggestions on earlier drafts of this book or on related articles. I would like to thank all of them, in particular, Estibaliz Amorrortu, Melanie Baljko, Alexander Budanitsky, Mark Chignell, Derek Corneil, Michael Cummings, Chrysanne DiMarco, Toby Donaldson, Phil Edmonds, Ulrich Germann, Steve Green, Sanda Harabagiu, Ulf Hermjakob, Kevin Knight, Chin-Yew Lin, Marzena Makuta, Inderjeet Mani, Marilyn Mantei, Kathleen McKeown, David Mitchell, Ion Muşlea, Magdalena Romera, Holger Schauer, Kevin Schlueter, Jeff Siskind, Ron Smyth, Manfred Stede, and Marilyn Walker. I am also grateful to the anonymous reviewers who commented on articles related to this book and to Regina Barzilay, Michael Elhadad, Hongyan Jing, and Kathleen McKeown for sharing their TREC summarization data.

1 Introduction

1.1 Motivation

Researchers in linguistics and computational linguistics have long pointed out that text is not just a simple sequence of clauses and sentences, but rather follows a highly elaborate structure. Still, a formal theory of free, unrestricted text, one that can be easily implemented in computational systems, is yet to be developed. In fact, the lack of such a theory is reflected by current natural language systems: most of them process text on a sentence-by-sentence basis. For example, if they were given the sequences of words shown in Figures 1.1 and 1.2, which differ only in the order of the sentences, they would, most likely, derive in both cases syntactic trees and construct semantic representations for each of the individual sentences, or fill up template slots for the entire texts without noticing any anomalies. Yet, only the sequence shown in Figure 1.1 is coherent, i.e., is understandable text. The sequence shown in Figure 1.2 does not make too much sense; consider just its first sentence: it does not seem right to start a text with an explicitly marked example.

The fact that the sequence in Figure 1.1 is coherent text, while the sequence in Figure 1.2 is merely a collection of sentences, although each is exemplary when taken in isolation, suggests that extra-sentential factors play a significant role in text understanding. If we are to build proficient natural language systems, it seems, therefore, obvious that we also need to enable these systems to derive inferences that pertain not only to the intra-sentential level, but to the extra-sentential level as well.

The inferences that I have in mind here are primarily of a rhetorical and intentional nature. Such inferences would enable a system to understand how the information given in different sentences and clauses is related, where the textual segments are, what the arguments that support a certain claim are, what the important clauses and sentences in a text are, etc. For example, with respect to the text in Figure 1.1, such inferences will explain that the statement Surface temperatures typically average about −60 degrees Celsius (−76 degrees Fahrenheit) at the equator and can dip to −123 degrees C near the poles provides EVIDENCE for the fact that Mars experiences frigid weather conditions; that 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 is an EXAMPLE of the statement most Martian weather involves blowing dust or carbon dioxide; that it is the low atmospheric pressure that CAUSES the liquid water to evaporate; and that 50 percent farther from the sun than Earth is some parenthetical information that is not central to the understanding of the sentence it belongs to.

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.

Figure 1.1
Example of coherent text (*Scientific American, November 1996*)

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. With its distant orbit—50 percent farther from the sun than Earth—and slim atmospheric blanket, Mars experiences frigid weather conditions. 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.

Yet even on the summer pole, where the sun remains in the sky all day long, temperatures never warm enough to melt frozen water. Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian weather involves blowing dust or carbon dioxide. Surface temperatures typically average about −60 degrees Celsius (−76 degrees Fahrenheit) at the equator and can dip to −123 degrees C near the poles.

Figure 1.2
Example of incoherent text

One possible way to represent these inferences explicitly is by means of a tree-like structure such as that shown in Figure 1.3.[1] In Figure 1.3, each leaf of the tree is associated with a contiguous textual span. The parenthetical units are enclosed within curly brackets. 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

1. This representation employs the conventions proposed by Mann and Thompson [1988] (see Section 2.2 for a detailed discussion).

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.

For example, the textual unit most Martian weather involves blowing dust or carbon dioxide is at the end of an arc that originates from textual unit Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, 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 spans that subsume the texts 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.

During the continuous refinement of the text and discourse theories that have been proposed to date, it has become clear that an adequate linguistic, formal, and computational account of text structures would have to provide answers to questions such as these:

- What is the abstract structure of text? Does it resemble the tree-structure shown in Figure 1.3? If so, what are the constraints that characterize this structure?
- What are the elementary units of texts?
- What are the relations that could hold between two textual units and what is the nature of these relations? Is the semantics of these relations grounded into the events and the world that the text describes? Or is it grounded into general principles of rhetoric, argumentation, and linguistics? Or both?
- Is there any correlation between these relations and the concrete lexicogrammatical realization of texts?
- How can text structures be determined automatically?
- Is there any correlation between the structure of text and what readers perceive as being important?

This book tries to answer these questions. In its first part, it examines the linguistic properties of naturally occurring texts with respect to their high-level discourse structures, and studies the formal properties of such structures. In the second part, it shows how the theoretical framework developed in Part I can be used in order to produce rhetorical parsing algorithms that derive the discourse structures of free texts. In the last part, it shows how discourse parsers can be employed in the context of building high-performance summarization systems.

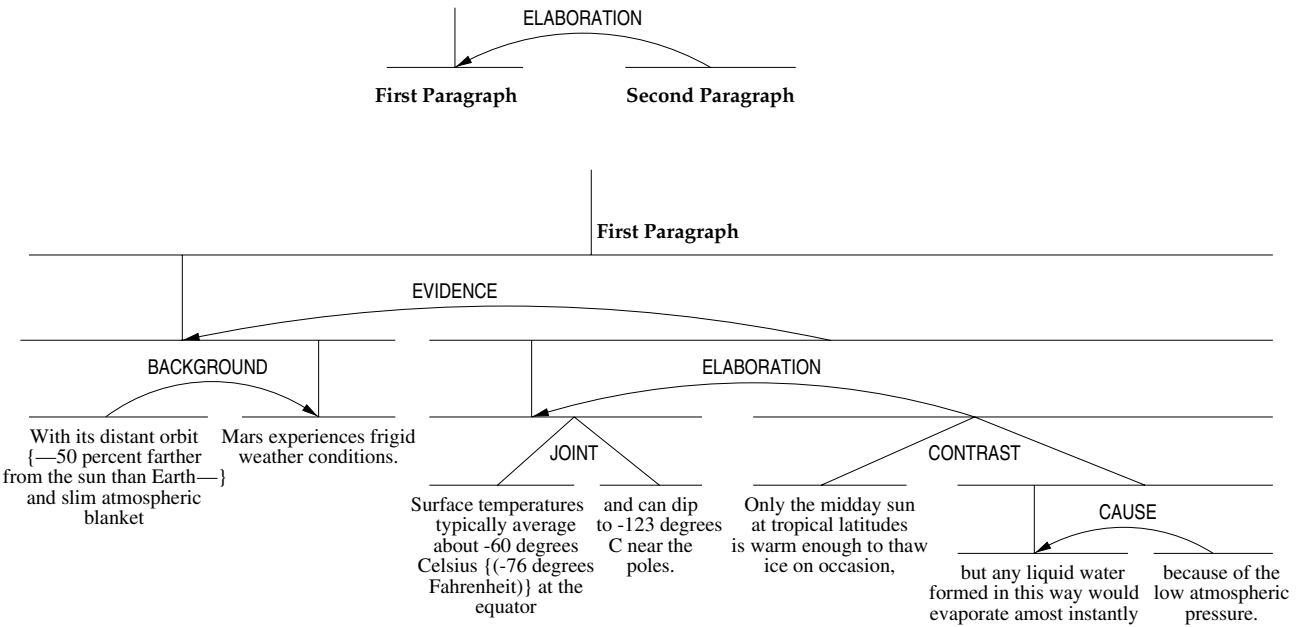

Figure 1.3
A tree-like structure that shows the rhetorical relations between the textual units of the text in Figure 1.1

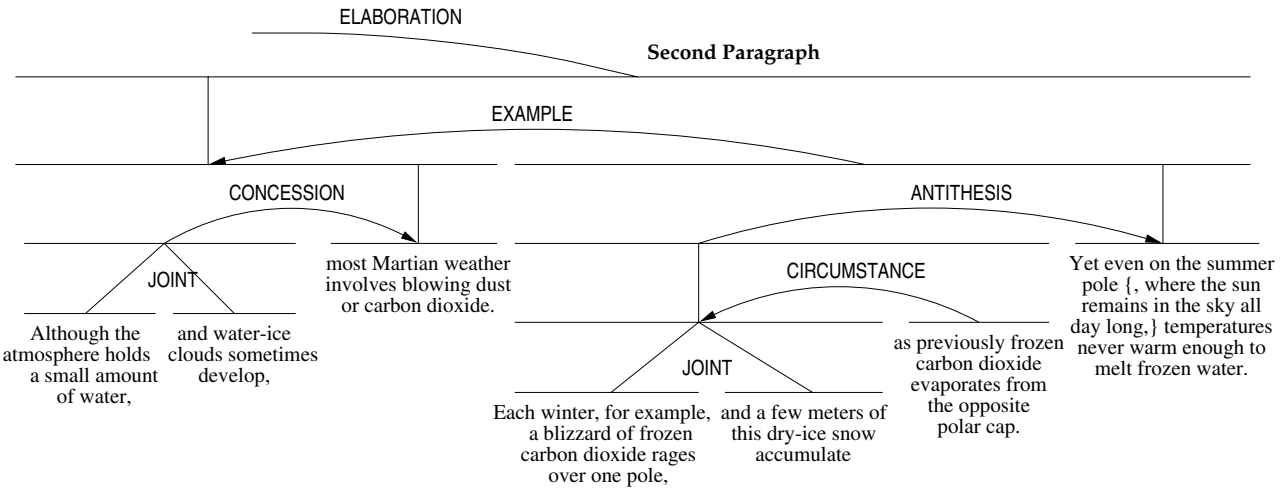

Figure 1.3 *(continued)*

1.2 Overview of the Book

1.2.1 Theoretical Foundations

In the first part of the book, I lay out some theoretical foundations: I propose a first-order formalization of the mathematical properties of valid text structures and I present, at the theoretical level, a set of non-incremental rhetorical parsing algorithms that can be used to derive *some* or *all* valid discourse structures of a text.

In formalizing the structure of unrestricted texts, I first distill the linguistic features that are common to previous approaches and show that most discourse theories assume that text can be sequenced into elementary units; that discourse relations of various natures hold between textual units of various sizes; that some textual units are more essential to the writer's purpose than others; and that trees are a good approximation of the abstract structure of text. Then I show that in order to create algorithms that produce discourse structures of texts we need to worry not only about providing unambiguous definitions of rhetorical relation, but also about explaining the relationship between discourse relations that hold between large spans and discourse relations that hold between elementary discourse units. Explaining this relationship amounts to proposing a compositionality criterion for discourse structures.

In Section 3.1, Chapter 3, I formalize this compositionality criterion and the features listed above in the language of first-order logic. The resulting formalization is independent with respect to the set of rhetorical relations that it can rely upon. As an example, I show how one can obtain, as a by-product, a formalization of the structural properties that are specific to Rhetorical Structure Theory (RST) [Mann and Thompson, 1988].

The formalization proposed in Section 3.1 focuses only on the mathematical properties of text structures, but says nothing about how discourse structures can be derived. I focus on the following theoretical problem of text structure derivation: given a sequence of elementary textual units, i.e., a text, and a set of rhetorical relations that hold among these units and spans of units, find all valid text structures that characterize the sequence. In Section 3.2, I provide a proof theory for this problem and I prove that it is sound and complete with respect to the formalization in Section 3.1.

In Chapter 4, I present three algorithms that can be used to derive some or all valid structures of a text. Two of them employ model-theoretic techniques and rely upon the formalization given in Section 3.1. The other one employs proof-theoretic techniques; it implements the proof theory developed in Section 3.2. The performance of the algorithms is compared empirically on a benchmark of discourse problems.

Like other structural approaches to discourse [van Dijk, 1972, Longacre, 1983, Grosz and Sidner, 1986, Cohen, 1987, Mann and Thompson, 1988, Polanyi, 1988, Moser and

Moore, 1996], this book also assumes that trees are adequate representations of discourse structures. However, unlike previously proposed logical approaches that focus primarily on the semantic grounding of rhetorical, temporal, intentional, and causal relations [Hobbs, 1990, Lascarides and Asher, 1991, Lascarides et al., 1992, Lascarides and Oberlander, 1992, Lascarides and Asher, 1993, Asher, 1993, Kamp and Reyle, 1993, Hobbs et al., 1993, Asher and Lascarides, 1994, Hobbs, 1996], the approach in this book uses logic in order to distinguish between discourse structures that are valid and discourse structures that are not, and to determine all valid discourse structures of a text.

1.2.2 The Rhetorical Parsing of Free Texts

In the second part of the book, I present two approaches to deriving valid discourse structures for free texts. One approach relies primarily on discourse markers: it employs manually written rules, which were derived from a corpus analysis of more than 450 cue phrases. The other approach relies both on discourse markers and on robust syntactic and semantic knowledge sources: it employs rules that are derived automatically by applying machine learning techniques on data obtained from three corpora of manually annotated discourse trees.

The empirical investigation of the role that discourse markers play in segmenting texts into elementary discourse units and in signaling rhetorical relations that hold between discourse segments contributes to discourse research aimed at explicating the connection between lexicogrammar and discourse relations [Halliday and Hasan, 1976, Halliday, 1994, Martin, 1992, Lascarides and Asher, 1993, Asher, 1993, Asher and Lascarides, 1994, Hobbs, 1990, Ono and Thompson, 1996, Webber, 1998]. The corpus analysis does not employ unambiguous ontological principles in order to study the relation between discourse markers and discourse relations; nevertheless, it takes advantage of suggestions that were made by proponents of functional and psycholinguistic approaches to discourse [Halliday, 1994, Martin, 1992, Givón, 1995, Traxler and Gernsbacher, 1995, Hoover, 1997, Gernsbacher, 1997].

The algorithms proposed in the second part of the book contribute to computational accounts of discourse processing. They demonstrate an approach to language engineering that combines empirical linguistics with formal theories in order to produce programs that derive the discourse structures of free texts.

1.2.3 Summarization

In the third part of the book, I explore the utility of discourse structures in the context of text summarization. I present experiments and discourse-based summarization methods

and programs that confirm that the structure of discourse can be successfully exploited in a practical summarization setting.

In Chapter 9, I describe a psycholinguistic experiment that shows that text structures can be used effectively in order to select the most important units in a text; I propose a discourse-based summarization algorithm; and I evaluate an implementation of it. The implemented discourse-based summarizer derives the structure of the text given as input and then, on the basis of this structure, associates an importance score to each unit in the text. The units with highest score provide a summary of the text.

In Chapter 10, I show how other summarization heuristics can be tightly integrated into the discourse-based summarization algorithm and I present a simple learning mechanism that uses manually annotated extracts in order to increase the performance of the discourse-based summarizer. An evaluation of the discourse-based summarization program shows that it significantly outperforms both two baseline algorithms and Microsoft's Office97 summarizer.

1.3 Rhetorical Organization of the Book

General Remarks on the Layout of the Book In the previous section, I presented a part by part overview of the main issues that I address in this book. The book dwells on topics that range from formal, knowledge representation issues in text theory to issues in algorithms, linguistics, psycholinguistics, and language engineering. Because the book addresses diverse research topics, I decided to discuss the relevant research in connection with each particular topic. I hope that this will enable readers who are interested in only a particular aspect of the book to find their way around more easily. For the same reason, I have included a preamble and a summary at the beginning and end of each part.

A Rhetorical Map of the Book In order to facilitate navigation through this book, I also provide a rhetorical map in Figure 1.4, using the same conventions as in Figure 1.3.

According to the map, the "Introduction" chapter MOTIVATES the material presented in the rest of the book. Part I consists of four main segments. Chapter 2 provides the linguistic BACKGROUND for the material presented in the rest of Part I. Sections 3.1 and 3.2 describe ALTERNATIVE ways of formalizing the valid structures of texts: Section 3.1 formalizes what discourse structures are, while Section 3.2 formalizes how discourse structures can be derived. Chapter 4 ELABORATES on the material presented in Chapter 3 by providing algorithms that can be used to derive structures that are consistent with the axioms of valid text structures.

The second part of the book is divided into four segments. The corpus analysis of cue phrases in Section 6.2 ENABLES the development of the cue-phrase-based rhetorical parser

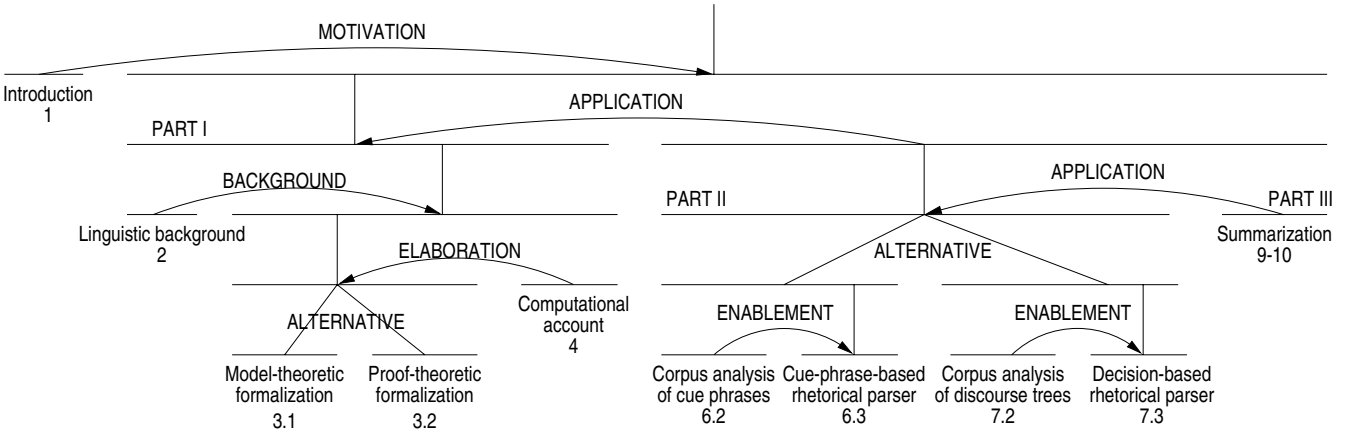

Figure 1.4
A rhetorical map of the book

in Section 6.3. An ALTERNATIVE approach to discourse parsing is to manually annotate a corpus of discourse trees (Section 7.2), which ENABLES the application of learning-based methods to rhetorical parsing (Section 7.3). A direct APPLICATION of the rhetorical parsers developed in Part II is the summarization work described in Part III (Chapters 9 and 10). In fact, both rhetorical parsers and the discourse-based summarizer can be seen as APPLICATIONS of the mathematical theory developed in Part I.

I THEORETICAL FOUNDATIONS

Preamble

One of the goals of this book is to provide a theory of text structures[1] that is general enough to be applicable to free texts, and simple enough to yield tractable discourse parsing algorithms. In what follows, I first discuss some of the most widely accepted linguistic properties of discourse structures that have been proposed previously. I then provide, from first principles, a first-order axiomatization of these properties. The axiomatization introduces a compositionality criterion that explains the relationship between discourse relations that hold between large textual spans in terms of discourse relations that hold between elementary discourse units. In addition to the axiomatization, I provide a proof theory that can be used as a backbone by a variety of implementations that are aimed at deriving the discourse structure of texts. The axiomatization and implementations of the proof theory can be used not only to label a given discourse structure as valid or invalid with respect to the theory presented here, but also to determine some or all valid discourse structures of a text.

1. In this book, I use the terms *discourse* and *text structure* interchangeably.

2 The Linguistics of Text Structures

2.1 Toward Formalizing the Structure of Free, Unrestricted Texts

2.1.1 The Linguistic Features of Text Structures

If we examine carefully the claims that current discourse theories make with respect to the *structure* of text and discourse, we will find significant commonalities. Essentially, most of these theories assume that the elementary textual units are non-overlapping spans of text; that discourse relations hold between textual units of various sizes; that some textual units play a more important role in text that others; and that the abstract structure of most texts is a tree. I now discuss each of these features in turn.

THE ELEMENTARY UNITS OF COMPLEX TEXT STRUCTURES ARE NON-OVERLAPPING SPANS OF TEXT. Although some researchers take the elementary units to be clauses [Grimes, 1975, Givón, 1983, Longacre, 1983], while others take them to be prosodic units [Hirschberg and Litman, 1987], turns of talk [Sacks et al., 1974], sentences [Polanyi, 1988], intentionally defined discourse segments [Grosz and Sidner, 1986], or the "contextually indexed representation of information conveyed by a semiotic gesture, asserting a single state of affairs or partial state of affairs in a discourse world" [Polanyi, 1996, p. 5], all agree that the elementary textual units are non-overlapping spans of text.

For example, if we take clauses to be the elementary units of text, the text fragment in 2.1 can be broken into four units, as shown below. The elementary units are delimited by square brackets.

[No matter how much one wants to stay a nonsmoker,^{A1}] [the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life.^{B1}] [We know that 3,000 teens start smoking each day,^{C1}] [although it is a fact that 90% of them once thought that smoking was something that they'd never do.^{D1}] [Pfau, 1995] (2.1)

DISCOURSE RELATIONS HOLD BETWEEN TEXTUAL UNITS OF VARIOUS SIZES. The nature, number, and taxonomy of the discourse relations that hold between textual units are controversial issues. At one end of a spectrum of influential proposals, we have the groundbreaking research that catalogued for the first time the "deep" relations that underlie the surface syntactic relations between clauses in complex sentences [Ballard et al., 1971, Grimes, 1975] (see also [Hovy and Maier, 1993] for an overview). Although they were not derived from first principles, these approaches provided the first "complete" taxonomies of discourse relations [Grimes, 1975]. At the other end of the spectrum, we have the approaches that take the position that taxonomies of relations should be created on the basis of some unambiguous principles. Such principles are derived from the lexicogrammatical resources that explicitly signal cohesive and rhetorical relations [Halliday and Hasan, 1976,

Martin, 1992]; from the types of inferences that the reader needs to draw in order to make sense of a text [Hobbs, 1990]; from the intentions that the writer had when she wrote the text [Grosz and Sidner, 1986]; from the effects that the writer intends to achieve [Mann and Thompson, 1988]; from the general cognitive resources that readers use when they process text [Sanders et al., 1992, Sanders et al., 1993, Spooren, 1997, Sanders, 1997]; from the linguistic evidence (such as cue phrases) of some linguistic psychological constructs that are used during text processing [Knott, 1995, Knott and Mellish, 1996, Knott and Dale, 1996, Knott and Sanders, 1998, Pander Maat, 1998]; from the presuppositions that rhetorical relations yield [Oversteegen, 1997]; from the propositional attitudes that underlie the communicative process [Bateman and Rondhuis, 1997]; and from a relational criterion that posits that relations should be included in a taxonomy only if they add some extra meaning to the meaning derivable from the textual units that they connect [Nicholas, 1994]. In spite of the heterogeneity of these approaches, one aspect is common to all of them: the assumption that relations *need* to be considered if one is to account for the meaning of text.

For example, if we adopt the set of relations proposed by Mann and Thompson [1988], we can say that a rhetorical relation of JUSTIFICATION holds between units A_1 and B_1 in text 2.1 because unit A_1 justifies the writer's right to present unit B_1. And an EVIDENCE relation holds between the first sentence (units A_1–B_1) and the second sentence (units C_1–D_1), because the second sentence supports the argument presented in the first.

SOME TEXTUAL UNITS PLAY A MORE IMPORTANT ROLE IN THE TEXT THAN OTHERS. The difference in importance between the roles played by the textual units that pertain to a given relation has been acknowledged from the beginning: in fact, the most important classification criterion in Grimes's [1975] taxonomy of relations is the distinction between *paratactic* relations, which are relations between units of equal importance, and *hypotactic* relations, which are relations between a unit that plays a central role and one that is subsidiary to the role played by the other unit. The distinction between paratactic and hypotactic relations is also explicitly acknowledged by Halliday and Hasan [1976] and Martin [1992]. The distinction between units that are important and units that are less important can be also induced from a discourse representation according to Grosz and Sidner [1986], by capitalizing on the dominance relations that hold between the intentions associated with embedded discourse segments. The same distinction is central to Mann and Thompson's theory [1988], in which rhetorical relations hold between so-called *nuclei (N)* and *satellites (S)*. The coordination and subordination structures in Polanyi's theory [1988, 1996] and the distinction between *core* and *contributor* in Moser and Moore's approach [1996, 2000] reflect the same difference in the relative importance of the units that are members of these structures.

For example, units A_2 and B_2 in text 2.2 below convey information pertaining to the average surface temperatures on Mars at the equator and at the poles respectively. In other words, each unit "talks about" a particular instance of the same thing—the average surface temperature. Therefore, we can say that a paratactic relation of LIST holds between units A_2 and B_2.

[Surface temperatures typically average about −60 degrees Celsius (−76 degrees Fahrenheit) at the equator^{A2}] [and can dip to −123 degrees C near the poles^{B2}] (2.2)

In contrast, if we reconsider units A_1 and B_1 in text 2.1, we easily notice that unit B_1 expresses what is most essential to the writer's purpose: the role that unit A_1 plays is subsidiary to the role played by unit B_1. Hence, we can say that a hypotactic relation of JUSTIFICATION holds between units A_1 and B_1.

THE ABSTRACT STRUCTURE OF MOST TEXTS IS A TREE. Most discourse and text theories mention explicitly or implicitly that trees are good mathematical abstractions of discourse and text structures [van Dijk, 1972, Longacre, 1983, Grosz and Sidner, 1986, Mann and Thompson, 1988, Polanyi, 1988, Lascarides and Asher, 1993, Polanyi, 1996, Moser and Moore, 1996, Walker, 1998]. For example, a possible tree-like representation of the discourse structure that pertains to units A_1–D_1 in text 2.1 is shown in Figure 2.1: the leaves of the tree correspond to elementary units and the internal nodes correspond to textual spans that are obtained through the juxtaposition of the immediate subspans; the satellite nodes are surrounded by dotted lines, nuclei by solid lines.

Unlike the other three features of discourse structures that we have discussed so far, the assumption that trees are adequate abstractions of discourse structures is the only assumption that has received serious criticism: it seems that certain classes of texts, such as argumentative texts [Toulmin et al., 1979, Birnbaum et al., 1980, Birnbaum, 1982, Zukerman and McConachy, 1995] and certain dialogues [Carberry et al., 1993] are better represented using graphs. Although I subscribe to the position that some texts are better represented using graph-based structures, I believe that trees are an adequate approximation in the majority of the cases. (In fact, Cohen [1983, 1987] shows that even arguments can be modeled as trees.) Since tree-based structures are also easier to formalize and derive automatically, it is such structures that I will concentrate my attention on for the rest of the book.

2.1.2 The Problem of Formalizing Text Structures

The four features that I have discussed in Section 2.1.1 constitute the foundations of my formalization. In other words, I take as axiomatic that any text can be partitioned into a

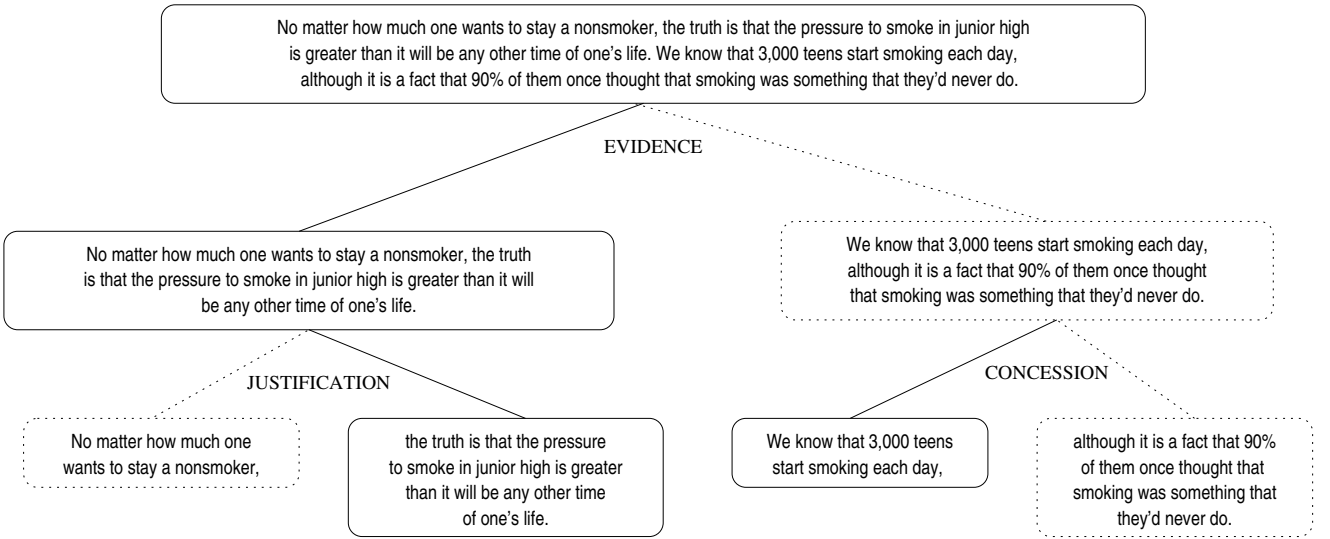

Figure 2.1
An example of a tree-like discourse structure that corresponds to text 2.1

sequence of non-overlapping, elementary textual units and that a text structure, i.e., a tree, can be associated with the text such that:

- The elementary textual units constitute the leaves of the tree.
- The leaves in the tree are in the same order as the elementary units in the text. In other words, an in-order traversal of the leaves of a text structure tree yields the sequence of elementary units in the original text.
- The tree obeys some well-formedness constraints that could be derived from the semantics and pragmatics of the elementary units and the relations that hold among these units. Were such constraints not obeyed, any tree would be appropriate to account for the rhetorical relations that hold between textual units of different sizes, which is obviously unreasonable.
- The relations that are used to connect textual units of various sizes fall into two categories: paratactic and hypotactic.

The formalization of text structures can then be equated with the problem of finding a declarative specification of the constraints that characterize well-formed text trees.

Before getting into the details of the formalization, I would like to draw the attention of the reader to the fact that the formalization is independent of the set of relations that it relies upon. The only assumption behind the formalization is that such a set exists and that relations in this set are either paratactic or hypotactic. Obviously, when the formalization discussed here is grounded in a set of rhetorical relations with a clearly defined semantics, this semantics will impose further constraints on discourse structures, which are not captured by the formalization discussed here.

Presenting the formalization only in abstract terms will make the reading difficult. To avoid this, I will mainly use in my examples the set of relations that was developed by Mann and Thompson [1988]. In what follows, I will primarily refer to the relations that hold between textual units as *rhetorical* or *discourse relations*. For the uninitiated reader, I first provide a short introduction to Mann and Thompson's theory and taxonomy of relations. Readers familiar with Mann and Thompson's theory may skip to Section 2.2.2.

2.2 Rhetorical Structure Theory

2.2.1 Background Information

Driven mostly by research in natural language generation, Rhetorical Structure Theory (RST) [Mann and Thompson, 1988] has become one of the most popular discourse theories of the last decade [Hovy, 1988, Scott and de Souza, 1990, Moore and Swartout, 1991,

Relation name:	EVIDENCE
Constraints on N:	The reader *R* might not believe the information that is conveyed by the nucleus *N* to a degree satisfactory to the writer *W*.
Constraints on S:	The reader believes the information that is conveyed by the satellite *S* or will find it credible.
Constraints on N + S combination:	*R*'s comprehending *S* increases *R*'s belief of *N*.
The effect:	*R*'s belief of *N* is increased.
Locus of the effect:	*N*.
Example:	[The truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life:^{B_1}] [we know that 3, 000 teens start smoking each day.^{C_1}]

Figure 2.2
The definition of the EVIDENCE relation in Rhetorical Structure Theory [Mann and Thompson, 1988, p. 251]

Cawsey, 1991, McCoy and Cheng, 1991, Horacek, 1992, Hovy, 1993, Moore and Paris, 1993, Vander Linden and Martin, 1995]. In fact, even the critics of the theory are not interested in rejecting it so much as in fixing unsettled issues such as the ontology of the relations [Hovy, 1990, Rösner and Stede, 1992, Maier, 1993, Hovy and Maier, 1993], the problematic mapping between rhetorical relations and speech acts [Hovy, 1990] and between intentional and informational levels [Moore and Pollack, 1992, Moore and Paris, 1993], and the inability of the theory to account for interruptions [Cawsey, 1991].

Central to Rhetorical Structure Theory is the notion of *rhetorical relation*, which is a relation that holds between two non-overlapping text spans called *nucleus* (*N*) and *satellite* (*S*). There are a few exceptions to this rule: some relations, such as CONTRAST, are multinuclear. The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's purpose than the satellite; and that the nucleus of a rhetorical relation is comprehensible independent of the satellite, but not vice versa.

Text coherence in RST is assumed to arise due to a set of constraints and an overall effect that are associated with each relation. The constraints operate on the nucleus, on the satellite, and on the combination of nucleus and satellite. For example, an EVIDENCE relation (see Figure 2.2) holds between the nucleus B_1 and the satellite C_1, because the nucleus B_1 presents some information that the writer believes to be insufficiently supported to be accepted by the reader; the satellite C_1 presents some information that is thought to be believed by the reader or that is credible to her; and the comprehension of the satellite increases the reader's belief in the nucleus. The effect of the relation is that the reader's belief in the information presented in the nucleus is increased.

Rhetorical relations can be assembled into rhetorical structure trees (RS-trees) on the basis of five structural constituency schemata, which are reproduced in Figure 2.3 from

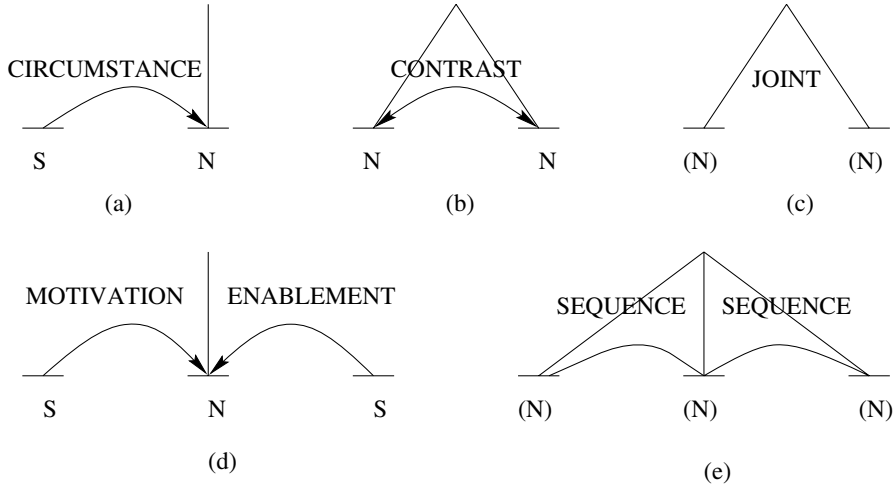

Figure 2.3
Examples of the five types of schema that are used in RST [Mann and Thompson, 1988, p. 247]. The arrows link the satellite to the nucleus of a rhetorical relation. Arrows are labeled with the name of the rhetorical relation that holds between the units over which the relation spans. The horizontal lines represent text spans and the vertical and diagonal lines represent identifications of the nuclear spans. In the SEQUENCE and JOINT relations, the vertical and diagonal lines identify nuclei by convention only, since there are no corresponding satellites.

Mann and Thompson [1988]. The large majority of rhetorical relations are assembled according to the pattern given in Figure 2.3a. Schema 2.3d covers the cases in which a nucleus is connected with multiple satellites by possibly different rhetorical relations. Schemata 2.3b, 2.3c, and 2.3e cover the multinuclear (paratactic) relations.

According to Mann and Thompson [1988], a canonical analysis of a text is a set of schema applications for which the following constraints hold:

$$\left\{\begin{array}{ll} \textbf{Completeness:} & \text{One schema application (the root) spans the entire text.} \\ \textbf{Connectedness:} & \text{Except for the root, each text span in the analysis is either} \\ & \text{a minimal unit or a constituent of another schema} \\ & \text{application of the analysis.} \\ \textbf{Uniqueness:} & \text{Each schema application involves a different set of text spans.} \\ \textbf{Adjacency:} & \text{The text spans of each schema application constitute one} \\ & \text{contiguous text span.} \end{array}\right. \tag{2.3}$$

Obviously, the formulation of the constraints that Mann and Thompson put on the discourse structure 2.3 is just a sophisticated way of saying that rhetorical structures are trees in which sibling nodes represent contiguous text. The distinction between the nucleus and

the satellite of a rhetorical relation is their acknowledgment that some textual units play a more important role in text than others, i.e., some relations are hypotactic, while others are paratactic. Because each textual span can be connected to another span by only one rhetorical relation, each unit plays either a nucleus or a satellite role. Since Mann and Thompson also take the elementary units to be non-overlapping spans of text, RST is fully compatible with the essential features of text structures that I discussed in Section 2.1.1.

2.2.2 Compositionality in RST

Despite its popularity, RST still lacks two things:

• A formal specification that would allow one to distinguish between well- and ill-formed rhetorical structure trees;

• Algorithms that would enable one to determine all the possible rhetorical analyses of a given discourse.

In this section, I show that these problems are primarily due to a lack of "compositionality" in RST, which would explain the relationship between rhetorical relations that hold between large textual spans and rhetorical relations that hold between elementary units and would enable an unambiguous determination of span boundaries. In order to ground the discussion, consider again text 2.1, which I reproduce for convenience below.

[No matter how much one wants to stay a nonsmoker,^{A1}] [the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life.^{B1}] [We know that 3,000 teens start smoking each day,^{C1}] [although it is a fact that 90% of them once thought that smoking was something that they'd never do.^{D1}] (2.4)

Assume, for the moment, that we do not analyze this text as a whole, but, rather, that we determine what rhetorical relations could hold between every pair of elementary textual units. When we apply Mann and Thompson's definitions [1988], we obtain the set given below.

$$RR = \begin{cases} rhet\_rel(\text{JUSTIFICATION}_0, \text{A}_1, \text{B}_1) \\ rhet\_rel(\text{JUSTIFICATION}_1, \text{D}_1, \text{B}_1) \\ rhet\_rel(\text{EVIDENCE}, \text{C}_1, \text{B}_1) \\ rhet\_rel(\text{CONCESSION}, \text{D}_1, \text{C}_1) \\ rhet\_rel(\text{RESTATEMENT}, \text{D}_1, \text{A}_1) \end{cases} \quad (2.5)$$

These relations hold because the understanding of both A_1 (teens want to stay nonsmokers) and D_1 (90% of the teens think that smoking is something that they would never do) will

increase the reader's readiness to accept the writer's right to present B_1 (the pressure on teens to start smoking is greater than it will be any other time of their lives); the understanding of C_1 (3,000 teens start smoking each day) will increase the reader's belief of B_1; the recognition of D_1 as something compatible with the situation presented in C_1 will increase the reader's negative regard for the situation presented in C_1; and the situation presented in D_1 is a restatement of the situation presented in A_1. Throughout this book, I use the convention that rhetorical relations are represented as sorted, first-order predicates having the form *rhet_rel*(*name*, *satellite*, *nucleus*) and multinuclear relations are represented as predicates having the form *rhet_rel*(*name*, $nucleus_1$, $nucleus_2$). To avoid confusion, rhetorical relation names are associated with unique identifiers, i.e., subscripts that range over the set of natural numbers. When the subscript is not shown, it is assumed to be 0.

Assume now that one is given the task of building an RS-tree for text 2.4, according to the constraints put forth by RST [Mann and Thompson, 1988]. That is, assume that one is asked to build trees whose leaves are the elementary units A_1–D_1 and whose internal nodes subsume contiguous spans of text.

Consider now that one produces the candidates in Figure 2.4. Any student in RST would notice from the beginning that the tree in Figure 2.4d is illegal with respect to the requirements specified by Mann and Thompson [1988] because C_1 belongs to more than one text span, namely [A_1, C_1] and [C_1, D_1]. However, even a specialist in RST will have trouble determining whether the trees in Figure 2.4a–c represent *all* the possible ways in which a rhetorical structure could be assigned to text 2.4, and moreover, in determining if these trees are *correct* with respect to the requirements of RST. To my knowledge, neither the description provided by Mann and Thompson nor any other formalization that has been proposed for RST is capable of providing sufficient help in resolving these problems. Even if we choose a different discourse theory, such as those proposed by Hobbs [1990], Martin [1992], Grosz and Sidner [1986], or Polanyi [1988], we will still be unable to enumerate *all* valid discourse interpretations and *determine formally and computationally whether a given representation is valid with respect to the constraints put forth by these theories*.

The reason for this is that the discourse structures in all these theories are either underspecified or incomplete with respect to some compositionality requirements that would be necessary in order to formulate precisely the conditions that have to be satisfied if two adjacent spans are to be put together. Assume, for example, that an analyst is given text 2.4 and the set of rhetorical relations that pertain to the minimal units 2.5, and that that analyst takes the reasonable decision to build the spans [A_1, B_1] and [C_1, D_1], as shown in Figure 2.5. To complete the construction of the discourse tree, the analyst will have to decide what the best relation is that could span over [A_1, B_1] and [C_1, D_1]. If he considers the

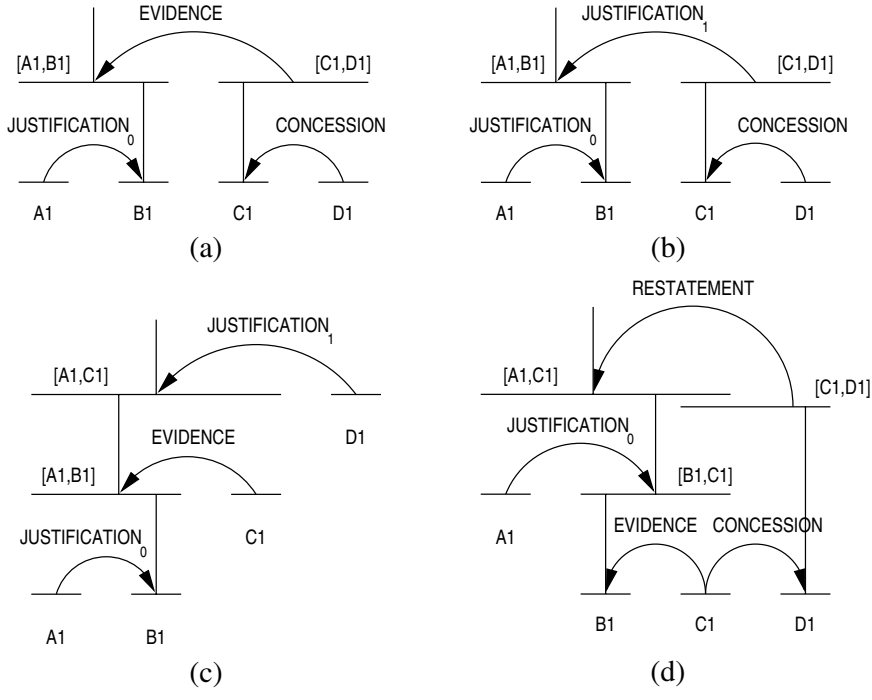

Figure 2.4
A set of possible rhetorical analyses of text 2.4

elementary relations 2.5 that hold across the two spans, she has three choices, which correspond to the relations *rhet_rel*(JUSTIFICATION$_1$, D$_1$, B$_1$), *rhet_rel*(EVIDENCE, C$_1$, B$_1$), and *rhet_rel*(RESTATEMENT, D$_1$, A$_1$). Which is the correct one to choose?

More generally, suppose that the analyst has already built two partial discourse trees on the top of two adjacent spans that consist of ten and twenty minimal units, respectively. Is it correct to join the two partial trees in order to create a bigger tree just because there is a rhetorical relation that holds between two arbitrary minimal units that happen to belong to those spans? One possible answer is to say that rhetorical relations are defined over spans that are larger than one unit too; therefore, in our case, it is correct to put the two partial trees together if there is a rhetorical relation that holds between the two spans that we have considered. But if this is the case, how did we determine the precise boundaries of the spans over which that relation holds? And how do the rhetorical relations that hold between minimal units relate to the relations that hold between larger text spans? Current discourse theories that are aimed to be applicable to the study of unrestricted texts [Mann and

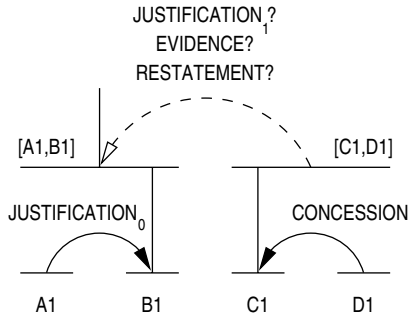

Figure 2.5
An example of the ambiguity that pertains to the construction of RS-trees

Thompson, 1988, Grosz and Sidner, 1986, Hobbs, 1990, Polanyi, 1996] provide no explicit answer for these questions. And no explicit answers are given by logic-based [Zadrozny and Jensen, 1991, Lascarides and Asher, 1993, Asher, 1993] and grammar-based [van Dijk, 1972, Polanyi, 1988, Scha and Polanyi, 1988, Gardent, 1994, Hitzeman et al., 1995, Polanyi and van den Berg, 1996, van den Berg, 1996, Gardent, 1997, Schilder, 1997, Cristea and Webber, 1997] approaches to discourse either.

2.3 The Formulation of a Compositionality Criterion of Valid Text Structures

2.3.1 A Weak Compositionality Criterion

Despite the lack of a formal specification of the conditions that must hold in order to join two adjacent textual units, I believe that some of the discourse theories contain such a condition implicitly. As I have mentioned before, during the development of RST, Mann and Thompson [1988] and Matthiessen and Thompson [1988] noticed that what is expressed by the nucleus of a rhetorical relation is more essential to the writer's purpose than the satellite; and that the satellite of a rhetorical relation is incomprehensible independent of the nucleus, but not vice versa. Consequently, it has been often argued [Mann and Thompson, 1988, Moser and Moore, 2000] that deleting the nuclei of the rhetorical relations that hold among all textual units in a text yields an incomprehensible text, while deleting the satellites of the rhetorical relations that hold among all textual units in a text yields a text that is still comprehensible. In fact, as Matthiessen and Thompson put it, "the nucleus-satellite relations are pervasive in texts independently of the grammar of clause combining" [Matthiessen and Thompson, 1988, p. 290].

[Smart cards are not a new phenomenon.^{A3}] [They have been in development since the late 1970s and have found major applications in Europe, with more than a quarter of a billion cards made so far.^{B3}] [The vast majority of chips have gone into prepaid, disposable telephone cards, but even so the experience gained has reduced manufacturing costs, improved reliability and proved the viability of smart cards.^{C3}] [International and national standards for smart cards are well under development to ensure that cards, readers and the software for the many different applications that may reside on them can work together seamlessly and securely.^{D3}] [Standards set by the International Organization for Standardization (ISO), for example, govern the placement of contacts on the face of a smart card so that any card and reader will be able to connect.^{E3}]

Figure 2.6
Fragment of the "Smart cards" text *(Scientific American, August 1996)*

In my own work (see Section 6.2), I have analyzed 2100 text fragments that were extracted from the Brown corpus and investigated the relationship between discourse marker occurrences, the text segments they relate, and the rhetorical relations they signal. And in collaboration with Estibaliz Amorrortu and Magdalena Romera [Marcu et al., 1999a], we constructed a corpus of discourse trees for 90 texts that spanned the news, editorial, and scientific genres (see Section 7.2). During these analyses, it became apparent that in order to determine the types of relations that held between large spans and the boundaries of these spans, it was often useful to mentally associate abstractions to the spans under consideration. Instead of assessing whether a relation held between two large textual segments, it was often easier to associate some unique abstractions with the segments under consideration and then determine the rhetorical relation between these abstractions. These unique abstractions were most of the time derivable from the nuclei of the segments under consideration.

For example, in building the discourse structure of the text shown in Figure 2.6, one can easily notice that sentences A_3–C_3 "talk about" smart cards, and sentences D_3–E_3 about standards for smart cards.[1] The discourse tree in Figure 2.7 reflects this: it has one subtree that subsumes sentences A_3–C_3, and one subtree that subsumes sentences D_3–E_3. Sentence B_3 is EVIDENCE to the assertion in sentence A_3, sentence C_3 ELABORATES on sentence B_3, and sentence E_3 is an EXAMPLE with respect to the information presented in sentence D_3. In assessing the relationship between segments A_3–C_3 and D_3–E_3, we need not consider the information pertaining to sentences B_3, C_3, and E_3 because this information is subsidiary to sentences A_3 and D_3 respectively. In this case, it seems sufficient if we focus our attention on what the most important sentences in these two segments are about. Hence,

1. For simplicity, I consider here that the elementary units of the discourse representation are sentences.

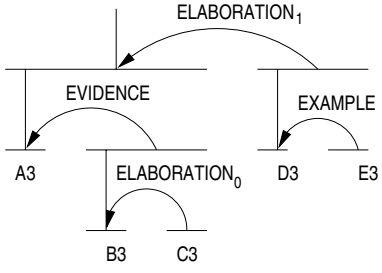

Figure 2.7
A rhetorical analysis of the text in Figure 2.6

we can, for example, determine that an ELABORATION relation holds between segments A_3–C_3 and D_3–E_3 because sentence A_3, which is the most important sentence of segment A_3–C_3, "talks about" smart cards, while sentence D_3, which is the most important sentence of segment D_3–E_3, "talks about" standards for smart cards. And the standards for smart cards ELABORATE on smart cards.

The idea that rhetorical relations that hold between large spans can be explained in terms of relations that hold between salient abstractions that characterize those spans is also implicit in most of the work in text planning, in the field of natural language generation [Hovy, 1993, Moore and Paris, 1993, Moore and Swartout, 1991, Cawsey, 1991, Maybury, 1992, Meteer, 1992]. Hierarchical text planning algorithms usually start with a high-level communicative goal, which is incrementally expanded into a tree. As a consequence, the rhetorical relations that hold between large textual spans, which eventually correspond to large subtrees, also hold between the abstractions that were associated with those subtrees before their expansion. In the natural language generation literature, these abstractions usually correspond to beliefs, intentions, and communicative action effects.

A careful analysis of the discourse structures that Mann, Thompson, Grosz, Sidner, Hobbs, and many others built and my own empirical investigations of naturally occurring texts have led me to formulate the following compositionality criterion:

PROPOSITION 2.1 **A weak compositionality criterion of valid text structures:** *If a relation R holds between two nodes of the tree structure of a text, then R can be explained in terms of a similar relation R that holds between two linguistic or nonlinguistic constructs that pertain to the most important constituents of those nodes.*

An alternative phrasing for "then *R* can be explained in terms of a similar relation *R* that holds between" could be "then R also holds between." I chose the first phrasing because linguistically and mathematically a relation that holds between two elementary textual

units is not the same with a relation that holds between two textual spans. The phrasing "linguistic or nonlinguistic constructs" in proposition 2.1 is meant to be general enough to cover all the possible elements that could be used in the definition of the taxonomy of relations that one adopts. For example, intentions are the nonlinguistic constructs that underlie Grosz and Sidner's Theory [1986] (GST)—all relations in GST are defined in terms of the intentions that are associated with the discourse segments. Knowledge about the world provides grounding for the nonlinguistic constructs that are used by Hobbs. In RST the relations make reference both to linguistic constructs that pertain to the semantics of the spans and to nonlinguistic constructs, such as beliefs, attitudes, and goals.

To understand better the claim that proposition 2.1 makes, let us restrict again our attention to the taxonomy of relations that was proposed by Mann and Thompson and reconsider the trees in Figure 2.4. If we examine tree 2.4a, we can notice that this tree is consistent with the compositionality criterion: the EVIDENCE relation that holds between text spans [C_1, D_1] and [A_1, B_1] holds between their most salient parts as well, i.e., between the nuclei C_1 and B_1. In this case, the linguistic constructs that the compositionality criterion refers to are clauses C_1 and B_1. Both of these clauses are the most important constituents (nuclei) of the spans that they belong to and an EVIDENCE relation holds between them.

As we have already seen, in the general case, the constructs that the compositionality criterion refers to need not be clauses. Consider the following example:

[He wanted to play squash with Janet,^{A4}] [but he also wanted to have dinner with Suzanne.^{B4}] [He went crazy.^{C4}] (2.6)

The RS-tree in Figure 2.8 shows the RST analysis of text 2.6, in which units A_4 and B_4 are connected through a CONTRAST relation. The text span that results, [A_4, B_4], is further connected with textual unit C_4 through a NONVOLITIONAL CAUSE relation. Note, however, that in this case, the NONVOLITIONAL CAUSE relation holds neither between A_4 and C_4, nor

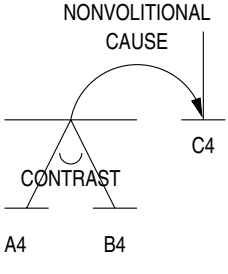

Figure 2.8
A rhetorical analysis of text 2.6

[There are two houses you might be interested in:^{A5}]
[House A is in Palo Alto.^{B5}] [It's got 3 bedrooms and 2 baths,^{C5}] [and was built in 1950.^{D5}] [It's on a quarter acre, with a lovely garden,^{E5}] [and the owner is asking $425K.^{F5}] [But **that**'s all I know about it.^{G5}]
[House B is in Portola Valley.^{H5}] [It's got 3 bedrooms, 4 baths and a kidney-shaped pool,^{I5}] [and was also built in 1950.^{J5}] [It's on 4 acres of steep wooded slope, with a view of the mountains.^{K5}] [The owner is asking $600K.^{L5}] [I heard all **this** from a friend,^{M5}] [who saw the house yesterday.^{N5}]
[Is **that** enough information for you to decide which to look at?^{P5}]

Figure 2.9
A text example used by Webber [1988, p. 115]

between B_4 and C_4. Rather, the relation shows that the incompatibility between the two desires, which is subsumed by the CONTRAST between A_4 and B_4, caused the situation presented in C_4. In this case, the constructs that the compositionality criterion refers to are the textual unit C_4, and the incompatibility between the two desires, which is made explicit by the CONTRAST relation that holds between units A_4 and B_4. If C_4 were "This indecisiveness drove him crazy" instead of "He went crazy", the "indecisiveness" would even refer to this incompatibility. Note also that the CONTRAST relation is a multinuclear (or paratactic) relation that assigns the rhetorical status of NUCLEUS to both units A_4 and B_4. Since both A_4 and B_4 are the most important units of span $[A_4, B_4]$, it follows that the incompatibility between the desires expressed in A_4 and B_4 is also an important construct of the span, which is consistent with the compositionality criterion given in proposition 2.1.

The linguistic constructs that proposition 2.1 mentions could take a wide range of forms. Consider the text in Figure 2.9 which was first used by Webber [1988a, p. 115]. One of Webber's main claims is that some discourse segments are characterized by "entities" that are distinct from the entities that are expressed explicitly therein. The fact that naturally occurring texts contain references to such entities proves the validity of Webber's proposal. For example, the first boldfaced "that" in Figure 2.9 refers not to house A, an entity explicitly mentioned in the discourse, but to the *description* of that house. Similarly, the boldfaced "this" refers to the description of house B. And the last boldfaced "that" refers to the description of the two houses taken together.

Figure 2.10 shows the RST analysis of the text in Figure 2.9. To demonstrate that this RST analysis and the kind of discourse deixis proposed by Webber [1988a, 1991] are consistent with the compositionality criterion given in proposition 2.1, I will use an informal, "bottom-up" analysis: each of the textual spans $[B_5, F_5]$ and $[H_5, L_5]$ contains a set of elementary units that are connected by a LIST relation. The linguistic constructs that these sets of units induce are the descriptions of the two houses; these constructs

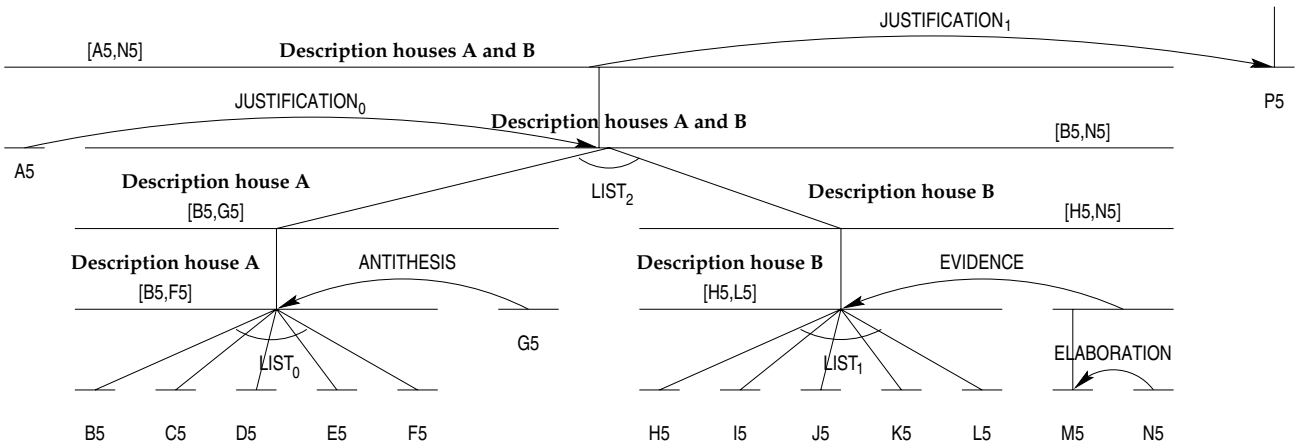

Figure 2.10
A rhetorical analysis of the text in Figure 2.9

are shown in boldface fonts in Figure 2.10. Text unit G_5 specifies only that the content presented in units B_5–F_5 is all that the writer knows. At the time unit G_5 is produced, the construct **Description house A** is already available for reference, so this explains why the first boldfaced "that" in the text in Figure 2.9 makes sense. Because **Description house A** is an important construct of span [B_5, F_5], and because [B_5, F_5] is the nucleus of the span [B_5, G_5], it is natural to consider that **Description house A** is an important construct for span [B_5, G_5] as well. Reasoning similarly, we can explain why the boldfaced "this" makes sense and why **Description house B** is an important construct for span [H_5, N_5]. Because spans [B_5, G_5] and [H_5, N_5] are connected through a LIST relation, i.e., a multinuclear relation that lists the description of the two houses, the important constructs of each of them could be promoted to the higher level span, [B_5, N_5]. This explains why **Description houses A and B** is an important construct of span [B_5, N_5]. Following the same procedure, **Description houses A and B** becomes an important construct for span [A_5, N_5], which explains why the second boldfaced "that" in the text in Figure 2.9 makes sense.

Again, as in the previous cases, the interpretation given above is consistent with the compositionality criterion. For example, the ANTITHESIS relation between span [B_5, F_5] and unit G_5 can be explained by a relation that holds between the construct **Description house A** and unit G_5. The LIST relation between spans [B_5, G_5] and [H_5, N_5] is also consistent with the fact that the descriptions of the two houses are listed after justification A_5 is given.

Formalization of the compositionality criterion given in proposition 2.1 would require the existence of well-developed formalisms that accommodate beliefs, intentions, and goals, and a full account of the relation between these constructs and their linguistic representations. Unfortunately, such an account is beyond the current state of the art of computational linguistics and artificial intelligence. Since my purpose is to provide a working theory of the structure of *unrestricted texts,* I cannot take compositionality criterion 2.1 as foundational because it is too underspecified.

2.3.2 A Strong Compositionality Criterion

Although compositionality criterion 2.1 is too weak to be exploited in a formal or computational model, I believe that one can still contribute to the general understanding of text by constructing a theory that takes as foundational a stronger criterion, which approximates criterion 2.1. The intuitive notion behind the stronger criterion is that, after all, all the linguistic and nonlinguistic constructs that are used as arguments of rhetorical relations can be derived from the textual units and the relations that pertain to those units. Since we do not know how to properly represent and reason about the linguistic and nonlinguistic constructs that we brought up in the previous section and since we do not know how to derive the nonlinguistic ones from the linguistic ones, we will simply ignore them for the moment. Textual units, i.e., clauses, sentences, and paragraphs, are constructs that we are familiar

with and that we do know how to handle. Therefore, I will use only these constructs in the formalization. These assumptions strengthen the weak compositionality criterion, as shown in proposition 2.2, below.

PROPOSITION 2.2 **A strong compositionality criterion of valid text structures:** *If a rhetorical relation* R *holds between two textual spans of the tree structure of a text, then it can be explained by a similar relation R that holds between at least two of the most important textual units of the constituent spans.*

If we reconsider text 2.4 and the tree in Figure 2.4a from the perspective of the strong compositionality criterion, we get the same interpretation as in the case of the weak compositionality criterion: the tree is consistent with the strong compositionality criterion because the EVIDENCE relation that holds between text spans [C_1, D_1] and [A_1, B_1] is explained by an EVIDENCE relation that holds between their most important subspans, i.e., between the spans C_1 and B_1.

In the case of text 2.6, whose RS-tree is given in Figure 2.8, the NONVOLITIONAL CAUSE relation that holds between spans [A_4, B_4] and C_4 is explained by a NONVOLITIONAL CAUSE relation that holds either between A_4 and C_4, B_4 and C_4, or both A_4, B_4 and C_4—the most important units of span [A_4, B_4] are both A_4 and B_4. Note that although, in this case, the strong compositionality criterion does not spell out precisely the elements between which the NONVOLITIONAL CAUSE relation holds, a potential reader of text structure 2.8 could identify that by herself because both units A_4 and B_4 are considered important for span [A_4, B_4] and therefore, these important units implicitly represent the incompatibility between the desires expressed in them.

In the case of the text in Figure 2.9, whose RS-tree is shown in Figure 2.10, the rhetorical relation between spans [B_5, G_5] and [H_5, N_5] is explained by a relation that holds between the most important units of subspans [B_5, F_5] and [H_5, L_5]. As in the previous cases, this constraint is stronger than that postulated by the weak compositionality criterion, i.e., it enables automatic inferences to be drawn, although it does not mention explicitly the constructs between which the relation holds. However, the information that pertains to the weak compositionality criterion is still implicit in the representation because the constructs **Description house A** and **Description house B** are implicitly encoded in the important units of spans [B_5, F_5] and [H_5, L_5], respectively.

Note that the strong compositionality criterion does not employ an "if and only if" condition because such a condition would prevent one from treating properly cases in which more than one rhetorical relation is hypothesized to hold between two elementary units. Because rhetorical relation definitions are sometimes ambiguous and sensitive to context and user beliefs, situations with contradictory hypotheses cannot be excluded.

Moore and Pollack [1992] have already shown that different analyses (intentional and semantic) could yield different rhetorical judgments. For example, in text 2.7, which is reproduced from [Moore and Pollack, 1992, p. 542], a semantic reading will assign a rhetorical relation of CONDITION to hold between satellite A_6 and nucleus B_6, while an intentional reading will assign a rhetorical relation of MOTIVATION to hold between satellite B_6 and nucleus A_6.

[Come home by 5:00.^{A6}] Then we can go to the hardware store before it closes.^{B6}] [That way we can finish the bookshelves tonight.^{C6}] (2.7)

If we used an "if and only if" condition in proposition 2.2, we would not be able to derive a valid discourse tree for text 2.7 because in any valid tree two spans can be connected by, at most, one relation. By using, in the compositionality criterion 2.2, only an "if" condition, we merely stipulate that the relations that hold between large spans must have counterpart relations that hold between elementary units as well. This formulation leaves room for producing valid discourse interpretations even in cases in which different elementary relations are hypothesized to hold between the same two elementary units.

2.4 From Texts to Discourse Structures

So far, I have focused only on the linguistic properties of valid discourse structures. Before I lay out the axioms that concern these properties, it would be useful to elaborate on how the resulting theory is going to be used in conjunction with real texts. To simplify the presentation of the structural properties of discourse structures, I have assumed, so far, that rhetorical relations that hold between large spans should be always explained in terms of rhetorical relations that hold between elementary units; and that rhetorical relations that hold between elementary units can be determined unambiguously. Obviously, both these assumptions are too strong.

THE NEED FOR DEALING WITH EXTENDED RHETORICAL RELATIONS. In presenting the weak and strong compositionality criteria, I have focused on the need to explain the rhetorical relations that hold between large spans in terms of rhetorical relations that hold between elementary units. Nevertheless, in some cases, human judges (as well as computational systems) can determine the rhetorical relations that hold between large spans without reasoning about the abstractions that are salient in the spans. For example, a text fragment may consist of three paragraphs, clearly marked by the connectives *First, Second,* and *Third*. For such a fragment, it is likely that the three paragraphs are in a LIST or SEQUENCE relation. If a computer program exploits the occurrences 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 book 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.

THE NEED FOR DEALING WITH NONDETERMINISM. So far, I emphasized that the need for a compositionality criterion and a clear formalization of the mathematical properties of valid discourse structures is orthogonal to the need of employing unambiguous definitions of rhetorical relations. In the examples presented so far, I have assumed that the rhetorical relations that hold between elementary units can be determined precisely. Again, this may prove to be too strong an assumption.

For example, consider a scenario in which a human judge hypothesizes that certain rhetorical relations hold between various textual segments and then a computational system chooses the combination of hypotheses that yields the "best" discourse interpretation. In analyzing the text shown in 2.8 below, an analyst may have trouble determining whether the rhetorical relation that holds between unit A_7 and segment $[B_7, C_7]$ is CONCESSION or ANTITHESIS.

["We are striving to have a strong renewed creative partnership with Coca-Cola," Mr. Dooner says.^{A7}] [However, odds of that happening are slim^{B7}] [since word from Coke headquarters in Atlanta is that CAA and other ad agencies, such as Fallon McElligott, will continue to handle Coke advertising.^{C7}] (2.8)

In such a case, it may be adequate if the analyst could specify that one (and only one) of these two relations holds between the two segments. The hypothesis shown in 2.9 reflects this ambiguity by means of an exclusive disjunction, $\oplus$.

$$\textit{rhet\_rel\_ext}(\text{ANTITHESIS}, A_7, [B_7, C_7]) \oplus \textit{rhet\_rel\_ext}(\text{CONCESSION}, A_7, [B_7, C_7]) \quad (2.9)$$

In the case a computational system attempts to build a discourse structure with no help from a human judge, the problem of hypothesizing unambiguously rhetorical relations is even less feasible. Consider, for example, that an automatic analyzer is given as input the text shown in 2.10, below.

[John likes sweets.^{A8}] [Most of all, John likes ice cream and chocolate.^{B8}] [*In constrast,* Mary likes fruits.^{C8}] [*Especially* bananas and strawberries.^{D8}] (2.10)

For the sake of argument, assume that an automatic analyzer identifies the elementary units of text 2.10 as labeled above and that it tries to hypothesize rhetorical relations between these units. During a corpus study that is to be discussed in Section 6.2, I have

noticed that in all its occurrences in a sample of texts, the connective *In constrast* signaled a CONTRAST relation. Hence, it is likely that *In contrast* signals a CONTRAST relation in text 2.10 as well. Unfortunately, although we know the relation that *In contrast* signals, we do not know what the spans that the CONTRAST relation holds between are: does the relation hold between spans [A_8, B_8] and [C_8, D_8]; or between unit A_8 and span [C_8, D_8]; or between span [A_8, B_8] and unit C_8; or between units A_8 and C_8; or between other units and spans? The best thing 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.

Given that the strong compositionality criterion 2.2 enables us to determine rhetorical relations that hold between large segments on the basis of rhetorical relations that hold between elementary units, it is sufficient if we hypothesize, with respect to the occurrence of the connective *In constrast*, the exclusively disjunctive hypothesis *rhet_rel*(CONTRAST, A_8, C_8) $\oplus$ *rhet_rel*(CONTRAST, A_8, D_8) $\oplus$ *rhet_rel*(CONTRAST, B_8, C_8) $\oplus$ *rhet_rel*(CONTRAST, B_8, D_8), because this hypothesis subsumes all the other possible rhetorical relations that may be signaled by the connective.

Instead of imposing that rhetorical relations must be hypothesized precisely, we should then more reasonably accept that discourse is ambiguous and that disjunctive hypotheses of rhetorical relations provide a better model of it. Given the text in 2.10 for example, a computer program may be able to hypothesize the relations shown in 2.11 using only knowledge of cue phrases and cohesion. The exclusive disjunction that uses the relation CONTRAST can be hypothesized on the basis of the occurrence of the marker *In contrast*. The rhetorical relation of ELABORATION between sentence A_8 and B_8 may be hypothesized using cohesion, by noticing that both sentences "talk about" John. And the exclusive disjunctive relation of ELABORATION may be hypothesized on the basis of the occurrence of the cue phrase *Especially*.

$$RR = \begin{cases} \textit{rhet\_rel}(\text{CONTRAST}, A_8, C_8) \oplus \textit{rhet\_rel}(\text{CONTRAST}, A_8, D_8) \oplus \\ \quad \textit{rhet\_rel}(\text{CONTRAST}, B_8, C_8) \oplus \textit{rhet\_rel}(\text{CONTRAST}, B_8, D_8) \\ \textit{rhet\_rel}(\text{ELABORATION}, B_8, A_8) \\ \textit{rhet\_rel}(\text{ELABORATION}, D_8, A_8) \oplus \textit{rhet\_rel}(\text{ELABORATION}, D_8, B_8) \oplus \\ \quad \textit{rhet\_rel}(\text{ELABORATION}, D_8, C_8) \end{cases} \tag{2.11}$$

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 usages, etc. only suggest that some rhetorical relations hold between some textual units; rarely can hypotheses be made with 100% confidence.

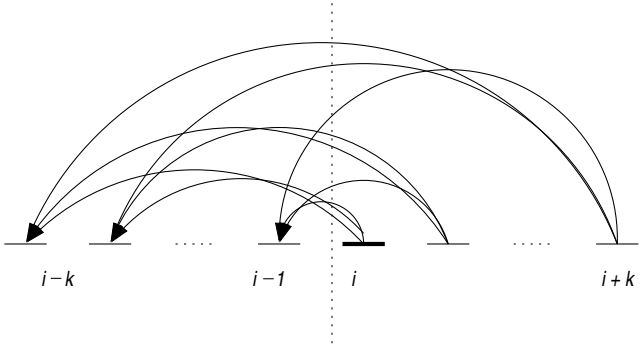

Figure 2.11
A graphical representation of the disjunctive hypothesis that is triggered by the occurrence of the marker *But* at the beginning of unit *i* of a text

When a computer program processes free texts and comes across a connective such as *But* at the beginning of a sentence, 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 one knows for sure though is that *But*, the *hypothesis trigger* in this example, can signal *at most one* such relation—in my empirical work (see Section 6.2), I have never come across a case in which a 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 2.11 provides a graphical representation of the simple rhetorical relations that can be hypothesized on the basis of the connective *however* in unit i.

In this book, I will focus on formalizing and exploiting only this sort of exclusively disjunctive hypotheses, i.e., hypotheses whose disjuncts subsume text spans that overlap. For example, in Figure 2.11, 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 hypotheses different from that

shown in Figure 2.11.[2] Definition 2.1 spells out the constraints we put on the exclusively disjunctive relations that are accommodated by the formalization and algorithms proposed in this book.

DEFINITION 2.1 *An exclusively disjunctive hypothesis of rhetorical relations is* well formed *if all textual spans that have as boundaries the units found in each disjunct overlap.*

According to definition 2.1, hypothesis 2.12 is well formed because spans [2, 4], [3, 4], and [3, 5] overlap, while hypothesis 2.13 is ill formed because spans [2, 3] and [4, 5] do not overlap. In both examples, hypotheses were defined over the sequence of elementary discourse units 2, 3, 4, 5.

$$rhet\_rel(\text{R}, 2, 4) \oplus rhet\_rel(\text{R}, 3, 4) \oplus rhet\_rel(\text{R}, 2, 4) \oplus rhet\_rel(\text{R}, 3, 5) \tag{2.12}$$

$$rhet\_rel(\text{R}, 2, 3) \oplus rhet\_rel(\text{R}, 4, 5) \tag{2.13}$$

In conclusion, the axiomatization proposed in this book can be used in the context of rhetorically parsing free, unrestricted texts because it assumes that rhetorical hypotheses can be ambiguous. Formally, the problem that I want to solve is that given in definition 2.2.

DEFINITION 2.2 ***The problem of text structure derivation:*** *Given a sequence of textual units $U = u_1, u_2, \ldots, u_n$ and a set RR of simple, extended, and well-formed exclusively disjunctive rhetorical relations that hold among these units and among contiguous textual spans that are defined over U, find all valid text structures of the linear sequence U.*

2. See Chapter 6 for details.

3 The Mathematics of Text Structures

3.1 A Model-Theoretic Account of Valid Text Structures

3.1.1 Introduction

The formalization of text structures that I propose assumes a set *Rels* of well-defined rhetorical relations that is partitioned into two subsets: the set of paratactic and the set of hypotactic relations ($Rels = Rels_{paratactic} \cup Rels_{hypotactic}$). Throughout the book, I will also use the terms "multinuclear" to refer to paratactic relations and "mononuclear" to refer to hypotactic relations.

I take the essential features of text structures given in Section 2.1.1, the strong compositionality criterion given in proposition 2.2, and the need to accommodate extended and exclusively disjunctive rhetorical relations to be the foundations of my formal treatment of text structures. More specifically, I will formalize the idea that two adjacent spans can be joined in a larger span by a given rhetorical relation only if one of the following holds:

- A similar relation holds also between at least two of the most salient units of those spans.
- An extended rhetorical relation holds between the two adjacent spans.

Obviously, the formalization will also specify the rules according to which the most salient units of text spans are determined.

In the rest of this section, I provide a formalization of the mathematical properties of valid text structures in the context of the text structure derivation problem 2.2. Subformalizations that disallow the use of extended rhetorical relations or exclusively disjunctive relations can be obtained by simply ignoring the axioms that pertain to these types of relations.

NOTATION. The formalization that I propose here uses the following predicates, with the following intended semantics:

- Predicate $position(u_i, j)$ is true for a textual unit u_i in sequence U if and only if u_i is the j-th element in the sequence.
- Predicate $rhet\_rel(name, u_i, u_j)$ is true for textual units u_i and u_j with respect to rhetorical relation *name* if and only if the definition D of rhetorical relation *name* is consistent with the relation between textual units u_i, in most cases a satellite, and u_j, a nucleus. The definition D could be part of any consistent theory of rhetorical relations. For example, from the perspective of RST, text 2.4 is completely described at the minimal unit level by the following set of predicates, in which the set of predicates *rhet_rel* is the same as that given in 2.5:

$$\begin{cases} \textit{rhet\_rel}(\text{JUSTIFICATION}_0, \text{A}_1, \text{B}_1) \\ \textit{rhet\_rel}(\text{JUSTIFICATION}_1, \text{D}_1, \text{B}_1) \\ \textit{rhet\_rel}(\text{EVIDENCE}, \text{C}_1, \text{B}_1) \\ \textit{rhet\_rel}(\text{CONCESSION}, \text{D}_1, \text{C}_1) \\ \textit{rhet\_rel}(\text{RESTATEMENT}, \text{D}_1, \text{A}_1) \\ \textit{position}(\text{A}_1, 1), \textit{position}(\text{B}_1, 2) \\ \textit{position}(\text{C}_1, 3), \textit{position}(\text{D}_1, 4) \end{cases} \tag{3.1}$$

• Predicate $\textit{rhet\_rel\_ext}(\textit{name}, s_s, s_e, n_s, n_e)$ is true for textual spans $[s_s, s_e]$ and $[n_s, n_e]$ with respect to rhetorical relation *name* if and only if the definition D of rhetorical relation *name* is consistent with the relation between the textual spans that ranges over units s_s–s_e, in most cases a satellite, and units n_s–n_e, a nucleus. Hence the five arguments of the predicate *rhet_rel_ext* denote the name of the rhetorical relation; the name of the elementary unit that is on the leftmost position in the satellite span, s_s; the name of the elementary unit that is on the rightmost position in the satellite span, s_e; the name of the elementary unit that is on the leftmost position in the nucleus span, n_s; and the name of the elementary unit that is on the rightmost position in the nucleus span, n_e. For example, from the perspective of RST, we can say that extended rhetorical relation $\textit{rhet\_rel\_ext}(\text{JUSTIFICATION}_2, \text{A}_1, \text{A}_1, \text{B}_1, \text{D}_1)$ holds between unit A_1 and span $[\text{B}_1, \text{D}_1]$.

In this book, I will also use the notation $\textit{rhet\_rel}(\textit{name}, [s_s, s_e], [n_s, n_e])$ as an abbreviation of $\textit{rhet\_rel\_ext}(\textit{name}, s_s, s_e, n_s, n_e)$ in the case $s_s \neq s_e$ and $n_s \neq n_e$, and $\textit{rhet\_rel}(\textit{name}, s_s, [n_s, n_e])$ as an abbreviation of $\textit{rhet\_rel\_ext}(\textit{name}, s_s, s_s, n_s, n_e)$ in the case the satellite is elementary ($s_s = s_e$). When the nucleus is elementary, I will use the notation $\textit{rhet\_rel}(\textit{name}, [s_s, s_e], n_s)$ as an abbreviation of $\textit{rhet\_rel\_ext}(\textit{name}, s_s, s_s, n_s, n_s)$. For example, $\textit{rhet\_rel}(\text{JUSTIFICATION}_2, \text{A}_1, [\text{B}_1, \text{D}_1])$ is nothing but a more intuitive representation of the predicate $\textit{rhet\_rel\_ext}(\text{JUSTIFICATION}_2, \text{A}_1, \text{A}_1, \text{B}_1, \text{D}_1)$ while $\textit{rhet\_rel}(\text{JUSTIFICATION}_3, [\text{C}_1, \text{D}_1], [\text{A}_1, \text{B}_1])$ is a more intuitive representation of the predicate $\textit{rhet\_rel\_ext}(\text{JUSTIFICATION}_3, \text{C}_1, \text{D}_1, \text{A}_1, \text{B}_1)$.

FEATURES OF THE FORMALIZATION. To simplify my formalization, I assume without restricting the generality of the problem that text trees are binary trees—the same assumption is employed in a number of other recent approaches to discourse processing [Polanyi and van den Berg, 1996, van den Berg, 1996, Gardent, 1997, Cristea and Webber, 1997, Schilder, 1997, Cristea et al., 1998]. A binary representation for a text tree maps each textual unit into a leaf and each rhetorical relation into an internal node whose children are the units between which that rhetorical relation holds. The mapping preserves the labeling associated with the nuclear status of each node. For example, a binary representation of the RS-tree in Figure 2.4a is given in Figure 3.1.

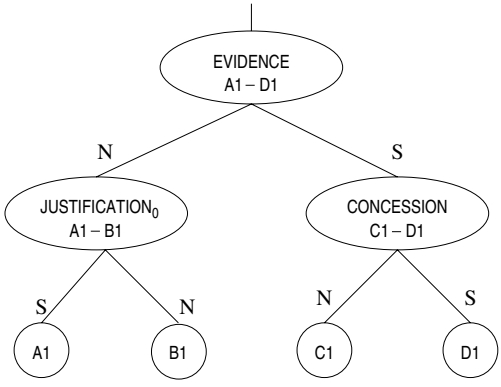

Figure 3.1
A binary representation isomorphic to the RS-tree shown in Figure 2.4a

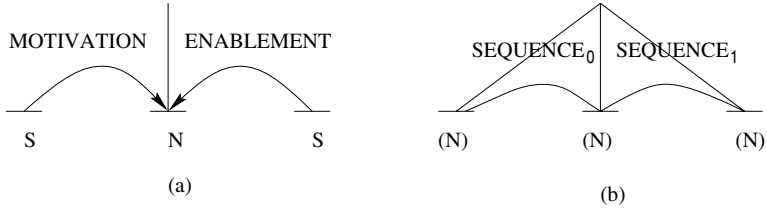

Figure 3.2
Examples of nonbinary discourse trees

In fact, we can interpret non-binary trees, such as those shown in Figure 3.2, as being collapsed versions of binary trees. For example, the tree in Figure 3.2a can be derived either from the tree in Figure 3.3a or that in 3.3b; and the tree in Figure 3.2b can be derived from the tree in Figure 3.3c. This view is also sympathetic with functional theories of language [Halliday, 1994] that stipulate that "rhetorical units defined by an enhancing nucleus-satellite relation have only one satellite. This satellite may be realized by a list (joint) of rhetorical units, but is still a single satellite" [Matthiessen and Thompson, 1988, p. 303].

Although the use of binary trees simplifies the formalization, it also has one disadvantage: binary representations, such as those in Figure 3.3a,b, explicitly induce an extra level of embedding on elementary units. In the original tree in Figure 3.2a, the satellites of the MOTIVATION and ENABLEMENT relations are equal with respect to their contribution to the

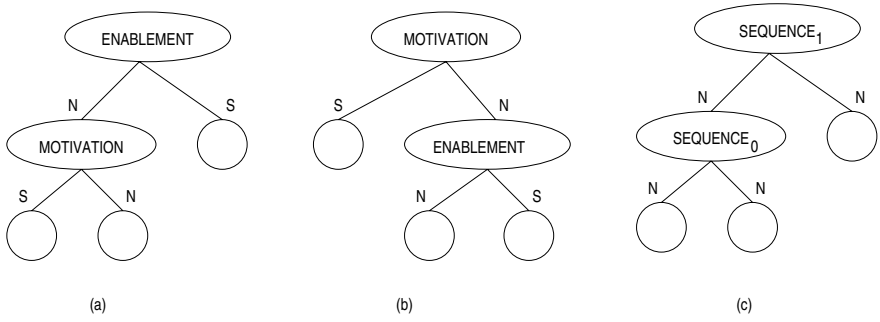

Figure 3.3
Binary trees equivalent with the nonbinary trees shown in Figure 3.2a,b

common nucleus and the structure has only one level of embedding. However, the binary representations in Figure 3.3a,b have two levels of embedding, which induce an ordering on the prominences of the two satellites. Similarly, if a multinuclear relation of LIST holds across seven units of a text, for example, a nonbinary representation will explicitly represent the fact that the seven units are "equal" with respect to their contribution to the span that subsumes them; the corresponding tree will have only one level of embedding. In contrast, the binary representation can be characterized by as many as six levels of embedding, depending on how the nuclei are grouped in the representation. In Chapter 7, I discuss an alternative approach to that discussed here; in the alternative approach, discourse structures are represented as nonbinary trees.

The formalization that I propose here is built on the following features:

- A text tree is a binary tree whose leaves denote elementary textual units.
- Each node has associated a *status* (nucleus or satellite), a *type* (the rhetorical relation that holds between the text spans that that node spans over), and a *salience* or *promotion set* (the set of units that constitute the most "important" part of the text that is spanned by that node). By convention, for each leaf node, the type is LEAF and the promotion set is the textual unit that it corresponds to.

A representation for the tree in Figure 2.4a, which reflects these characteristics, is given in Figure 3.4. The status, the type, and the salience unit that are associated with each leaf follow directly from the convention that I have given above. The status and the type of each internal node are a one-to-one mapping of the status and the rhetorical relation that are associated with each non-minimal text span from the original representation, respectively. The status of the root as {NUCLEUS, SATELLITE} reflects the fact that text span $[A_1, D_1]$ could play either a NUCLEUS or a SATELLITE role in any larger span that contains it.

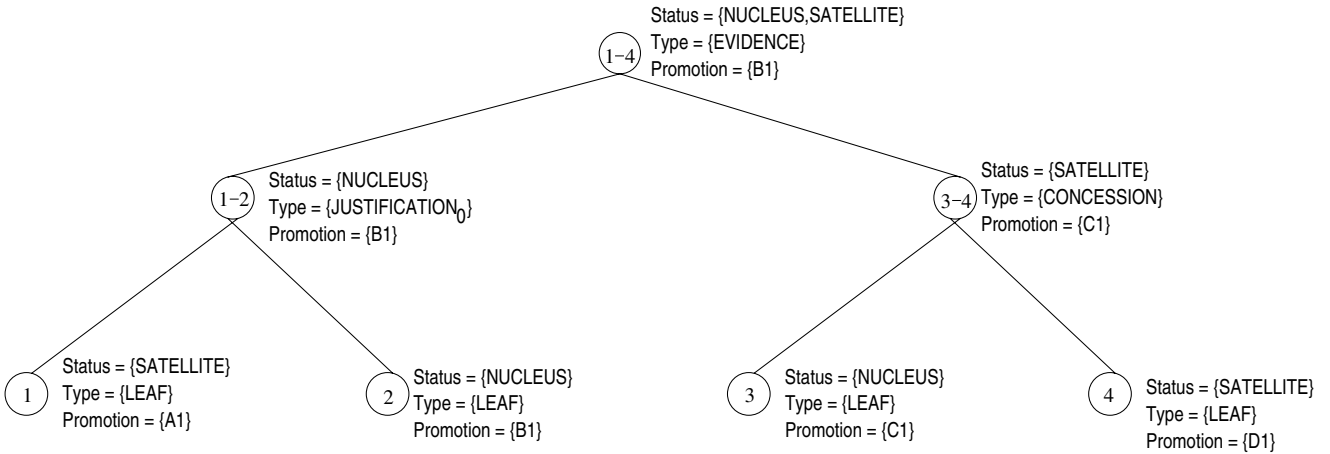

Figure 3.4
An isomorphic representation of tree in Figure 2.4a according to the status, type, and promotion features that characterize every node. The numbers associated with each node denote the limits of the text span that that node characterizes.

The most significant differences between the tree in Figure 3.4 and the tree in Figure 2.4a pertain to the promotion sets that are associated with every internal node. These promotion sets play a major role in determining the validity of a text tree. The tree in Figure 3.4 is valid, because the EVIDENCE relation that holds between spans [C_1, D_1] and [A_1, B_1] also holds between their most salient units, i.e., C_1 and B_1.

The status, the type, and the promotion set that are associated with each node in a text tree provide sufficient information for a full description of an instance of a text structure. Given the linear nature of text and the fact that we cannot predict in advance where the boundaries between various text spans will be drawn, we should provide a methodology that permits one to enumerate all possible ways in which a tree could be built on the top of a linear sequence of textual units. The solution that I propose relies on the same intuition that constitutes the foundation of chart parsing: just as a chart parser is capable of considering all possible ways in which different words in a text could be clustered into higher-order grammatical units, so my formalization would be capable of considering all the possible ways in which different text spans could be joined into larger spans.[1]

Let $span_{i,j}$, or simply $[i, j]$, denote a text span that includes all the textual units between positions i and j. Then, if we consider a sequence of textual units $u_1, u_2, \ldots, u_n$, there are n ways in which spans of length one could be built, $span_{1,1}, span_{2,2}, \ldots, span_{n,n}$; $n-1$ ways in which spans of length two could be built, $span_{1,2}, span_{2,3}, \ldots, span_{n-1,n}$; $n-2$ ways in which spans of length three could be built, $span_{1,3}, span_{2,4}, \ldots, span_{n-2,n}$; . . .; and one way in which a span of length n could be built, $span_{1,n}$. Since it is impossible to determine a priori the text spans that will be used to make up a text tree, I will associate with each text span that could possibly become part of a text tree a status, a type, and a promotion relation and let the constraints that pertain to the essential features of text structures and the strong compositionality criterion generate the valid text trees. In fact, my intent is to determine, from the set of $n + (n-1) + (n-2) + \cdots + 1 = n(n+1)/2$ potential text spans that pertain to a sequence of n textual units, the subset that adheres to the constraints that I have mentioned above. For example, for text 2.4, there are $4 + 3 + 2 + 1 = 10$ potential spans, i.e., $span_{1,1}$, $span_{2,2}$, $span_{3,3}$, $span_{4,4}$, $span_{1,2}$, $span_{2,3}$, $span_{3,4}$, $span_{1,3}$, $span_{2,4}$, and $span_{1,4}$, but only seven of them play an active role in the representation given in Figure 3.4, i.e., $span_{1,1}$, $span_{2,2}$, $span_{3,3}$, $span_{4,4}$, $span_{1,2}$, $span_{3,4}$, and $span_{1,4}$.

In formalizing the constraints that pertain to a text tree, I assume that each possible text span, $span_{l,h}$,[2] which will or will not eventually become a node in the final discourse tree, is characterized by the following relations:

1. I am grateful to Jeff Siskind for bringing to my attention the similarity between charts and text spans.

2. In what follows, l and h always denote the left and right boundaries of a text span.

• $S(l, h, status)$ provides the status of $span_{l,h}$, i.e., the text span that contains units l to h; *status* can take one of the values NUCLEUS, SATELLITE, or NONE according to the role played by that span in the final text tree. For example, for the RS-tree depicted in Figure 3.4, the following relations hold: $S(1, 2, \text{NUCLEUS})$, $S(3, 4, \text{SATELLITE})$, $S(1, 3, \text{NONE})$.

• $T(l, h, relation\_name)$ provides the name of the rhetorical relation that holds between the text spans that are immediate subordinates of $span_{l,h}$ in the text tree.[3] If the text span is not used in the construction of the final text tree, the type assigned by convention is NONE. For example, for the RS-tree in Figure 3.4, the following relations hold: $T(1, 1, \text{LEAF})$, $T(1, 2, \text{JUSTIFICATION}_0)$, $T(3, 4, \text{CONCESSION})$, $T(1, 3, \text{NONE})$.

• $P(l, h, unit\_name)$ provides one of the units of the set of units that are salient for $span_{l,h}$ and that can be used to connect this text span with adjacent text spans in the final RS-tree. The set of relations $P(l, h, unit\_name)$ that hold with respect to a span $[l, h]$ gives the promotion set of that span. If $span_{l,h}$ is not used in the final text tree, by convention, the set of salient units is NONE. For example, for the RS-tree in Figure 3.4, the following relations hold: $P(1, 1, \text{A}_1)$, $P(1, 2, \text{B}_1)$, $P(1, 3, \text{NONE})$, $P(3, 4, \text{C}_1)$.

3.1.2 A Complete Formalization of Text Trees

Using the conventions that I have discussed above, I present now a complete first-order formalization of the mathematical properties of valid text trees. In this formalization, I assume a universe that consists of the set of natural numbers from 1 to N, where N represents the number of textual units in the text that is considered; the set of names that were defined by a discourse theory for each rhetorical relation, indexed with a unique natural number identifier; the set of unit names that are associated with each textual unit; and four extra constants: NUCLEUS, SATELLITE, NONE, and LEAF. Unique name axioms for the constants are considered to be part of the formalization. The only function symbols that operate over this domain are the traditional $+$ and $-$ functions that are associated with the set of natural numbers. The formalization uses the traditional predicate symbols that pertain to the set of natural numbers ($<, \leq, >, \geq, =, \neq$) and six other predicate symbols: S, T, and P to account for the status, the type, and the salient units that are associated with every text span; *rhet_rel* to account for the rhetorical relations that hold between different textual units; *position* to account for the index of the textual units in the text that one considers; and *Exclusive* to account for exclusively disjunctive hypotheses. I use interchangeably the terms *text tree* or *discourse tree* whenever I refer to a general abstract structure, which is built using some taxonomy of relations $Rels = Rels_{hypotactic} \cup Rels_{paratactic}$. I use the term

3. The names of the rhetorical relations are dependent on the set of relations that one uses.

RS-tree whenever I refer to a text structure that uses the taxonomy of relations defined by Mann and Thompson [1988].

Throughout this book, I apply the convention that all unbound variables are universally quantified and that variables are represented in lowercase letters while constants appear in small capitals. I also make use of two extra relations (*relevant_rel* and *relevant_unit*), which I define here as follows: for every text span $span_{l,h}$, $relevant\_rel(l, h, name)$ 3.2 describes the set of simple, exclusively disjunctive, and extended rhetorical relations that are relevant to that text span, i.e., the set of rhetorical relations that span over units from the interval $[l, h]$. It is only these relations that can be used to label the type of a tree that subsumes all units in the interval $[l, h]$.

$$
\begin{aligned}
&relevant\_rel(l, h, name)\\
&\quad\equiv (\exists s, n, sp, np)[\\
&\qquad position(s, sp) \wedge position(n, np) \wedge (l \leq sp \leq h) \wedge (l \leq np \leq h)\\
&\qquad \wedge rhet\_rel(name, s, n)]\\
&\quad\vee (\exists s_1, n_1, sp_1, np_1, name_1 \ldots, s_k, n_k, sp_k, np_k, name_k)[\\
&\qquad position(s_1, sp_1) \wedge position(n_1, np_1)\\
&\qquad \wedge (l \leq sp_1 \leq h) \wedge (l \leq np_1 \leq h) \wedge name_1 = name\\
&\qquad \wedge rhet\_rel(name_1, s_1, n_1) \oplus \ldots \oplus rhet\_rel(name_k, s_k, n_k)]\\
&\quad\vee (\exists s_s, s_e, n_s, n_e, split\_point)[\\
&\qquad position(s_s, l) \wedge position(s_e, split\_point)\\
&\qquad \wedge position(n_s, split\_point + 1) \wedge position(n_e, h)\\
&\qquad \wedge (l \leq split\_point \leq h)\\
&\qquad \wedge (rhet\_rel\_ext(name, s_s, s_e, n_s, n_e) \vee rhet\_rel\_ext(name, n_s, n_e, s_s, s_e))]\\
&\quad\vee (\exists s_{s1}, s_{e1}, n_{s1}, n_{e1}, split\_point_1, name_1, \ldots, s_{sk}, s_{ek}, n_{sk}, n_{ek}, split\_point_k, name_k)[\\
&\qquad position(s_{s1}, l) \wedge position(s_{e1}, split\_point_1)\\
&\qquad \wedge position(n_{s1}, split\_point + 1) \wedge position(n_{e1}, h)\\
&\qquad \wedge (l \leq split\_point_1 \leq h) \wedge name = name_1\\
&\qquad \wedge (rhet\_rel\_ext(name, s_{s1}, s_{e1}, n_{s1}, n_{e1}) \oplus \cdots\\
&\qquad\qquad \oplus rhet\_rel\_ext(name, s_{sk}, s_{ek}, n_{sk}, n_{ek})\\
&\qquad\quad \vee rhet\_rel\_ext(name, n_{s1}, n_{e1}, s_{s1}, s_{e1}) \oplus \cdots\\
&\qquad\qquad \oplus rhet\_rel\_ext(name, n_{sk}, n_{ek}, s_{sk}, s_{ek}))]
\end{aligned}
\tag{3.2}
$$

The definition of *relevant_rel* employs one disjunct for each type of rhetorical hypothesis that the book focuses on. A relation *name* can be used to label the type of a tree that subsumes all units in the interval $[l, h]$ if and only if:

• A simple relation *name* holds between two units in the interval $[l, h]$ (the first disjunct in definition 3.2).

• A simple relation *name* of an exclusively disjunctive hypothesis holds between two units in the interval $[l, h]$ (the second disjunct in definition 3.2).

• An extended rhetorical relation *name* holds between two adjacent spans that subsume the whole span $[l, h]$ (the third disjunct in definition 3.2).

• An extended relation *name* of an exclusively disjunctive hypothesis holds between two adjacent spans that subsume the whole span $[l, h]$ (the fourth disjunct in definition 3.2).

For every text span $span_{l,h}$, $relevant\_unit(l, h, u)$ 3.3 describes the set of textual units that are relevant for that text span, i.e., the units whose positions in the initial sequence are numbers in the interval $[l, h]$. It is only these units that can be used to label the promotion set associated with a tree that subsumes all units in the interval $[l, h]$.

$$relevant\_unit(l, h, u) \equiv (\exists x)[position(u, x) \wedge (l \leq x \leq h)] \tag{3.3}$$

For example, for text 2.4, which is described formally in 3.1, the following is the set of all *relevant_rel* and *relevant_unit* relations that hold with respect to text segment [1, 3] and with respect to the relation definitions proposed by RST:

$\{relevant\_rel(1, 3, \text{JUSTIFICATION}_0), relevant\_rel(1, 3, \text{EVIDENCE}),$
$relevant\_unit(1, 3, \text{A}_1), relevant\_unit(1, 3, \text{B}_1), relevant\_unit(1, 3, \text{C}_1)\}$

For text 2.10, which is described formally in 2.11, the following is the set of all *relevant_rel* and *relevant_unit* relations that hold with respect to text segment [1, 3] and with respect to the relation definitions proposed by RST:

$\{relevant\_rel(1, 3, \text{CONTRAST}), relevant\_rel(1, 3, \text{ELABORATION}),$
$relevant\_unit(1, 3, \text{A}_8), relevant\_unit(1, 3, \text{B}_8), relevant\_unit(1, 3, \text{C}_8)\}$

The constraints that pertain to the structure of a text tree can be partitioned into constraints related to the objects over which each predicate ranges and constraints related to the structure of the tree. I describe each set of constraints in turn.

Constraints that Concern the Objects over which the Predicates that Describe Every Span $[l, h]$ of a Text Tree Range

• **For every span $[l, h]$, the set of objects over which predicate S ranges is the set** NUCLEUS, SATELLITE, NONE**.** Since every textual unit has to be part of the final RS-tree, the elementary text spans, i.e., those spans for which $l = h$, constitute an exception to this rule, i.e., they could play only a NUCLEUS or SATELLITE role.

$$[(1 \leq h \leq \text{N}) \wedge (1 \leq l \leq h)] \\ \rightarrow \big\{[l = h \rightarrow (S(l, h, \text{NUCLEUS}) \vee S(l, h, \text{SATELLITE}))] \\ \wedge [l \neq h \rightarrow (S(l, h, \text{NUCLEUS}) \vee S(l, h, \text{SATELLITE}) \vee S(l, h, \text{NONE}))]\big\} \tag{3.4}$$

• **The status of any text span is unique.**

$$[(1 \leq h \leq \text{N}) \wedge (1 \leq l \leq h)] \\ \rightarrow [(S(l, h, status_1) \wedge S(l, h, status_2)) \rightarrow status_1 = status_2] \tag{3.5}$$

• **For every span $[l, h]$, the set of objects over which predicate T ranges is the set of rhetorical relations that are relevant to that span.** By convention, the rhetorical relation associated with a leaf is LEAF.

$$[(1 \leq h \leq \text{N}) \wedge (1 \leq l \leq h)] \\ \rightarrow \big\{[l = h \rightarrow T(l, h, \text{LEAF})] \\ \wedge [l \neq h \rightarrow (T(l, h, \text{NONE}) \\ \vee (T(l, h, name) \rightarrow relevant\_rel(l, h, name)))]\big\} \tag{3.6}$$

As one can see, axiom 3.6 formalizes part of the strong compositionality criterion because it restricts the set of rhetorical relations that can characterize an internal node to the set of relations that span over units in the interval $[l, h]$. Axiom 3.12, which is discussed below, will formalize the remaining part of the strong compositionality criterion, i.e., it will require these units to be salient.

• **At most one rhetorical relation can connect two adjacent text spans.**

$$[(1 \leq h \leq \text{N}) \wedge (1 \leq l < h)] \\ \rightarrow [(T(l, h, name_1) \wedge T(l, h, name_2)) \rightarrow name_1 = name_2] \tag{3.7}$$

• **For every span $[l, h]$, the set of objects over which predicate P ranges is the set of units that make up that span.**

$$[(1 \leq h \leq \text{N}) \wedge (1 \leq l \leq h)] \\ \rightarrow [P(l, h, \text{NONE}) \vee (P(l, h, u) \rightarrow relevant\_unit(l, h, u))] \tag{3.8}$$

Constraints that Concern the Structure of the Text Trees The following constraints are derived from the essential features of text structures that were discussed in Section 2.1.1 and from the strong compositionality criterion given in proposition 2.2.

• **Text spans do not overlap.**

$$
\begin{aligned}
&[(1 \le h_1 \le \text{N}) \wedge (1 \le l_1 \le h_1) \wedge (1 \le h_2 \le \text{N}) \wedge (1 \le l_2 \le h_2) \\
&\quad \wedge (l_1 < l_2) \wedge (h_1 < h_2) \wedge (l_2 \le h_1)] \\
&\quad \rightarrow [\neg S(l_1, h_1, \text{NONE}) \rightarrow S(l_2, h_2, \text{NONE})]
\end{aligned} \tag{3.9}
$$

• **A text span with status** NONE **does not participate in the tree at all.**

$$
\begin{aligned}
&[(1 \le h \le \text{N}) \wedge (1 \le l < h)] \\
&\quad \rightarrow [(S(l, h, \text{NONE}) \wedge P(l, h, \text{NONE}) \wedge T(l, h, \text{NONE})) \\
&\qquad \vee (\neg S(l, h, \text{NONE}) \wedge \neg P(l, h, \text{NONE}) \wedge \neg T(l, h, \text{NONE}))]
\end{aligned} \tag{3.10}
$$

• **There exists a text span, the root, that spans over the entire text.**

$$
\neg S(1, \text{N}, \text{NONE}) \wedge \neg P(1, \text{N}, \text{NONE}) \wedge \neg T(1, \text{N}, \text{NONE}) \tag{3.11}
$$

• **The status, the type, and the promotion set that are associated with a text span reflect the strong compositionality criterion.**

$$
\begin{aligned}
&[(1 \le h \le \text{N}) \wedge (1 \le l < h) \wedge \neg S(l, h, \text{NONE})] \\
&\quad \rightarrow (\exists name, split\_point, s, n)[(l \le split\_point \le h) \\
&\qquad \wedge (Nucleus\_first(name, split\_point, s, n) \\
&\qquad\quad \vee Satellite\_first(name, split\_point, s, n))] \\
&\qquad \vee (\exists name, split\_point, s_s, s_e, n_s, n_e)[(l \le split\_point \le h) \\
&\qquad \wedge (Nucleus\_first\_ext(name, split\_point, s_s, s_e, n_s, n_e) \\
&\qquad\quad \vee Satellite\_first\_ext(name, split\_point, s_s, s_e, n_s, n_e))]
\end{aligned} \tag{3.12}
$$

$$
\begin{aligned}
&Nucleus\_first(name, split\_point, s, n) \\
&\quad \equiv rhet\_rel(name, s, n) \wedge T(l, h, name) \\
&\qquad \wedge position(s, sp) \wedge position(n, np) \\
&\qquad \wedge l \le np \le split\_point \wedge split\_point < sp \le h \\
&\qquad \wedge P(l, split\_point, n) \wedge P(split\_point + 1, h, s) \\
&\qquad \wedge \{(name \in Rels_{paratactic}) \\
&\qquad\quad \rightarrow S(l, split\_point, \text{NUCLEUS}) \wedge S(split\_point + 1, h, \text{NUCLEUS}) \\
&\qquad\qquad \wedge (\forall p)[P(l, h, p) \equiv (P(l, split\_point, p) \vee P(split\_point + 1, h, p))]\} \\
&\qquad \wedge \{(name \in Rels_{hypotactic}) \\
&\qquad\quad \rightarrow S(l, split\_point, \text{NUCLEUS}) \wedge S(split\_point + 1, h, \text{SATELLITE}) \\
&\qquad\qquad \wedge (\forall p)(P(l, h, p) \equiv P(l, split\_point, p))\}
\end{aligned} \tag{3.13}
$$

$$
\begin{aligned}
&\mathit{Satellite\_first}(\mathit{name}, \mathit{split\_point}, s, n)\\
&\quad \equiv \mathit{rhet\_rel}(\mathit{name}, s, n) \wedge T(l, h, \mathit{name})\\
&\qquad \wedge \mathit{position}(s, sp) \wedge \mathit{position}(n, np)\\
&\qquad \wedge l \leq sp \leq \mathit{split\_point} \wedge \mathit{split\_point} < np \leq h\\
&\qquad \wedge P(l, \mathit{split\_point}, s) \wedge P(\mathit{split\_point}+1, h, n)\\
&\qquad \wedge \{(\mathit{name} \in \mathit{Rels}_{\mathit{paratactic}})\\
&\qquad\quad \rightarrow S(l, \mathit{split\_point}, \textsc{nucleus}) \wedge S(\mathit{split\_point}+1, h, \textsc{nucleus})\\
&\qquad\qquad \wedge (\forall p)[P(l, h, p) \equiv (P(l, \mathit{split\_point}, p) \vee P(\mathit{split\_point}+1, h, p))]\}\\
&\qquad \wedge \{(\mathit{name} \in \mathit{Rels}_{\mathit{hypotactic}})\\
&\qquad\quad \rightarrow S(l, \mathit{split\_point}, \textsc{satellite}) \wedge S(\mathit{split\_point}+1, h, \textsc{nucleus})\\
&\qquad\qquad \wedge (\forall p)(P(l, h, p) \equiv P(\mathit{split\_point}+1, h, p))\}
\end{aligned}
\tag{3.14}
$$

$$
\begin{aligned}
&\mathit{Nucleus\_first\_ext}(\mathit{name}, \mathit{split\_point}, s_s, s_e, n_s, n_e)\\
&\quad \equiv \{\mathit{rhet\_rel\_ext}(\mathit{name}, s_s, s_e, n_s, n_e) \wedge T(l, h, \mathit{name})\\
&\qquad \wedge \mathit{position}(s_s, \mathit{split\_point}+1) \wedge \mathit{position}(s_e, h)\\
&\qquad \wedge \mathit{position}(n_s, l) \wedge \mathit{position}(n_e, \mathit{split\_point})\\
&\qquad \wedge \{(\mathit{name} \in \mathit{Rels}_{\mathit{paratactic}})\\
&\qquad\quad \rightarrow S(l, \mathit{split\_point}, \textsc{nucleus}) \wedge S(\mathit{split\_point}+1, h, \textsc{nucleus})\\
&\qquad\qquad \wedge (\forall p)[P(l, h, p) \equiv (P(l, \mathit{split\_point}, p) \vee P(\mathit{split\_point}+1, h, p))]\}\\
&\qquad \wedge \{(\mathit{name} \in \mathit{Rels}_{\mathit{hypotactic}})\\
&\qquad\quad \rightarrow S(l, \mathit{split\_point}, \textsc{nucleus}) \wedge S(\mathit{split\_point}+1, h, \textsc{satellite})\\
&\qquad\qquad \wedge (\forall p)(P(l, h, p) \equiv P(l, \mathit{split\_point}, p))\}
\end{aligned}
\tag{3.15}
$$

$$
\begin{aligned}
&\mathit{Satellite\_first\_ext}(\mathit{name}, \mathit{split\_point}, s_s, s_e, n_s, n_e)\\
&\quad \equiv \{\mathit{rhet\_rel\_ext}(\mathit{name}, s_s, s_e, n_s, n_e) \wedge T(l, h, \mathit{name})\\
&\qquad \wedge \mathit{position}(n_s, \mathit{split\_point}+1) \wedge \mathit{position}(n_e, h)\\
&\qquad \wedge \mathit{position}(s_s, l) \wedge \mathit{position}(s_e, \mathit{split\_point})\\
&\qquad \wedge \{(\mathit{name} \in \mathit{Rels}_{\mathit{paratactic}})\\
&\qquad\quad \rightarrow S(l, \mathit{split\_point}, \textsc{nucleus}) \wedge S(\mathit{split\_point}+1, h, \textsc{nucleus})\\
&\qquad\qquad \wedge (\forall p)[P(l, h, p) \equiv (P(l, \mathit{split\_point}, p) \vee P(\mathit{split\_point}+1, h, p))]\}\\
&\qquad \wedge \{(\mathit{name} \in \mathit{Rels}_{\mathit{hypotactic}})\\
&\qquad\quad \rightarrow S(l, \mathit{split\_point}, \textsc{satellite}) \wedge S(\mathit{split\_point}+1, h, \textsc{nucleus})\\
&\qquad\qquad \wedge (\forall p)(P(l, h, p) \equiv P(\mathit{split\_point}+1, h, p))\}
\end{aligned}
\tag{3.16}
$$

Formula 3.12 specifies that whenever a text span $[l, h]$ denotes an internal node ($l < h$) in the final text tree, i.e., its status is not NONE, the span $[l, h]$ is built on the top of two text spans that meet at index *split_point* and there either exists an elementary relation that holds between two units that are salient in the adjacent spans (*Nucleus_first* $\vee$ *Satellite_first*) or an extended rhetorical relation that holds between the two spans (*Nucleus_first_ext* $\vee$ *Satellite_first_ext*).

Formula 3.13 specifies that there is a rhetorical relation with name *name*, from a unit s (in most cases a satellite) that belongs to span [*split_point* + 1, h] to a unit n, the nucleus, that belongs to span [l, *split_point*]; that unit n is salient with respect to text span [l, *split_point*] and unit s is salient with respect to text span [*split_point* + 1, h]; and that the type of span $[l, h]$ is given by the name of the rhetorical relation. If the relation is paratactic (multinuclear), the status of the immediate subspans is NUCLEUS and the set of salient units for text span $[l, h]$ consists of all the units that make up the set of salient units that are associated with the two subspans. If the relation is hypotactic, the status of text span [l, *split_point*] is NUCLEUS, the status of text span [*split_point* + 1, h] is SATELLITE and the set of salient units for text span $[l, h]$ is given by the salient units that are associated with the subordinate nucleus span. The $\in$ symbol in formulas 3.13 and 3.16 is just an abbreviation of a disjunction over all the relation names that belong to the paratactic and hypotactic partitions respectively. Formula *Satellite_first*(*name*, *split_point*, s, n), 3.14, is a mirror image of 3.13 and it describes the case when the satellite that pertains to rhetorical relation *rhet_rel*(*name*, s, n) belongs to text span [l, *split_point*], i.e., when the satellite goes before the nucleus.

Formula 3.15 specifies that there is an extended rhetorical relation with name *name*, which holds between two textual spans that meet at *split_point*, and that the nucleus of the rhetorical relation goes before the satellite. In such a case, the type of span $[l, h]$ is given by the name of the extended rhetorical relation. If the relation is paratactic (multinuclear), the status of the immediate subspans is NUCLEUS and the set of salient units for text span $[l, h]$ consists of all the units that make up the set of salient units that are associated with the two subspans. If the relation is hypotactic, the status of text span [l, *split_point*] is NUCLEUS, the status of text span [*split_point* + 1, h] is SATELLITE and the set of salient units for text span $[l, h]$ is given by the salient units that are associated with the subordinate nucleus span. Formula 3.16 is a mirror image of 3.15 and it describes the case when the satellite goes before the nucleus. That is, the case when the satellite span s_s–s_e belongs to text span [l, *split_point*].

For the rest of the book, the set of axioms 3.2–3.16 will be referred to as *the axiomatization of valid text structures*.

3.1.3 A Formalization of RST

The axiomatization of valid text structures given in Section 3.1.2 can be tailored to any set of relations. If we choose to work with the set of rhetorical relations proposed by Mann and Thompson [1988], the only thing that we need to do is specify what the hypotactic and paratactic relations are. We can do this explicitly, by instantiating in axioms 3.13, 3.14, 3.15, and 3.16 the sets of hypotactic and paratactic relations that are proposed in RST. For example, axiom 3.17 is the RST instantiation of axiom 3.13.

$$
\begin{aligned}
&\mathit{Nucleus\_first}(\mathit{name}, \mathit{split\_point}, s, n)\\
&\quad \equiv \mathit{rhet\_rel}(\mathit{name}, s, n) \wedge T(l, h, \mathit{name})\\
&\qquad \wedge \mathit{position}(s, sp) \wedge \mathit{position}(n, np)\\
&\qquad \wedge l \leq np \leq \mathit{split\_point} \wedge \mathit{split\_point} < sp \leq h\\
&\qquad \wedge P(l, \mathit{split\_point}, n) \wedge P(\mathit{split\_point}+1, h, s)\\
&\qquad \wedge \{(\mathit{name} = \textsc{contrast} \vee \mathit{name} = \textsc{joint} \vee \mathit{name} = \textsc{sequence}) \qquad (3.17)\\
&\qquad\quad \rightarrow S(l, \mathit{split\_point}, \textsc{nucleus}) \wedge S(\mathit{split\_point}+1, h, \textsc{nucleus})\\
&\qquad\qquad \wedge (\forall p)[P(l, h, p) \equiv (P(l, \mathit{split\_point}, p) \vee P(\mathit{split\_point}+1, h, p))]\}\\
&\qquad \wedge \{(\mathit{name} \neq \textsc{sequence} \wedge \mathit{name} \neq \textsc{contrast} \wedge \mathit{name} \neq \textsc{joint})\\
&\qquad\quad \rightarrow S(l, \mathit{split\_point}, \textsc{nucleus}) \wedge S(\mathit{split\_point}+1, h, \textsc{satellite})\\
&\qquad\qquad \wedge (\forall p)(P(l, h, p) \equiv P(l, \mathit{split\_point}, p))\}
\end{aligned}
$$

In a similar manner, we can instantiate axioms 3.14, 3.15, and 3.16 as well. For the rest of the book, axioms 3.2–3.12 and the set of axioms that are derived from axioms 3.13–3.16 by instantiating the set of relations proposed by RST will be referred to as *the axiomatization of RST*.

If we evaluate now the RS-trees in Figure 2.4 against the axiomatization of RST, we can determine immediately that the structures of the trees in Figure 2.4a and 2.4c satisfy all the axioms, while the structure of the tree in Figure 2.4b does not satisfy axiom 3.12. More precisely, the rhetorical relation of CONCESSION between D_1 and C_1 projects C_1 as the salient unit for text span $[\text{C}_1, \text{D}_1]$. The initial set of rhetorical relations 3.1 depicts a JUSTIFICATION relation only between units D_1 and B_1 and not between C_1 and B_1. Since the nuclearity requirements make it impossible for D_1 both to play a satellite role in the span $[\text{C}_1, \text{D}_1]$, and to be, at the same time, a salient unit for it, it follows that tree 2.4b is incorrect.

3.2 A Proof-Theoretic Account of Valid Text Structures

In Section 3.1, I have discussed only what discourse structures are. The formalization provides a mathematical description of the valid text structures, i.e., an expression of

the properties of the class of structures that are licensed by the essential features that were put forth in Section 2.1.1 and by the strong compositionality criterion 2.2. In the rest of this chapter, I will discuss how discourse structures can be derived. In doing so, I start with a proof-theoretic account of the problem of text structure derivation defined in 2.2.

The proof-theoretic account 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 status either 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 built 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.

The proof theory is defined on a universe Ω that consists of the set of natural numbers from 1 to N, the set of unit names $U_N = \{u_1, u_2, \ldots, u_N\}$ associated with the elementary units in a text, the set of constants NUCLEUS, SATELLITE, LEAF, NULL, and the names of all rhetorical relations in the taxonomy under consideration. The universe also contains objects of the form *tree*(*status*, *type*, *promotion*, *left*, *right*), where *status* can be either NUCLEUS or SATELLITE; *type* can be a name of a rhetorical relation; *promotion* can be a subset of elementary units from U_N; and *left* and *right* can be either NULL or recursively defined objects of type *tree*. Sets of simple, extended, and exclusively disjunctive rhetorical relations are considered legal objects as well. We assume that the language defined over the universe Ω supports the traditional function symbols $+$ and $-$ and operations that are typical to sets.

The objects having the form *tree*(*status*, *type*, *promotion*, *left*, *right*) provide a functional representation of valid text structures. Assume, for example, that a rhetorical relation *rhet_rel*(JUSTIFICATION, A_1, B_1) holds between the units of a text with two elementary units. Then, the valid tree structure shown in Figure 3.5a can be represented using an object of type *tree* as shown in 3.18. Although the objects of type *tree* can represent valid text structures, their syntax does not impose sufficient constraints on the semantics of the structures that they correspond to. For example, the structure shown in Figure 3.5b can also be represented as an object of type *tree*, as shown in 3.19, but, obviously, it is not a valid text structure: the JUSTIFICATION relation is hypotactic, so, assigning the status NUCLEUS to both elementary units is incorrect.

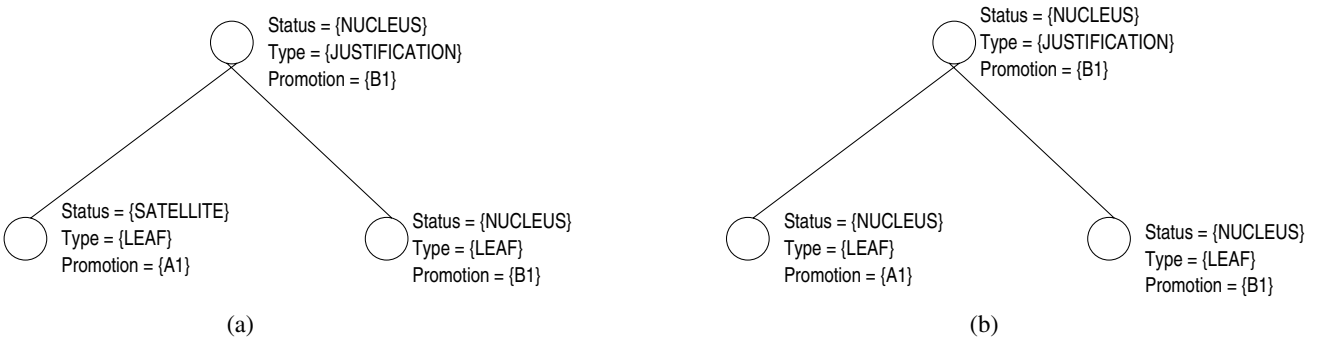

Figure 3.5
Examples of valid and invalid text structures

$$\begin{aligned}&\textit{tree}(\text{NUCLEUS, JUSTIFICATION, } \{\text{B}_1\},\\ &\quad \textit{tree}(\text{SATELLITE, LEAF, } \{\text{A}_1\}, \text{NULL, NULL}),\\ &\quad \textit{tree}(\text{NUCLEUS, LEAF, } \{\text{B}_1\}, \text{NULL, NULL}))\end{aligned} \tag{3.18}$$

$$\begin{aligned}&\textit{tree}(\text{NUCLEUS, JUSTIFICATION, } \{\text{B}_1\},\\ &\qquad \textit{tree}(\text{NUCLEUS, LEAF, } \{\text{A}_1\}, \text{NULL, NULL}),\\ &\qquad \textit{tree}(\text{NUCLEUS, LEAF, } \{\text{B}_1\}, \text{NULL, NULL}))\end{aligned} \tag{3.19}$$

The definition below makes explicit the correspondence between valid text structures and objects of type *tree*.

DEFINITION 3.1 *An object tree*(*status*, *type*, *promotion*, *left*, *right*) corresponds *to a valid text structure if and only if the status*, *type*, *and promotion arguments of the tree have the same values as those of the root of the text structure and if the left and right arguments* correspond *to the left and right subtrees of the valid text structure.*

The language that I describe here in conjunction with universe Ω accepts only five predicate symbols:

• Predicate *position*(u_i, j) is true if unit u_i is the jth unit in the sequence $U_\text{N} = u_1, u_2, \ldots, u_\text{N}$ that corresponds to the text under scrutiny. For example, for the text shown in 2.10, *position*(A_8, 1) and *position*(C_8, 3) are true. However, *position*(A_8, 2) is false.

• Predicate *hold*(*rr*) is true for a given text if and only if the rhetorical relations enumerated in the set *rr* hold among the units in that text. For example, for text 2.10, the predicate *hold*(*RR*) is true if *RR* contains the list of rhetorical relations shown in 2.11.

• Predicate $S(l, h, \textit{tree}(\ldots), R_{lh})$ is true when a valid text structure that corresponds to the argument *tree*(. . .) can be built on span $[l, h]$ using rhetorical relations that hold among units in the span. The argument R_{lh} denotes the set of rhetorical relations that can be used to extend the valid structure of span $[l, h]$, i.e., the rhetorical relations that hold among the units in the text that have not been used in the construction of the valid structure that corresponds to the object *tree*(. . .). For example, given text 2.10 and the set of rhetorical relations *RR* that hold among its units, 2.11, the predicate in 3.20 is true.

$$\begin{aligned}&S(1, 2, \textit{tree}(\text{NUCLEUS, ELABORATION, } \{\text{A}_8\},\\ &\qquad \textit{tree}(\text{NUCLEUS, LEAF, } \{\text{A}_8\}, \text{NULL, NULL}),\\ &\qquad \textit{tree}(\text{SATELLITE, LEAF, } \{\text{B}_8\}, \text{NULL, NULL})),\\ &\qquad\quad RR \setminus \{\textit{rhet\_rel}(\text{ELABORATION, } \text{B}_8, \text{A}_8)\})\end{aligned} \tag{3.20}$$

I say loosely that a predicate $S(l, h, tree(\ldots), R_{lh})$ *corresponds* to a valid text structure if its third argument corresponds to that structure.

• Predicate *hypotactic*(*name*) is true if *name* is a hypotactic relation in the taxonomy of rhetorical relations that is used. For example, if one uses RST, *hypotactic*(JUSTIFICATION) and *hypotactic*(CONCESSION) are both true.

• Predicate *paratactic*(*name*) is true if *name* is a paratactic relation in the taxonomy of rhetorical relations that is used. For example, if one uses RST, *paratactic*(CONTRAST) and *paratactic*(SEQUENCE) are both true.

I treat the following as axioms of a logical system that characterizes how text structures can be derived:

• Instantiations of schemata 3.21 and 3.22 with respect to the taxonomy of relations that is used.

hypotactic(*relation_name*) (3.21)

paratactic(*relation_name*) (3.22)

• Instantiations of schema 3.23 with respect to the rhetorical relations *RR* that hold among the units of the text under scrutiny.

hold(*RR*) (3.23)

• Instantiations of schema 3.24, with respect to each unit $u_i \in U_N$ in the text under scrutiny that occurs in position j.

position(u_i, j) (3.24)

I describe now a set of Horn-like axioms that characterize how textual structures that characterize textual spans can be joined to obtain textual structures for larger spans. For the limit case, I assume that for every textual unit u_i at position j in the initial sequence of textual units $u_1, \ldots, u_N$ there exists a textual span S that can be associated with a valid text structure that has status either NUCLEUS or SATELLITE, type LEAF, and promotion set $\{u_i\}$; any of the relations given in the initial set *RR* can be used to extend the span S into a larger one. A text of N units can therefore yield at most N axioms having the form 3.25 and N axioms having the form 3.26.

$$[position(u_i, j) \wedge hold(RR)]row \rightarrow S(j, j, tree(\text{NUCLEUS}, \text{LEAF}, \{u_i\}, \text{NULL}, \text{NULL}), RR) \quad (3.25)$$

$$[position(u_i, j) \wedge hold(RR)]row \rightarrow S(j, j, tree(\text{SATELLITE}, \text{LEAF}, \{u_i\}, \text{NULL}, \text{NULL}), RR) \quad (3.26)$$

The intuitive notion behind the use of the set *RR* of rhetorical relations that are available to extend a current span is the following: in the beginning, when we construct a tree structure for a text, we can use any of the relations that hold among the units of the text. However, since only one relation can be associated with a node and since each relation can be used at most once, as we proceed with the bottom-up construction of a tree structure, we can use fewer and fewer relations. The last argument of the predicate *S* keeps track of the relations that are still available for future use.

In fact, because the input to the text derivation problem may consist of exclusively disjunctive hypotheses as well, we will have to make sure that no hypothesis trigger is used more than once. To understand why this is the case, assume that we are interested in deriving the valid discourse structures of text 3.27. And assume that the rhetorical relations shown in 3.28 have been hypothesized to hold between the elementary units on the basis of the occurrence of discourse marker *In contrast*, and on the basis of the fact that both A9 and C9 "talk about" John.

[John likes ice cream and chocolate.^{A9}] [*In contrast*, Mary likes bananas and strawberries.^{B9}] [John likes everything that is sweet.^{B9}] (3.27)

$$RR = \begin{cases} rhet\_rel(\text{CONTRAST}, \text{A9}, \text{B9}) \oplus rhet\_rel(\text{CONTRAST}, \text{A9}, \text{C9}) \\ rhet\_rel(\text{ELABORATION}, \text{C9}, \text{A9}) \end{cases} \tag{3.28}$$

A possible discourse structure of text 3.27 is shown in Figure 3.6. Although this structure does not seem to violate any of the constraints of valid text structures we have discussed so far, it is obviously incorrect, because it uses the rhetorical relation of CONTRAST twice. To prevent such discourse structures from being constructed, we should make sure that at most one rhetorical relation of an exclusively disjunctive hypothesis can be used in a valid tree. In other words, no coherence trigger should be used more than once in building a discourse tree.

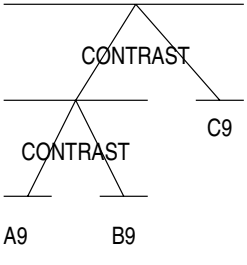

Figure 3.6
An incorrect rhetorical analysis of text 3.27

In order to avoid a hypothesis being used more than once, we only need to define one set-specific relation, $\in_{\oplus}$, and one set-specific operator (function), $\setminus_{\oplus}$. In explaining their semantics, we use the sets of rhetorical relations shown in 3.29 and 3.30 below.

$$rr_1 = \begin{cases} rhet\_rel(\text{CONTRAST}, 1, 2) \oplus rhet\_rel(\text{CONTRAST}, 1, 3) \\ rhet\_rel(\text{ELABORATION}, 3, 1) \end{cases} \tag{3.29}$$

$$rr_2 = \begin{cases} rhet\_rel(\text{CONTRAST}, 1, 2) \\ rhet\_rel(\text{ELABORATION}, 3, 1) \\ rhet\_rel(\text{CONCESSION}, 2, 3) \end{cases} \tag{3.30}$$

DEFINITION 3.2 *The relation* $rhet\_rel(name, s, n) \in_{\oplus} rr$ *(*$rhet\_rel\_ext(name, s_s, s_e, n_s, n_e) \in_{\oplus} rr$*) holds if and only if* $rhet\_rel(name, s, n)$ *(*$rhet\_rel\_ext(name, s_s, s_e, n_s, n_e)$*) occurs in set rr either as a simple (extended) relation, or as one of the disjuncts of an exclusive disjunction of rhetorical relations.*

For example, the following relations hold.

$rhet\_rel(\text{CONTRAST}, 1, 2) \in_{\oplus} rr_1$

$rhet\_rel(\text{CONTRAST}, 1, 2) \in_{\oplus} rr_2$

DEFINITION 3.3 *The elements of the set* $rr \setminus_{\oplus} \{rhet\_rel(name, s, n)\}$ *(*$rr \setminus_{\oplus} \{rhet\_rel\_ext(name, s_s, s_e, n_s, n_e)\}$*) are given by the simple, extended, and exclusively disjunctive rhetorical relations in rr that are not equal to* $rhet\_rel(name, s, n)$ *(*$rhet\_rel\_ext(name, s_s, s_e, n_s, n_e)$*) and that do not have a disjunct equal to* $rhet\_rel(name, s, n)$ *(*$rhet\_rel\_ext(name, s_s, s_e, n_s, n_e)$*). In the case in which one of the disjuncts is* $rhet\_rel(name, s, n)$ *(*$rhet\_rel\_ext(name, s_s, s_e, n_s, n_e)$*), the whole collection of related disjuncts is eliminated from the set.*

For example, the application of the $\setminus_{\oplus}$ operator yields the following results.

$$rr_1 \setminus_{\oplus} \{rhet\_rel(\text{CONTRAST}, 1, 2)\} = \{rhet\_rel(\text{ELABORATION}, 3, 1)\}$$

$$rr_2 \setminus_{\oplus} \{rhet\_rel(\text{CONTRAST}, 1, 2)\} = \{rhet\_rel(\text{ELABORATION}, 3, 1), rhet\_rel(\text{CONCESSION}, 2, 3)\}$$

Using the relation $\in_{\oplus}$ and the function $\setminus_{\oplus}$, we can now write a set of Horn axioms that explain how text spans can be assembled into larger spans. These axioms provide a procedural account of compositionality criterion 2.2.

Assume that there exist two spans: one from unit l to unit b that is characterized by valid text structure $tree_1(\ldots)$ and rhetorical relations rr_1, and the other from unit $b+1$ to unit h that is characterized by valid text structure $tree_2(\ldots)$ and rhetorical relations rr_2. Assume also that rhetorical relation $rhet\_rel(name, s, n)$ holds between a unit s that is in

the promotion set of span $[l, b]$ and a unit n that is in the promotion set of span $[b + 1, h]$, that $rhet\_rel(name, s, n)$ can still be used to extend both spans $[l, b]$ and $[b + 1, h]$ ($rhet\_rel(name, s, n) \in_{\oplus} rr_1$ and $rhet\_rel(name, s, n) \in_{\oplus} rr_2$), and assume 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 3.31) or SATELLITE (see axiom 3.32), type *name*, promotion set p_2, 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 $rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, s, n)\}$.

$$\begin{aligned}
&[S(l, b, tree_1(\text{SATELLITE}, type_1, p_1, left_1, right_1), rr_1)\\
&\wedge S(b + 1, h, tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2)\\
&\wedge rhet_rel(name, s, n) \in_{\oplus} rr_1 \wedge rhet_rel(name, s, n) \in_{\oplus} rr_2\\
&\wedge s \in p_1 \wedge n \in p_2 \wedge hypotactic(name)]\\
&\quad \rightarrow S(l, h, tree(\text{NUCLEUS}, name, p_2, tree_1(\ldots), tree_2(\ldots)),\\
&\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet_rel(name, s, n)\})
\end{aligned} \tag{3.31}$$

$$\begin{aligned}
&[S(l, b, tree_1(\text{SATELLITE}, type_1, p_1, left_1, right_1), rr_1)\\
&\wedge S(b + 1, h, tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2)\\
&\wedge rhet_rel(name, s, n) \in_{\oplus} rr_1 \wedge rhet_rel(name, s, n) \in_{\oplus} rr_2\\
&\wedge s \in p_1 \wedge n \in p_2 \wedge hypotactic(name)]\\
&\quad \rightarrow S(l, h, tree(\text{SATELLITE}, name, p_2, tree_1(\ldots), tree_2(\ldots)),\\
&\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet_rel(name, s, n)\})
\end{aligned} \tag{3.32}$$

Hence, axioms 3.31 and 3.32 treat each exclusive disjunction as a whole, thus ensuring that no hypothesis trigger is used more than once in building a discourse structure. Similarly, we define rules of inference for the cases in which an extended rhetorical relation holds across spans $[l, b]$ and $[b + 1, h]$(3.33–3.34); for the cases in which the nucleus goes before the satellite (3.35–3.38); and for the cases in which the relation under scrutiny is paratactic (3.39–3.42).

$$\begin{aligned}
&[S(l, b, tree_1(\text{SATELLITE}, type_1, p_1, left_1, right_1), rr_1)\\
&\wedge S(b + 1, h, tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2)\\
&\wedge rhet_rel_ext(name, l, b, b + 1, h) \in_{\oplus} rr_1\\
&\wedge rhet_rel_ext(name, l, b, b + 1, h) \in_{\oplus} rr_2 \wedge hypotactic(name)]\\
&\quad \rightarrow S(l, h, tree(\text{NUCLEUS}, name, p_2, tree_1(\ldots), tree_2(\ldots)),\\
&\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet_rel(name, l, b, b + 1, h)\})
\end{aligned} \tag{3.33}$$

$$
\begin{aligned}
&[S(l, b, tree_1(\text{SATELLITE}, type_1, p_1, left_1, right_1), rr_1) \\
&\wedge S(b+1, h, tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2) \\
&\wedge rhet\_rel\_ext(name, l, b, b+1, h) \in_{\oplus} rr_1 \\
&\wedge rhet\_rel\_ext(name, l, b, b+1, h) \in_{\oplus} rr_2 \wedge hypotactic(name)] \\
&\quad \rightarrow S(l, h, tree(\text{SATELLITE}, name, p_2, tree_1(\ldots), tree_2(\ldots)), \\
&\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, l, b, b+1, h)\})
\end{aligned}
\tag{3.34}
$$

$$
\begin{aligned}
&[S(l, b, tree_1(\text{NUCLEUS}, type_1, p_1, left_1, right_1), rr_1) \\
&\wedge S(b+1, h, tree_2(\text{SATELLITE}, type_2, p_2, left_2, right_2), rr_2) \\
&\wedge rhet\_rel(name, s, n) \in_{\oplus} rr_1 \wedge rhet\_rel(name, s, n) \in_{\oplus} rr_2 \\
&\wedge s \in p_2 \wedge n \in p_1 \wedge hypotactic(name)] \\
&\quad \rightarrow S(l, h, tree(\text{NUCLEUS}, name, p_1, tree_1(\ldots), tree_2(\ldots)), \\
&\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, s, n)\})
\end{aligned}
\tag{3.35}
$$

$$
\begin{aligned}
&[S(l, b, tree_1(\text{NUCLEUS}, type_1, p_1, left_1, right_1), rr_1) \\
&\wedge S(b+1, h, tree_2(\text{SATELLITE}, type_2, p_2, left_2, right_2), rr_2) \\
&\wedge rhet\_rel(name, s, n) \in_{\oplus} rr_1 \wedge rhet\_rel(name, s, n) \in_{\oplus} rr_2 \\
&\wedge s \in p_2 \wedge n \in p_1 \wedge hypotactic(name)] \\
&\quad \rightarrow S(l, h, tree(\text{SATELLITE}, name, p_1, tree_1(\ldots), tree_2(\ldots)), \\
&\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, s, n)\})
\end{aligned}
\tag{3.36}
$$

$$
\begin{aligned}
&[S(l, b, tree_1(\text{NUCLEUS}, type_1, p_1, left_1, right_1), rr_1) \\
&\wedge S(b+1, h, tree_2(\text{SATELLITE}, type_2, p_2, left_2, right_2), rr_2) \\
&\wedge rhet\_rel\_ext(name, b+1, h, l, b) \in_{\oplus} rr_1 \\
&\wedge rhet\_rel\_ext(name, b+1, h, l, b) \in_{\oplus} rr_2 \wedge hypotactic(name)] \\
&\quad \rightarrow S(l, h, tree(\text{NUCLEUS}, name, p_1, tree_1(\ldots), tree_2(\ldots)), \\
&\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, b+1, h, l, b)\})
\end{aligned}
\tag{3.37}
$$

$$
\begin{aligned}
&[S(l, b, tree_1(\text{NUCLEUS}, type_1, p_1, left_1, right_1), rr_1) \\
&\wedge S(b+1, h, tree_2(\text{SATELLITE}, type_2, p_2, left_2, right_2), rr_2) \\
&\wedge rhet\_rel\_ext(name, b+1, h, l, b) \in_{\oplus} rr_1 \\
&\wedge rhet\_rel\_ext(name, b+1, h, l, b) \in_{\oplus} rr_2 \wedge hypotactic(name)]
\end{aligned}
$$

$$\rightarrow S(l, h, tree(\text{SATELLITE}, name, p_1, tree_1(\ldots), tree_2(\ldots)), \\ rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, b+1, h, l, b)\}) \tag{3.38}$$

$$[S(l, b, tree_1(\text{NUCLEUS}, type_1, p_1, left_1, right_1), rr_1) \\ \wedge S(b+1, h, tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2) \\ \wedge rhet\_rel(name, n_1, n_2) \in_{\oplus} rr_1 \wedge rhet\_rel(name, n_1, n_2) \in_{\oplus} rr_2 \\ \wedge n_1 \in p_1 \wedge n_2 \in p_2 \wedge paratactic(name)] \\ \rightarrow S(l, h, tree(\text{NUCLEUS}, name, p_1 \cup p_2, tree_1(\ldots), tree_2(\ldots)), \\ rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, n_1, n_2)\}) \tag{3.39}$$

$$[S(l, b, tree_1(\text{NUCLEUS}, type_1, p_1, left_1, right_1), rr_1) \\ \wedge S(b+1, h, tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2) \\ \wedge rhet\_rel(name, n_1, n_2) \in_{\oplus} rr_1 \wedge rhet\_rel(name, n_1, n_2) \in_{\oplus} rr_2 \\ \wedge n_1 \in p_1 \wedge n_2 \in p_2 \wedge paratactic(name)] \\ \rightarrow S(l, h, tree(\text{SATELLITE}, name, p_1 \cup p_2, tree_1(\ldots), tree_2(\ldots)), \\ rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, n_1, n_2)\}) \tag{3.40}$$

$$[S(l, b, tree_1(\text{NUCLEUS}, type_1, p_1, left_1, right_1), rr_1) \\ \wedge S(b+1, h, tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2) \\ \wedge rhet\_rel\_ext(name, l, b, b+1, h) \in_{\oplus} rr_1 \\ \wedge rhet\_rel\_ext(name, l, b, b+1, h) \in_{\oplus} rr_2 \wedge paratactic(name)] \\ \rightarrow S(l, h, tree(\text{NUCLEUS}, name, p_1 \cup p_2, tree_1(\ldots), tree_2(\ldots)), \\ rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, l, b, b+1, h)\}) \tag{3.41}$$

$$[S(l, b, tree_1(\text{NUCLEUS}, type_1, p_1, left_1, right_1), rr_1) \\ \wedge S(b+1, h, tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2) \\ \wedge rhet\_rel\_ext(name, l, b, b+1, h) \in_{\oplus} rr_1 \\ \wedge rhet\_rel\_ext(name, l, b, b+1, h) \in_{\oplus} rr_2 \wedge paratactic(name)] \\ \rightarrow S(l, h, tree(\text{SATELLITE}, name, p_1 \cup p_2, tree_1(\ldots), tree_2(\ldots)), \\ rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, l, b, b+1, h)\}) \tag{3.42}$$

1. *hold*(*RR*)	Axiom (3.23)
2. *position*(A9, 1)	Axiom (3.24)
3. *position*(B9, 2)	Axiom (3.24)
4. *position*(C9, 3)	Axiom (3.24)
5. *S*(1, 1, *tree*(NUCLEUS, LEAF, {A9}, NULL, NULL), *RR*)	1, 2, Axiom (3.25), MP
6. *S*(2, 2, *tree*(NUCLEUS, LEAF, {B9}, NULL, NULL), *RR*)	1, 3, Axiom (3.25), MP
7. *S*(1, 2, *tree*(NUCLEUS, CONTRAST, {A9, B9}, *tree*(NUCLEUS, LEAF, {A9}, NULL, NULL), *tree*(NUCLEUS, LEAF, {B9}, NULL, NULL)), {*rhet_rel*(ELABORATION, C9, A9)})	5, 6, Axiom (3.39), MP
8. *S*(3, 3, *tree*(SATELLITE, LEAF, {C9}, NULL, NULL), *RR*)	1, 4, Axiom (3.26) , MP
9. *S*(1, 3, *tree*(NUCLEUS, ELABORATION, {A9, B9}, *tree*(NUCLEUS, CONTRAST, {A9, B9}, *tree*(NUCLEUS, LEAF, {A9}, NULL, NULL), *tree*(NUCLEUS, LEAF, {B9}, NULL, NULL)), *tree*(SATELLITE, LEAF, {C9}, NULL, NULL)), {∅})	7, 8, Axiom (3.35), MP

Figure 3.7
A derivation of the theorem that corresponds to the valid text structure shown in Figure 3.8

The instantiations of schemata 3.21 and 3.22 over the set of rhetorical relations used by the rhetorical parser, of axiom schemata 3.24 over the units in the text under consideration, and of axioms 3.23, 3.25, 3.26, and 3.31–3.42 provide a *proof theory* for the problem of text structure derivation.

Example of a Derivation of a Valid Text 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 text structures of that text. Assume, for example, that we are given text 3.27 and assume that the rhetorical relations *RR* in 3.28 (page 57) have been hypothesized to hold among the units in the text. In Figure 3.7, we sketch the derivation of the theorem that corresponds to the valid text structure that is shown in Figure 3.8.

The derivation starts with one instantiation of axiom 3.23 and three instantiations of axiom 3.24. Using the axioms in lines 1 and 2, axiom 3.25, and the Modus Ponens rule, we derive the theorem in line 5. Using the axioms in lines 1 and 3, axiom 3.25, and Modus Ponens, we derive the theorem in line 6. Both theorems correspond to valid text structures that can be built on top of elementary units. Using the theorems in lines 5 and 6, axiom 3.39, and Modus Ponens, we derive the theorem in line 7. It corresponds to a valid text structure that can be built across span [1, 2]. Since this structure uses rhetorical relation *rhet_rel*(CONTRAST, A9, B9), the set of rhetorical relations that can be used to expand further the text structure will contain only the relation *rhet_rel*(ELABORATION, C9, A9). Line 8

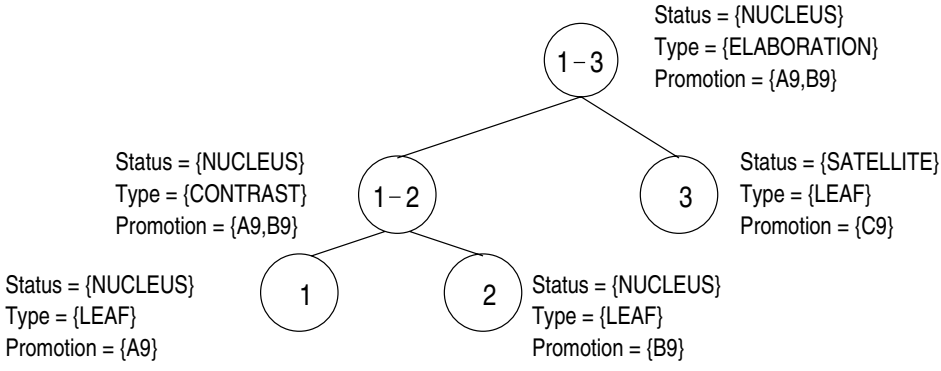

Figure 3.8
The valid text structure that corresponds to the last theorem of the derivation shown in Figure 3.7

corresponds to a valid text structure that can be built on the top of elementary unit c_9. Using the theorems derived in lines 7 and 8, axiom 3.35, and Modus Ponens gives us a theorem that corresponds to a valid structure for the whole text, the structure shown in Figure 3.8.

3.3 The Relation between the Axiomatization of Valid Text Structures and Its Proof Theory

Given the proof theory presented above, it is natural to inquire about its relation to the axiomatization of valid text structures that we have presented in Section 3.1. The theorem below explains the nature of this relation.

THEOREM 3.1 *Given a text T that is characterized by a set of rhetorical relations RR that may be exclusively disjunctive, the application of the proof theory is both sound and complete with respect to the axiomatization of valid text structures. That is, all theorems that are derived using the disjunctive proof theory correspond to valid text structures; and any valid text structure can be derived through the successive application of Modus Ponens and the axioms of the proof theory.*

Proof Since axioms 3.21–3.42 are essentially Horn clauses, for the purpose of this proof, I will treat them in the same way Prolog does. More precisely, instead of focusing on their fixed-point semantics, I will treat axioms 3.21–3.42 of the proof theory from a procedural perspective and consider them to be a Prolog program that, like any other Prolog program, computes inferences only in minimal models [Lloyd, 1987]. Hence, I will show that the

procedural semantics of axioms 3.21–3.42 is consistent with the constraints described in Section 3.1.

In order to prove the theorem, we first observe that the objects of type *tree* that are accepted by the logical language described in the proof theory obey, by definition, most of the constraints specific to a valid text structure. Each of the objects of type *tree* essentially encodes a binary text structure whose nodes are characterized by a promotion set, and a unique status and type. Therefore, by definition, the objects of type *tree* obey the shape of a valid text structure. In order to prove that the axioms are both sound and complete, we only need to prove that the values that are associated with the status, type, and promotion set of each node are consistent with the constraints that characterize the structures that are valid.

Proof of soundness. By definition, given a text of N units $U_N = u_1, u_2, \ldots, u_N$ among which rhetorical relations *RR* hold, $position(u_1, 1), position(u_2, 2), \ldots, position(u_N, N)$ and $hold(RR)$ are the only atomic axioms that correspond to that text—the axioms pertaining to the set of hypotactic and paratactic relations are text-independent. In order to derive theorems, we need to apply one of axioms 3.25–3.42. These axioms fall into two categories. Axioms 3.25 and 3.26 can be applied only on elementary textual units. Their application yields theorems that are characterized by *tree* objects that are valid—these trees are the direct expression of the conventions that we use. Axioms 3.31–3.42 are nothing but a one-to-one translation of strong compositionality criterion 2.2, so the constraints specific to the criterion are always obeyed.

The only thing that may go wrong is that the repeated application of the axioms of the proof theory may create a tree that either uses the same rhetorical relation twice or uses two relations generated by the same rhetorical trigger. For simple and extended rhetorical relations it is impossible to use a relation more than once because a relation that connects two spans $[l, h_1]$ $[h_1 + 1, h]$ cannot connect any of their subspans. And as soon as such a relation is used, it is removed from the list of relations that can be used to extend the tree built up to that point. For well-formed exclusively disjunctive hypotheses, this is impossible as well because adjacent spans do not overlap; two disjuncts of a well-formed exclusively disjunctive hypothesis cannot be used in two spans that do not overlap. And as soon as one of them is used, all disjuncts are made unavailable for future use.

Proof of completeness. The proof follows immediately from lemma 3.1. Given any text T, the algorithm shown in Figure 3.9 derives all the valid discourse trees of any span $[l, h]$ in the text by means of the proof-theoretic account; so it follows that the algorithm also derives all the valid trees of the whole text T. Hence, there is no tree that cannot be derived using the proof-theoretic account. ■

Input: a text T of N units and a set RR of rhetorical relations that hold among these units.

Output: all the theorems that can can be derived by applying axioms of the proof theory of valid text structures and modus ponens.

1. apply axiom schema (3.23)
2. **for** $i := 1$ **to** N
3. apply axiom schema (3.24)
4. apply axiom schemata (3.25)–(3.26)
5. **for** *size_of_span* $:= 1$ **to** N $- 1$
6. **for** $l := 1$ **to** N $-$ *size_of_span*
7. $h = l +$ *size_of_span*
8. **for** $b := l$ **to** $h - 1$
9. **for each** theorem $S(l, b, tree_1, RR_1)$ of span $[l, b]$
10. **for each** theorem $S(b + 1, h, tree_2, RR_2)$ of span $[b + 1, h]$
11. **for each** relation r such that $r \in_{\oplus} RR_1$ and $r \in_{\oplus} RR_2$
12. apply all possible axioms (3.31)–(3.42)

Figure 3.9
An algorithm that applies the proof theory of valid text structures in order to derive all the theorems (valid discourse trees) that characterize a text T

LEMMA 3.1 *Given a text T of* N *elementary units among which rhetorical relations RR hold, the theorems derived by the algorithm in Figure 3.9 by means of the proof theory correspond to* all *valid structures that can be built for any span* $[l, h]$ *of T, where* $1 \leq l \leq h \leq$ N.

Proof The algorithm in Figure 3.9 derives first all theorems that correspond to all the valid text structures that can be built for each of the elementary textual units (lines 2–4). Then, it derives all the theorems that correspond to spans of size 2, 3, . . ., N (lines 5–12). The proof of the lemma reflects the main steps of the algorithm: it is inductive on the number of units in the span $[l, h]$.

Base case (*number_of_units_in_span* $= 1$)*:*
All the valid trees that can be built for any leaf u_i of the text are described by structures that correspond either to term *tree*(SATELLITE, LEAF, $\{u_i\}$, NULL, NULL) or to term *tree*(NUCLEUS, LEAF, $\{u_i\}$, NULL, NULL). Lines 2–4 of the algorithm in Figure 3.9 derive all these structures.

Induction step:
Assume that the lemma holds for all spans $[x, y]$ whose size is less than *number_of_units_in_span* $= k$, i.e., $y - x < k$. We prove now that the lemma holds for span $[l, h]$ of size k as well. By contradiction, assume that there exists a valid structure *vs* that spans across

units $[l, h]$ and assume that the algorithm in Figure 3.9 cannot derive any theorem that corresponds to *vs*. In looser terms, we assume that the algorithm cannot derive a theorem having the form $S(l, h, vs, rr)$.

According to the axiomatization given in Section 3.1, if a valid text structure can be associated with span $[l, h]$, it must be built on the top of two substructures of two adjacent subspans. Since the algorithm iterates over all possible combinations of subspans and over all possible valid structures that correspond to these subspans (lines 8–12), the only situations in which a theorem that corresponds to *vs* can fail to be derived is when one or more of the antecedents that characterize one of the axioms 3.31–3.42 do not hold; and when there exists no axiom to derive *vs*. If we consider in a proof by cases all the possible combinations that could be associated with the status, type, promotion units, and set of rhetorical relations of *vs*, it is trivial to show that, for each combination, there exists an axiom that in conjunction with Modus Ponens derives a theorem that corresponds to *vs*.

For example, assume that *vs* is isomorphic to the structure that corresponds to the third term of theorem 3.43.

$$\begin{aligned} S(l, h, tree(&\text{SATELLITE}, \text{NAME}, \text{P}_2,\\ &tree(\text{SATELLITE}, \text{NAME}_1, \text{P}_1, left_1, right_1),\\ &tree(\text{NUCLEUS}, \text{NAME}_2, \text{P}_2, left_2, right_2)),\\ &RR_{lh}) \end{aligned} \tag{3.43}$$

Since *vs* is valid, it follows that there exist spans $[l, b]$ and $[b + 1, h]$ that are characterized by valid text structures vs_1 and vs_2; these structures correspond to terms $tree(\text{SATELLITE}, \text{NAME}_1, \text{P}_1, left_1, right_1)$ and $tree(\text{NUCLEUS}, \text{NAME}_2, \text{P}_2, left_2, right_2)$ respectively. According to the induction hypothesis, this means that the theorems given in 3.44 and 3.45 hold for some $rr_1, rr_2 \subseteq RR$.

$$S(l, b, tree(\text{SATELLITE}, \text{NAME}_1, \text{P}_1, left_1, right_1), rr_1) \tag{3.44}$$

$$S(b + 1, h, tree(\text{NUCLEUS}, \text{NAME}_2, \text{P}_2, left_2, right_2), rr_2) \tag{3.45}$$

Also, since *vs* is a valid structure, this also means that rhetorical relation NAME is either a simple hypotactic relation that holds between two elementary units, one unit $\text{S} \in [l, b]$ and one unit $\text{N} \in [b + 1, h]$, or an extended hypotactic relation that holds between the two spans. In both cases, the relation can be a member of a well-formed exclusively disjunctive hypothesis. Assume that NAME is a simple relation (if NAME is an extended relation, the proof is similar). In order to be able to apply the axiom given in 3.31, we only need to prove that $rhet\_rel(\text{NAME}, \text{S}, \text{N}) \in_{\oplus} rr_1$ and $rhet\_rel(\text{NAME}, \text{S}, \text{N}) \in_{\oplus} rr_2$.

Now, all the sets of rhetorical relations that are associated with all theorems derived for all spans of size smaller that $h - l$ are either equal to *RR* or are obtained from *RR* through successive eliminations of relations that are used to build valid text structures. We consider two cases:

THE RELATION IS SIMPLE. Since *rhet_rel*(NAME, S, N) holds across two units that belong to spans $[l, b]$ and $[b + 1, h]$ respectively, it is obvious that this relation could not have been used to build either the tree structure for vs_1 or that for vs_2. Hence, *rhet_rel*(NAME, S, N) must be both in the set rr_1 and in the set rr_2.

THE RELATION IS THE MEMBER OF A WELL-FORMED EXCLUSIVELY DISJUNCTIVE HYPOTHESIS. Since *vs* is a valid discourse structure tree, it does not use a member of an exclusively disjunctive hypothesis more than once. Hence, if *name* is such a relation, neither vs_1, nor vs_2 are characterized by relations that pertain to the same well-formed exclusively disjunctive hypothesis. Hence, the disjunctive hypothesis that contains *name* must be a member of both rr_1 and rr_2.

Hence, all the antecedents that pertain to axiom 3.31 are true. Therefore, one can use axiom 3.31 and Modus Ponens to derive theorem 3.43, which contradicts our initial hypothesis that *vs* cannot be derived. The proof of the other cases is similar. ■

4 A Computational Account of the Axiomatization of Valid Text Structures and its Proof Theory

4.1 Introduction

In the previous chapter, I have shown that the axiomatization of valid text structures and the proof theory are equivalent. That is, given the same problem of text structure derivation (see definition 2.2), they can both be used to derive the same valid discourse structures. The axiomatization and the proof theory give rise to two alternatives to computing the valid structures of a text.

• The first alternative takes advantage of the declarative formalization and equates the process of tree derivation with the process of finding the models of a theory that enumerates the axioms that characterize the general constraints of a text structure and the axioms that characterize the text under scrutiny. This alternative employs model-theoretic techniques.

The major benefit of this alternative is that it enables a declarative, clear formulation of the linguistic constraints that characterize the structures that are valid; such a formulation is independent of the algorithms that derive the structures.

• The second alternative implements the rewriting rules of the proof theory. This alternative employs theorem-proving techniques.

The major benefit of this alternative is that it enables one to control directly the process of text structure derivation. When a text has thousands of valid text structures and one wants to derive only some of them, it is crucial to be able to control the discourse structure derivation process and focus only on a subset of preferred interpretations. (Section 6.3.5 and Chapter 10 elaborate more on this issue.)

In this chapter, I discuss implementations of both alternatives. I discuss briefly three paradigms for solving the problem of text structure derivation given in definition 2.2. In two of these paradigms I use model-theoretic techniques, i.e., I discuss how the problem of text structure derivation can be encoded as a classical constraint-satisfaction problem and as a propositional, satisfiability problem. For the third paradigm, I discuss a straightforward implementation of the proof theory of valid text structures. The interested reader can find details concerning all paradigms in [Marcu, 1998a].

4.2 Deriving Text Structures—A Constraint-Satisfaction Approach

Given a sequence U of N textual units, one can take advantage of the structure of the domain and associate with each of the $N(N+1)/2$ possible text spans a status and a type variable whose domains consist in the set of objects over which the corresponding predicates S and T range. Hence, the domains of these predicates are given by axioms 3.4 and 3.6. Axioms 3.5 and 3.7 need not be explicitly encoded: a solution to a constraint-satisfaction problem will choose by definition unique values from the domain of each variable. For each

of the $N(N+1)/2$ possible text spans $[l, h]$, one can also associate $h - l + 1$ promotion variables. These are boolean variables that specify whether units $l, l+1, \ldots, h$ belong to the promotion set of span $[l, h]$. These variables encode axiom 3.8.

Hence, each text of N units yields a constraint-satisfaction problem with $N(N+1)(N+11)/6$ variables ($N(N+1)(N+11)/6 = 2N(N+1)/2 + \sum_{l=1}^{l<=N} \sum_{h=l}^{h<=N}(h-l+1)$). The constraints associated with these variables are a one-to-one mapping of axioms 3.9–3.16. Finding the set of RS-trees that are associated with a given discourse amounts then to finding all the solutions of a traditional constraint-satisfaction problem.

4.3 Deriving Text Structures—A Propositional Logic, Satisfiability Approach

Recent successes in using greedy methods for solving large satisfiability problems [Selman et al., 1992, Selman et al., 1994, Kautz and Selman, 1996] prompted me to investigate their appropriateness for finding the discourse structure of text, which can be accomplished by representing the axioms of valid text structures in propositional logic. In order to do this, we need to provide a propositional representation of the status, type, and promotion variables of each of the $N(N+1)/2$ potential textual spans in a text and a propositional representation of the constraints between them.

Each potential textual span has a status that can be NUCLEUS, SATELLITE, or NONE. Two propositional variables suffice to encode these three possible values; for ease of reference, we label each pair of propositional variables that encode the status of each span $[l, h]$ with $S_{l,h,\text{NUCLEUS}}$ and $S_{l,h,\text{SATELLITE}}$. If a truth assignment assigns the value "true" to $S_{l,h,\text{NUCLEUS}}$, we consider that the status of span $[l, h]$ is NUCLEUS; if a truth assignment assigns the value "true" to $S_{l,h,\text{SATELLITE}}$, we consider that the status of span $[l, h]$ is SATELLITE; if a truth assignment assigns the value "false" both to $S_{l,h,\text{NUCLEUS}}$ and $S_{l,h,\text{SATELLITE}}$, we consider that the status of span $[l, h]$ is NONE. Since a textual span cannot play both a NUCLEUS and a SATELLITE role in the same text structure, no model will assign the value "true" both to $S_{l,h,\text{NUCLEUS}}$ and $S_{l,h,\text{SATELLITE}}$. Because the final representation is characterized by $N(N+1)/2$ potential spans, it follows that a text of N units will yield $N(N+1)$ propositional status variables.

Axioms 3.4 and 3.5 can be expressed in propositional logic as follows. For each leaf, an appropriate encoding consists of two conjunctive-normal-form formulas of size two, which are the expression of an exclusive "or" between the variables $S_{i,i,\text{NUCLEUS}}$ and $S_{i,i,\text{SATELLITE}}$:

$$S_{i,i,\text{NUCLEUS}} \vee S_{i,i,\text{SATELLITE}} \tag{4.1}$$

$$\neg S_{i,i,\text{NUCLEUS}} \vee \neg S_{i,i,\text{SATELLITE}} \tag{4.2}$$

For each non-elementary span $[l, h]$, $(h - l > 0)$, formula 4.3 expresses that the status of a span can be NUCLEUS, SATELLITE, or NONE.

$$\neg S_{l,h,\text{NUCLEUS}} \vee \neg S_{l,h,\text{SATELLITE}} \tag{4.3}$$

Similarly, one can represent in propositional logic the constraints specific to the type and promotion variables and the constraints on the overall structure of the text. Marcu [1998a] provides full details of all axioms of such an encoding; for a text derivation problem with N elementary units, the equivalent propositional representation is characterized by $O(N^3)$ variables and about $O(N^5)$ conjunctive-normal-form constraints. Finding a solution to the problem of text structure derivation amounts then to finding a model of a propositional logic theory.

4.4 Deriving Text Structures—A Proof-Theoretic Approach

There are many ways in which one can implement a set of rewriting rules of the kind described in the proof theory. For example, one can encode all the axioms as Horn clauses and let the Prolog inference mechanism derive the valid discourse structures of a text. In such a case, the valid discourse structures are built through unification-specific mechanisms that implement the rewriting rules presented above.

Another choice is to write a grammar having rules such as those shown in 4.4, where each grammar rule is associated with a set of semantic constraints in the style of Montague [1973].

$$\begin{aligned}
&S(sem) \rightarrow u_i\{sem = \{tree(\text{NUCLEUS, LEAF, } \{u_i\}, \text{ NULL, NULL}), RR\}\} \\
&S(sem) \rightarrow u_i\{sem = \{tree(\text{SATELLITE, LEAF, } \{u_i\}, \text{ NULL, NULL}), RR\}\} \\
&S(sem) \rightarrow S(sem_1)S(sem_2)\{sem = f(sem_1, sem_2)\}
\end{aligned} \tag{4.4}$$

The grammar-based approach assumes that the input is a sequence of textual units $u_1, u_2, \ldots, u_N$. Each nonterminal S in the grammar has associated a semantic representation that reflects the valid structure that corresponds to that derivation and the set of rhetorical relations that can be used for further derivations. The semantic constraints $sem = f(sem_1, sem_2)$ that characterize all juxtapositions of nonterminals are a one-to-one expression of the constraints expressed in axioms 3.31–3.42. For example, the semantic constraint associated with rule 3.31 is that shown in 4.5 below.

$$\begin{aligned}
&[sem_1 = \{tree_1(\text{SATELLITE}, type_1, p_1, left_1, right_1), rr_1\} \\
&\quad \wedge\ sem_2 = \{tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2\} \\
&\quad \wedge\ rhet_rel(name, s, n) \in_{\oplus} rr_1 \wedge rhet_rel(name, s, n) \in_{\oplus} rr_2 \\
&\quad \wedge\ s \in p_1 \wedge n \in p_2 \wedge hypotactic(name)] \\
&\hline \\
&sem = \{tree(\text{NUCLEUS}, name, p_2, tree_1(\ldots), tree_2(\ldots)), \\
&\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet_rel(name, s, n)\}
\end{aligned} \tag{4.5}$$

Once such a grammar is written, one can apply any parsing algorithm in order to produce all valid discourse trees of a text.

4.5 Implementation and Empirical Results

I have implemented algorithms that take as input a problem of text structure derivation and automatically derive constraint-satisfaction, propositional, and grammar-based representations of it. I have empirically compared the suitability of applying model- and proof-based techniques for deriving the valid structures of eight manually encoded text derivation problems. The simplest text had three elementary units among which four simple rhetorical relations held; the most complex had 19 elementary units among which 25 simple rhetorical relations held.

In order to find the valid text structures that corresponded to the eight problems, I have used the following software packages.

• For the constraint-satisfaction (CS) approach, I have used Lisp and Screamer [Siskind and McAllester, 1993a, Siskind and McAllester, 1993b], a macro package that provides constraint-satisfaction facilities.

• For the propositional-logic (PL) approach, I have used off-the-shelf, efficient implementations of the Davis-Putnam (*http://www.cirl.uoregon.edu/crawford/*) [Crawford and Auton, 1996], GSAT, and WALKSAT (*ftp://ftp.research.att.com/dist/ai/*) [Selman et al., 1992, Selman et al., 1994] procedures for finding models of propositional theories in conjunctive normal form.

The Davis-Putnam [1960] (DP) procedure backtracks over the space of all truth assignments, incrementally assigning truth values to variables and simplifying formulas. Backtracking occurs whenever no "new" variable can be assigned a truth value without producing inconsistency. In contrast, the GSAT procedure performs a greedy local search [Selman et al., 1992]. The procedure incrementally modifies a randomly generated truth assignment by "flipping" the assignment of the variable that leads to the largest increase in the total number of satisfied formulas. The "flipping" process is repeated until a truth assignment is found or until an upper threshold, MAX-FLIPS, is reached. If no satisfying truth assignment is found after MAX-FLIPS, the whole process is repeated. At most MAX-TRIES repetitions are allowed. WALKSAT [Selman et al., 1994] is a variant of GSAT that introduces some "noise" in the local search. With probability p, the WALKSAT algorithm picks a variable occurring in some unsatisfied clause and flips its truth assignment. With probability $1 - p$, WALKSAT follows the standard greedy schema of GSAT, i.e., it makes the best possible move.

Table 4.1
Performance of the constraint-satisfaction (CS), propositional logic (GSAT, WALKSAT, and DP), and proof-theory (PT) approaches to text structure derivation

		Time (in seconds) required to derive				
		all trees	one tree			all trees
Text	Number of valid trees	CS	GSAT	WALKSAT	DP	PT
1	3	<1	<1	<1	<1	<1
2	5	38	<1	<1	<1	<1
3	40	–	17	<1	<1	<1
4	8	–	–	<1	<1	<1
5	20	–	–	<1	<1	<1
6	816	–	–	–	4	19
7	2584	–	–	–	137	45
8	24055	–	–	–	9021	13227

• For the proof-theoretic (PT) approach, I have used a modified version of the bottom-up parser described by Norvig [1992]. The parser applies a memoization procedure[1] in order to avoid computing the same structure twice, being therefore equivalent to a chart parser.

Table 4.1 shows the amounts of time on a Sparc Ultra 2–2170 that were required by these implementations for determining one or all valid text structures of these texts. The dashed lines correspond to computations that did not terminate. When the proof-theory-based implementation was required to determine only one valid tree, it did so in less than a second for all eight text derivation problems. As one can see, the proof-theory-based implementation is by far the most efficient one, followed by the propositional- and constraint-satisfaction-based implementations. In spite of being able to rapidly compute all valid trees of a text, the proof-theory implementation does not seem to handle well the explosion of valid trees that may characterize some text derivation problems.

From a theoretical perspective, these results are not surprising. After all, Maxwell and Kaplan [1993] and Barton et al. [1985] have already shown that parsing phrase structure trees in the presence of functional constraints can be exponential in the worst case. Therefore, deriving the valid text structures of a text using a grammar-based implementation of the proof theory can be exponential in the worst case. This suggests that for highly rhetorically ambiguous texts, it may be useful to define preference criteria over the set of valid

1. A memoization procedure consists in creating dynamically a database of function input/output pairs; whenever a memoized function is called, the database is checked in order to avoid computing the same function more than once.

interpretations and to produce only some of the tree structures that are valid, i.e., the structures that are the most preferred. (See Sections 6.3.5 and Chapter 10 for ways to define preferences over discourse structures.)

The comparison of the model-theoretic implementations is interesting from two perspectives. On one hand, from a linguistic perspective, the propositional encoding shows a significant improvement over the constraint-satisfaction encoding: the Davis-Putnam implementation derived one valid text structure for each of the eight texts that we considered. However, since the number of conjunctive normal formulas is in the range of $O(N^5)$, it is obvious that a direct application of the method is ill-suited for real texts, where the number of elementary units is in the hundreds and even the thousands.

On the other hand, from a computational perspective, the encoding raises some interesting questions with respect to the adequacy of stochastic methods for finding models of propositional theories. Most of the research on greedy methods that was generated in the last five years is concerned with propositional satisfiability problems that are randomly generated. Empirical studies showed that, for such problems, the GSAT algorithm significantly outperforms the Davis-Putnam procedure. However, as Table 4.1 shows, for the propositional encoding of the problem of text structure derivation, which is highly structured, it seems that the reverse holds. It is surprising that even WALKSAT, which adds some noise to the GSAT procedure, fails to find satisfying truth assignments for problems on which DP succeeds. For example, Selman, Levesque, and Mitchell [1992] noticed that whenever a problem was easy to solve by the DP procedure, it was also easy to solve by GSAT. The results presented in this section do not seem to follow the same pattern. In addition, although empirical results showed repeatedly that the DP procedure is intractable for randomly generated propositional encodings that have more than 500 variables, in our case, it manages to find satisfying truth assignments in less than two and a half hours for propositional encodings of the problem of text structure derivation that have more than 4,000 variables and more than 140,000 clauses!

I believe that a much deeper investigation of the computational properties of exhaustive and stochastic procedures with respect to the class of problems that I presented in this section is required in order to derive valid conclusions. Such an investigation is beyond the scope of this book.

In addition to the three paradigms compared here, I have also investigated a paradigm in which the problem of text structure derivation is mapped into a parsing problem using grammars in Chomsky normal form [Marcu, 1998a]. When the problem of text structure derivation is characterized only by simple and extended rhetorical relations that do not cross, this paradigm yields an algorithm that can determine in polynomial time all valid trees of a text. Unfortunately, when exclusively disjunctive relations are considered, the paradigm is equivalent to the proof theory. Since the focus in this book is on deriving

the valid structures of unrestricted texts, which are usually formalized using exclusively disjunctive relations, I did not include here a discussion of this paradigm. The interested reader can find details about it in [Marcu, 1998a].

4.6 Applications

The axiomatization of valid text structure and its proof theory provide formal and computational means for deriving *all and only* the valid structures of a text. For example, if we go back now to our original example that was shown in 2.4, the implementation of the proof theory produces five distinct valid text structures, which are shown in Figure 4.1.

Among the set of trees in Figure 4.1, trees 4.1a and 4.1b match the trees given earlier in Figure 2.4a and 2.4c. Trees 4.1c–e represent trees that are not given in Figure 2.4. As discussed in Section 3.1.3, the tree in Figure 2.4b is not derived by the implementation because it is invalid: it does not enforce the compositionality constraints (axiom 3.12).

If the relations to the same text consisted of the relations given below in 4.6, the implementation of the proof theory would generate only one tree, the tree shown in Figure 4.1.e.

$$\begin{cases} \textit{rhet\_rel}(\text{JUSTIFICATION}, \text{D}_1, \text{B}_1) \\ \textit{rhet\_rel}(\text{EVIDENCE}, \text{C}_1, \text{B}_1) \\ \textit{rhet\_rel}(\text{JUSTIFICATION}, \text{A}_1, [\text{B}_1 - \text{D}_1]) \end{cases} \tag{4.6}$$

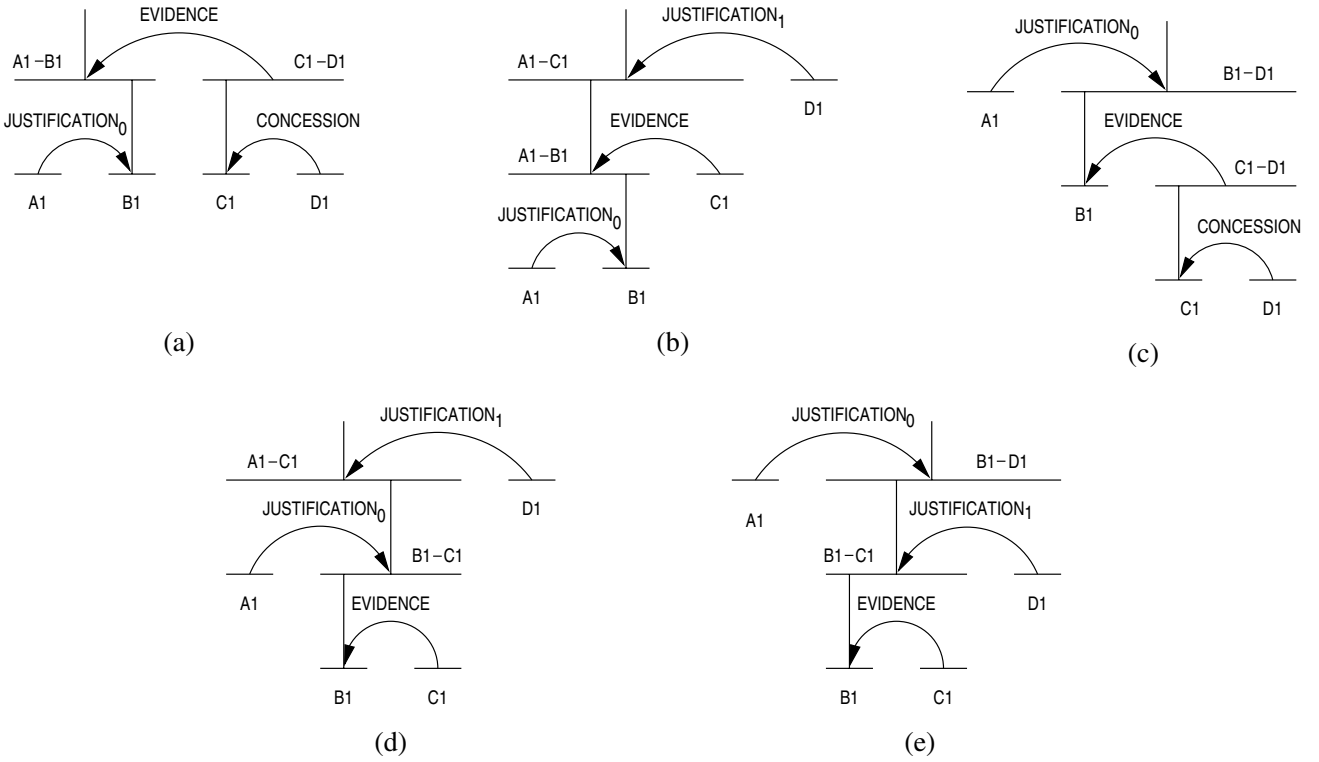

Figure 4.1
The set of all RS-trees that could be built for text 2.4

5 Discussion

5.1 Related Work

To my knowledge, all approaches to deriving discourse structures that were proposed previously were incremental. That is, they assumed that elementary discourse units are processed sequentially and that a discourse tree is created by incrementally updating a tree structure that corresponds to the discourse units that were processed up to the unit under scrutiny. The unit under scrutiny provides information about the way the updating operation should be performed. These approaches fall into two classes: they are either logic- or grammar-based.

In logic-based approaches [Zadrozny and Jensen, 1991, Lascarides and Asher, 1993, Asher, 1993], the idea of structure is only implicit. Discourse trees can be obtained by considering the coherence relations that hold among the discourse units, which are first-class entities in a logic that captures both the semantics of sentences and the semantics of discourse. Because the logic-based approaches are couched in terms of default logics and logics of beliefs, they are intractable.

In grammar-based approaches [van Dijk, 1972, Polanyi, 1988, Scha and Polanyi, 1988, Gardent, 1994, Hitzeman et al., 1995, Polanyi and van den Berg, 1996, van den Berg, 1996, Gardent, 1997, Schilder, 1997, Cristea and Webber, 1997, Webber et al., 1999], the structure of discourse is explicitly represented; it is assimilated with the parse tree of a sequence of discourse constituents. The first attempts to write discourse grammars [van Dijk, 1972] put very few constraints on the applicability of the rules. However, further developments brought in more and more constraints that were both semantic and structural in nature. The semantic constraints stipulate the conditions that must hold in order to join an incoming discourse unit to an existing discourse structure. For example, in order to substitute a unit on the right frontier[1] of an existing discourse tree with an incoming elementary discourse tree, the semantic information associated with the unit on the right frontier must unify with the semantic information associated with the elementary discourse tree [Gardent, 1997]. The structural constraints are a direct consequence of the assumption that discourse processing is incremental. To account for the sequentiality of text, grammar-based approaches allow only the nodes on the right frontier of a discourse tree to be updated.

Some of the grammar-based approaches to discourse are extensions of context-free and HPSG grammars [van Dijk, 1972, Scha and Polanyi, 1988, Hitzeman et al., 1995]. However, the most recent approaches [Gardent, 1994, van den Berg, 1996, Polanyi and van den Berg, 1996, Gardent, 1997, Schilder, 1997, Cristea and Webber, 1997, Webber

1. The right frontier is the set of nodes of the tree structure that are found on a path from the root to the right-most leaf.

et al., 1999] rely on extensions of tree-adjoining grammars (TAGs) [Joshi, 1987]. The appeal of using TAGs for discourse processing seems to follow from the power of the adjoining operations, which allow trees to be not only expanded, as in the case of context-free grammars, but also rewritten.

In contrast to this work, the formalization that I have presented in Section 3.1.2 provides a mathematical description of the valid text structures, i.e., an expression of the properties of the class of structures that are licensed by the essential features that were put forth in Section 2.1.1 and by the strong compositionality criterion 2.2. As such, the formalization in Section 3.1.2 can be interpreted as a sibling of model-theoretic frameworks that characterize the properties of the syntactic structures of sentences [Keller, 1993, Blackburn et al., 1995, Rogers, 1994, Rogers, 1996]. In contrast with model-theoretic approaches to syntax, the formalization presented in this chapter is much simpler. The constraints on the features of the trees (discourse structures) that the axiomatization of valid text structures captures are much simpler than the constraints that are used by syntactic theories. Because of this, unlike model-theoretic approaches to syntax, which use highly expressive languages with modal operators and second-order quantifiers, the formalization presented here can be couched into the language of first-order logic.

To my knowledge, the formalization of text structures provided in Section 3.1.2 is the first attempt to provide a model-theoretic framework for the study of discourse in general and the study of RST in particular. The proof theory presented in Section 3.2 differs from previous approaches to discourse parsing because it does not assume the derivation of discourse trees to be incremental. Rather it assumes that one first hypothesizes the rhetorical relations that hold between elementary and non-elementary discourse units, and then derives all discourse structures that are valid with respect to the axiomatization. This approach has both advantages and disadvantages.

Advantages

DEALING WITH EXPECTATIONS AND NONMONOTONICITY. Cristea and Webber [1997] introduced a mechanism that enables the incremental derivation of discourse structures in the presence of expectations. For example, the occurrence of the expression "on one hand" raises the expectation that the discourse will subsequently express some contrasting situation. In spite of this, incremental processing along the lines described in most current grammar-based approaches may be inefficient from a computational perspective. Consider example 5.1, which is reproduced from [Cristea and Webber, 1997].

[Because John is such a generous man^{A10}][—whenever he is asked for money,^{B10}] [he will give whatever he has, for example^{C10}][—he deserves the "Citizen of the Year" award.^{D10}] (5.1)

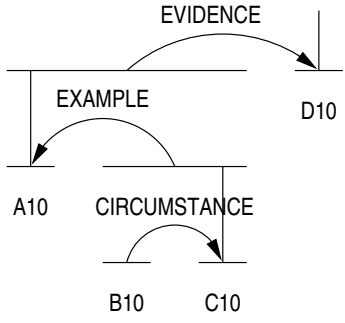

Figure 5.1
The valid text structure of text 5.1

As Cristea and Webber note, the fact that unit B_{10} provides together with unit C_{10} an example for A_{10}, rather than satisfying the expectation raised by "Because", becomes apparent only when unit C_{10} is processed—more specifically, when the discourse marker "for example" is considered. Obviously, in order to accommodate the finding that units B_{10} and C_{10} are an example for the idea presented in the first unit, we have to undo the adjoining of node B_{10}. Therefore, the incremental processing of discourse cannot be monotonic. In order to deal with the nonmonotonicity of incremental discourse derivation, we have to either consider, in the style of Tomita [1985], all possible ways in which a tree can be extended or allow for backtracking. Either approach negatively affects the computational properties of an incremental discourse parser.

Given the discourse parsing paradigm in this book, in order to derive the valid discourse structure of text 5.1, one would first determine that the relations given in 5.2 hold among the elementary units of the text. Then, one would use any of the implementations discussed in Chapter 4 in order to obtain the valid discourse structure shown in Figure 5.1.

$$\begin{cases} \textit{rhet\_rel}(\text{EVIDENCE}, A_{10}, D_{10}) \\ \textit{rhet\_rel}(\text{CIRCUMSTANCE}, B_{10}, C_{10}) \oplus \textit{rhet\_rel}(\text{CIRCUMSTANCE}, B_{10}, D_{10}) \\ \textit{rhet\_rel}(\text{EXAMPLE}, C_{10}, A_{10}) \oplus \textit{rhet\_rel}(\text{EXAMPLE}, C_{10}, B_{10}) \end{cases} \tag{5.2}$$

Since the derivation process starts only after all relations have been hypothesized, one should no longer worry about designing special mechanisms for dealing with expectations and nonmonotonicity.

Dealing with lack of information. In an experiment carried out with Estibaliz Amorrortu and Magdalena Romera [Marcu et al., 1999a], we attempted to build discourse trees incrementally using a discourse annotation tool. The tool allowed us to see only one

sentence at a time. As we identified the discourse units, we attempted to immediately attach them to the right frontier of a growing discourse structure. During the experiment, we noticed that, quite often, we did not have sufficient information in order to decide where to attach the units. The annotating style varied significantly among us: one of us postponed the attachment decisions 37%, one 18%, and one 9% of the time.

During the experiment, we noticed that managing multiple partial discourse trees during the annotation process was the norm rather than the exception. In fact it was not that elementary discourse units were attached incrementally to *one* partial discourse structure, although we did our best to do so, but rather that multiple partial discourse structures were created and then assembled using a rich variety of operations. Moreover, even this strategy proved to be somewhat inadequate, since we needed from time to time to change rhetorical relation labels (2–3% of the operations) and restructure completely the discourse (1–2% of the operations).

This data suggests that it is unlikely that we will be able to build perfect discourse parsers that can incrementally derive discourse trees without applying any form of backtracking. If humans are unable to decide incrementally, in 100% of the cases, where to attach elementary discourse units, it is unlikely we can build computer programs that are.

The approach to discourse parsing proposed in this part of the book does not suffer from this problem, since a human or automatic process has access to all units in order to hypothesize rhetorical relations.

Disadvantages

DEALING WITH RHETORICAL AMBIGUITY. Just as a syntactic parser may generate hundreds of thousands of correct syntactic trees, so can the model theoretic implementations generate hundreds of thousands of valid discourse trees. Finding appropriate mechanisms for dealing with this ambiguity is not a trivial issue, because in the approach taken here, it is treated globally. In contrast, incremental parsing models seem to suffer less from this problem because disambiguation decisions can be made locally, during the discourse structure derivation process.

PSYCHOLINGUISTIC PLAUSIBILITY. The nonincremental paradigm that I presented in this part of the book is not psycholinguistically plausible—after all, humans do process text in an incremental fashion. Given the psychological constraints and the limited resources that humans have, it is conceivable that incremental processing is impossible without backtracking—this would be consistent with the mistakes and reinterpretations that are observed in naturally occurring conversations [Hirst et al., 1993, McRoy, 1993]. Yet, a correct computational model of these processes seems to require more sophistication than current incremental models of discourse parsing.

5.2 Open Problems

The main shortcoming of my formal inquiry into the structure of text comes from its simplicity. The formalization discussed in this chapter ignores a wealth of linguistic phenomena that have been shown to be important in discourse understanding. These phenomena include focus, topic, cohesion, pragmatics, etc. Formalizing these linguistic dimensions of text and incorporating them into the formal model presented in this book is a research direction that promises to be extremely rewarding.

By incorporating a representation of intentions into the formal model discussed here, for example, one can investigate formally and computationally the relation between discourse structures and intentions. More precisely, one can determine what intentions are consistent with a given discourse structure and what discourse structures are consistent with a given intention (see [Marcu, 1999b] for such an approach). By enriching the formal model with logical objects that represent the referring expressions in a text (as Asher [1993], for example), one can also study the relation between discourse structure and reference.

All these are reasonable extensions, which can be easily couched into the formalism discussed here. But even if we end up extending the formalism to cover phenomena such as intentions, referring expressions, and focus, we can still attack some of the assumptions on which the formalization relies. For example, the formalization assumes that text can be sequenced into elementary units—however, as we will see in the next chapter, providing an objective definition for such units is not trivial. Quite often, the boundary between discourse and syntactic phenomena is blurry [Marcu, 1999c]. Eventually, the formalism will need to be refined to a level where the boundary between discourse and syntax becomes transparent (see Section 8.2 for a discussion).

Another assumption that can be challenged concerns the tree-like representation of discourse, which does not permit an explicit representation of multiple relations that hold between various textual units. For example, consider the text shown in 5.3, below, in which we have labeled not the elementary units, but the discourse segments of interest.

[According to engineering lore, the late Ermal C. Fraze, founder of Dayton Reliable Tool & Manufacturing Company in Ohio, came up with a practical idea for the pop-top lid after attempting with halting success to open a beer can on the bumper of his car.[1]] [For decades, inventors had been trying to devise a can with a self-contained opener. Their elaborate schemes had proved unworkable because they required complex manufacturing steps for the attachment of the pull tab—the element that exerts force to open the can top.[2]] [Fraze succeeded because he conceived of a simple and economical rivet to hold the tab in place. Unlike in previous approaches, the rivet was formed from the surface of the can top itself.[3]] (5.3)

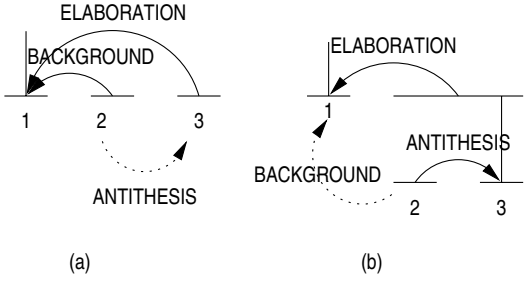

Figure 5.2
Limitations of the tree-like representation of discourse structures

One possible high-level rhetorical analysis of text 5.3 is that shown in Figure 5.2a. In this analysis, the emphasis is on representing explicitly that span 2 describes the historical context (BACKGROUND) in which Ermal Fraze discovered the pop-top lid (sentence 1); and that span 3 ELABORATES on sentence 1 by giving more details about the pop-top lid. In such an analysis, the ANTITHESIS between spans 2 and 3 cannot be explicitly represented, because unit 3 is already linked to another unit, unit 1, by means of an ELABORATION relation. In the analysis in Figure 5.2b, the emphasis is on representing the ANTITHESIS between spans 2 and 3. However, in this representation, it is impossible to represent that span 2 describes the context (BACKGROUND) in which the pop-top lid was invented. To explicitly represent all these rhetorical judgments, we would need to represent in the same structure both the BACKGROUND and ANTITHESIS relations, which are considered to be intentional by Mann and Thompson [1988].

In my corpus studies, I also came across cases in which I wanted to represent simultaneously two informational relations, or one informational and one intentional relation, as in the case of the example used by Moore and Pollack [1992]. My analyses suggest that the impossibility of representing explicitly non-isomorphic structures is not due to the difference between intentional and informational relations, but to the representational constraints of the tree structure that we chose as underlying discourse model. Future research will have to provide means for relaxing the assumption that discourse structures are trees, in order to enable one textual unit to be related to more than one unit in the formal representation of text.

5.3 Summary

In this part of the book, I have provided a first-order formalization of valid text structures and a proof theory for deriving such structures. The formalization relies on five essential features:

1. The elementary units of complex text structures are nonoverlapping spans of text.

2. Rhetorical, coherence, and cohesive relations hold between textual units of various sizes. These relations can be simple, extended, and exclusively disjunctive.

3. Some textual units play a more important role in text than others.

4. The abstract structure of most texts is a tree-like structure.

5. If a relation R holds between two textual spans of a tree structure of a text, that relation can be explained by a similar relation that holds between the most important units of the constituent spans. The most important units are determined recursively: they correspond to the union of the most important units of the immediate subspans when the relation that holds between these subspans is paratactic, and to the most important units of the nucleus subspan when the relation that holds between the immediate subspans is hypotactic.

The formalization and the algorithms proposed here enable an analyst or computer program to determine whether a discourse structure is valid with respect to the axiomatization and proof theory given in Chapter 3, and to derive some or all of the valid discourse structures of a text.

5 Discussion

5.1 Related Work

To my knowledge, all approaches to deriving discourse structures that were proposed previously were incremental. That is, they assumed that elementary discourse units are processed sequentially and that a discourse tree is created by incrementally updating a tree structure that corresponds to the discourse units that were processed up to the unit under scrutiny. The unit under scrutiny provides information about the way the updating operation should be performed. These approaches fall into two classes: they are either logic- or grammar-based.

In logic-based approaches [Zadrozny and Jensen, 1991, Lascarides and Asher, 1993, Asher, 1993], the idea of structure is only implicit. Discourse trees can be obtained by considering the coherence relations that hold among the discourse units, which are first-class entities in a logic that captures both the semantics of sentences and the semantics of discourse. Because the logic-based approaches are couched in terms of default logics and logics of beliefs, they are intractable.

In grammar-based approaches [van Dijk, 1972, Polanyi, 1988, Scha and Polanyi, 1988, Gardent, 1994, Hitzeman et al., 1995, Polanyi and van den Berg, 1996, van den Berg, 1996, Gardent, 1997, Schilder, 1997, Cristea and Webber, 1997, Webber et al., 1999], the structure of discourse is explicitly represented; it is assimilated with the parse tree of a sequence of discourse constituents. The first attempts to write discourse grammars [van Dijk, 1972] put very few constraints on the applicability of the rules. However, further developments brought in more and more constraints that were both semantic and structural in nature. The semantic constraints stipulate the conditions that must hold in order to join an incoming discourse unit to an existing discourse structure. For example, in order to substitute a unit on the right frontier[1] of an existing discourse tree with an incoming elementary discourse tree, the semantic information associated with the unit on the right frontier must unify with the semantic information associated with the elementary discourse tree [Gardent, 1997]. The structural constraints are a direct consequence of the assumption that discourse processing is incremental. To account for the sequentiality of text, grammar-based approaches allow only the nodes on the right frontier of a discourse tree to be updated.

Some of the grammar-based approaches to discourse are extensions of context-free and HPSG grammars [van Dijk, 1972, Scha and Polanyi, 1988, Hitzeman et al., 1995]. However, the most recent approaches [Gardent, 1994, van den Berg, 1996, Polanyi and van den Berg, 1996, Gardent, 1997, Schilder, 1997, Cristea and Webber, 1997, Webber

1. The right frontier is the set of nodes of the tree structure that are found on a path from the root to the right-most leaf.

et al., 1999] rely on extensions of tree-adjoining grammars (TAGs) [Joshi, 1987]. The appeal of using TAGs for discourse processing seems to follow from the power of the adjoining operations, which allow trees to be not only expanded, as in the case of context-free grammars, but also rewritten.

In contrast to this work, the formalization that I have presented in Section 3.1.2 provides a mathematical description of the valid text structures, i.e., an expression of the properties of the class of structures that are licensed by the essential features that were put forth in Section 2.1.1 and by the strong compositionality criterion 2.2. As such, the formalization in Section 3.1.2 can be interpreted as a sibling of model-theoretic frameworks that characterize the properties of the syntactic structures of sentences [Keller, 1993, Blackburn et al., 1995, Rogers, 1994, Rogers, 1996]. In contrast with model-theoretic approaches to syntax, the formalization presented in this chapter is much simpler. The constraints on the features of the trees (discourse structures) that the axiomatization of valid text structures captures are much simpler than the constraints that are used by syntactic theories. Because of this, unlike model-theoretic approaches to syntax, which use highly expressive languages with modal operators and second-order quantifiers, the formalization presented here can be couched into the language of first-order logic.

To my knowledge, the formalization of text structures provided in Section 3.1.2 is the first attempt to provide a model-theoretic framework for the study of discourse in general and the study of RST in particular. The proof theory presented in Section 3.2 differs from previous approaches to discourse parsing because it does not assume the derivation of discourse trees to be incremental. Rather it assumes that one first hypothesizes the rhetorical relations that hold between elementary and non-elementary discourse units, and then derives all discourse structures that are valid with respect to the axiomatization. This approach has both advantages and disadvantages.

Advantages

DEALING WITH EXPECTATIONS AND NONMONOTONICITY. Cristea and Webber [1997] introduced a mechanism that enables the incremental derivation of discourse structures in the presence of expectations. For example, the occurrence of the expression "on one hand" raises the expectation that the discourse will subsequently express some contrasting situation. In spite of this, incremental processing along the lines described in most current grammar-based approaches may be inefficient from a computational perspective. Consider example 5.1, which is reproduced from [Cristea and Webber, 1997].

[Because John is such a generous man^{A10}][—whenever he is asked for money,^{B10}] [he will give whatever he has, for example^{C10}][—he deserves the "Citizen of the Year" award.^{D10}] (5.1)

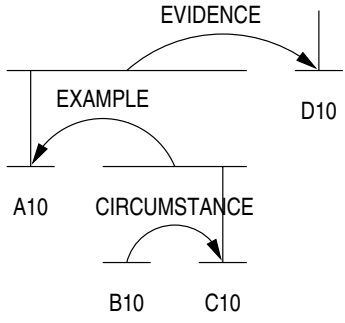

Figure 5.1
The valid text structure of text 5.1

As Cristea and Webber note, the fact that unit B_{10} provides together with unit C_{10} an example for A_{10}, rather than satisfying the expectation raised by "Because", becomes apparent only when unit C_{10} is processed—more specifically, when the discourse marker "for example" is considered. Obviously, in order to accommodate the finding that units B_{10} and C_{10} are an example for the idea presented in the first unit, we have to undo the adjoining of node B_{10}. Therefore, the incremental processing of discourse cannot be monotonic. In order to deal with the nonmonotonicity of incremental discourse derivation, we have to either consider, in the style of Tomita [1985], all possible ways in which a tree can be extended or allow for backtracking. Either approach negatively affects the computational properties of an incremental discourse parser.

Given the discourse parsing paradigm in this book, in order to derive the valid discourse structure of text 5.1, one would first determine that the relations given in 5.2 hold among the elementary units of the text. Then, one would use any of the implementations discussed in Chapter 4 in order to obtain the valid discourse structure shown in Figure 5.1.

$$\begin{cases} \textit{rhet\_rel}(\text{EVIDENCE}, A_{10}, D_{10}) \\ \textit{rhet\_rel}(\text{CIRCUMSTANCE}, B_{10}, C_{10}) \oplus \textit{rhet\_rel}(\text{CIRCUMSTANCE}, B_{10}, D_{10}) \\ \textit{rhet\_rel}(\text{EXAMPLE}, C_{10}, A_{10}) \oplus \textit{rhet\_rel}(\text{EXAMPLE}, C_{10}, B_{10}) \end{cases} \tag{5.2}$$

Since the derivation process starts only after all relations have been hypothesized, one should no longer worry about designing special mechanisms for dealing with expectations and nonmonotonicity.

Dealing with lack of information. In an experiment carried out with Estibaliz Amorrortu and Magdalena Romera [Marcu et al., 1999a], we attempted to build discourse trees incrementally using a discourse annotation tool. The tool allowed us to see only one

sentence at a time. As we identified the discourse units, we attempted to immediately attach them to the right frontier of a growing discourse structure. During the experiment, we noticed that, quite often, we did not have sufficient information in order to decide where to attach the units. The annotating style varied significantly among us: one of us postponed the attachment decisions 37%, one 18%, and one 9% of the time.

During the experiment, we noticed that managing multiple partial discourse trees during the annotation process was the norm rather than the exception. In fact it was not that elementary discourse units were attached incrementally to *one* partial discourse structure, although we did our best to do so, but rather that multiple partial discourse structures were created and then assembled using a rich variety of operations. Moreover, even this strategy proved to be somewhat inadequate, since we needed from time to time to change rhetorical relation labels (2–3% of the operations) and restructure completely the discourse (1–2% of the operations).

This data suggests that it is unlikely that we will be able to build perfect discourse parsers that can incrementally derive discourse trees without applying any form of backtracking. If humans are unable to decide incrementally, in 100% of the cases, where to attach elementary discourse units, it is unlikely we can build computer programs that are.

The approach to discourse parsing proposed in this part of the book does not suffer from this problem, since a human or automatic process has access to all units in order to hypothesize rhetorical relations.

Disadvantages

DEALING WITH RHETORICAL AMBIGUITY. Just as a syntactic parser may generate hundreds of thousands of correct syntactic trees, so can the model theoretic implementations generate hundreds of thousands of valid discourse trees. Finding appropriate mechanisms for dealing with this ambiguity is not a trivial issue, because in the approach taken here, it is treated globally. In contrast, incremental parsing models seem to suffer less from this problem because disambiguation decisions can be made locally, during the discourse structure derivation process.

PSYCHOLINGUISTIC PLAUSIBILITY. The nonincremental paradigm that I presented in this part of the book is not psycholinguistically plausible—after all, humans do process text in an incremental fashion. Given the psychological constraints and the limited resources that humans have, it is conceivable that incremental processing is impossible without backtracking—this would be consistent with the mistakes and reinterpretations that are observed in naturally occurring conversations [Hirst et al., 1993, McRoy, 1993]. Yet, a correct computational model of these processes seems to require more sophistication than current incremental models of discourse parsing.

5.2 Open Problems

The main shortcoming of my formal inquiry into the structure of text comes from its simplicity. The formalization discussed in this chapter ignores a wealth of linguistic phenomena that have been shown to be important in discourse understanding. These phenomena include focus, topic, cohesion, pragmatics, etc. Formalizing these linguistic dimensions of text and incorporating them into the formal model presented in this book is a research direction that promises to be extremely rewarding.

By incorporating a representation of intentions into the formal model discussed here, for example, one can investigate formally and computationally the relation between discourse structures and intentions. More precisely, one can determine what intentions are consistent with a given discourse structure and what discourse structures are consistent with a given intention (see [Marcu, 1999b] for such an approach). By enriching the formal model with logical objects that represent the referring expressions in a text (as Asher [1993], for example), one can also study the relation between discourse structure and reference.

All these are reasonable extensions, which can be easily couched into the formalism discussed here. But even if we end up extending the formalism to cover phenomena such as intentions, referring expressions, and focus, we can still attack some of the assumptions on which the formalization relies. For example, the formalization assumes that text can be sequenced into elementary units—however, as we will see in the next chapter, providing an objective definition for such units is not trivial. Quite often, the boundary between discourse and syntactic phenomena is blurry [Marcu, 1999c]. Eventually, the formalism will need to be refined to a level where the boundary between discourse and syntax becomes transparent (see Section 8.2 for a discussion).

Another assumption that can be challenged concerns the tree-like representation of discourse, which does not permit an explicit representation of multiple relations that hold between various textual units. For example, consider the text shown in 5.3, below, in which we have labeled not the elementary units, but the discourse segments of interest.

[According to engineering lore, the late Ermal C. Fraze, founder of Dayton Reliable Tool & Manufacturing Company in Ohio, came up with a practical idea for the pop-top lid after attempting with halting success to open a beer can on the bumper of his car.[1]] [For decades, inventors had been trying to devise a can with a self-contained opener. Their elaborate schemes had proved unworkable because they required complex manufacturing steps for the attachment of the pull tab—the element that exerts force to open the can top.[2]] [Fraze succeeded because he conceived of a simple and economical rivet to hold the tab in place. Unlike in previous approaches, the rivet was formed from the surface of the can top itself.[3]] (5.3)

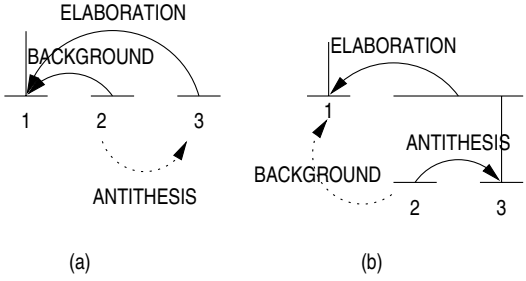

Figure 5.2
Limitations of the tree-like representation of discourse structures

One possible high-level rhetorical analysis of text 5.3 is that shown in Figure 5.2a. In this analysis, the emphasis is on representing explicitly that span 2 describes the historical context (BACKGROUND) in which Ermal Fraze discovered the pop-top lid (sentence 1); and that span 3 ELABORATES on sentence 1 by giving more details about the pop-top lid. In such an analysis, the ANTITHESIS between spans 2 and 3 cannot be explicitly represented, because unit 3 is already linked to another unit, unit 1, by means of an ELABORATION relation. In the analysis in Figure 5.2b, the emphasis is on representing the ANTITHESIS between spans 2 and 3. However, in this representation, it is impossible to represent that span 2 describes the context (BACKGROUND) in which the pop-top lid was invented. To explicitly represent all these rhetorical judgments, we would need to represent in the same structure both the BACKGROUND and ANTITHESIS relations, which are considered to be intentional by Mann and Thompson [1988].

In my corpus studies, I also came across cases in which I wanted to represent simultaneously two informational relations, or one informational and one intentional relation, as in the case of the example used by Moore and Pollack [1992]. My analyses suggest that the impossibility of representing explicitly non-isomorphic structures is not due to the difference between intentional and informational relations, but to the representational constraints of the tree structure that we chose as underlying discourse model. Future research will have to provide means for relaxing the assumption that discourse structures are trees, in order to enable one textual unit to be related to more than one unit in the formal representation of text.

5.3 Summary

In this part of the book, I have provided a first-order formalization of valid text structures and a proof theory for deriving such structures. The formalization relies on five essential features:

1. The elementary units of complex text structures are nonoverlapping spans of text.

2. Rhetorical, coherence, and cohesive relations hold between textual units of various sizes. These relations can be simple, extended, and exclusively disjunctive.

3. Some textual units play a more important role in text than others.

4. The abstract structure of most texts is a tree-like structure.

5. If a relation R holds between two textual spans of a tree structure of a text, that relation can be explained by a similar relation that holds between the most important units of the constituent spans. The most important units are determined recursively: they correspond to the union of the most important units of the immediate subspans when the relation that holds between these subspans is paratactic, and to the most important units of the nucleus subspan when the relation that holds between the immediate subspans is hypotactic.

The formalization and the algorithms proposed here enable an analyst or computer program to determine whether a discourse structure is valid with respect to the axiomatization and proof theory given in Chapter 3, and to derive some or all of the valid discourse structures of a text.

II THE RHETORICAL PARSING OF FREE TEXTS

Preamble

One of the main results in Part I concerned the resolution of the text structure derivation problem: given a sequence of elementary textual units and a set of exclusively disjunctive rhetorical relations that hold between the units and contiguous spans of units, the model- and proof-theoretic accounts provided means for deriving some or all of the valid text structures of the sequence. Hence, in order to automatically build the valid text structures of an arbitrary text, a rhetorical parser needs only

1. Determine the elementary units of that text;
2. And hypothesize the rhetorical relations that hold between those units.

Depending on the way rhetorical parsers address the specifics of these two tasks, they can be associated with different geometrical points in a hypercube whose axes reflect (at least) the types of knowledge, relations, and approaches that the parsers employ.

TYPE OF KNOWLEDGE. Rhetorical parsers can use one or more of the following types of knowledge: orthographic, cue-phrase-specific, lexical, semantic, pragmatic, domain-specific, and corpus-specific.

TYPE OF RELATIONS. Rhetorical parsers can hypothesize only simple rhetorical relations, i.e., relations that hold between elementary units, or simple and extended rhetorical relations, i.e., relations that hold between spans of elementary units, as well.

TYPE OF APPROACH. Rhetorical parsers can employ manually written, semi-automatically derived, automatically derived, or a combination of these rules.

Figure II.1 depicts graphically the space of choices that correspond to these three dimensions.

In this part of the book, I describe two rhetorical parsers:

• The *cue-phrase-based rhetorical parser*, which corresponds to the grey bullet in Figure II.1, relies primarily on cue phrases and manually written rules in order to identify the elementary units of texts and hypothesize rhetorical relations that hold between units and spans of texts. (In the third part of the book, which presents a discourse-based summarizer, I will show how the performance of this rhetorical parser can be improved by taking advantage of machine learning techniques and human constructed summaries. More precisely, I will show how the rhetorical parser can be trained in order to choose discourse interpretations that are not only valid, but also most likely to yield summaries that are similar to those built by humans.)

• The *decision-based rhetorical parser*, which corresponds to the black bullet in Figure II.1, relies on cue phrases, lexical, syntactic, and semantic information in order to

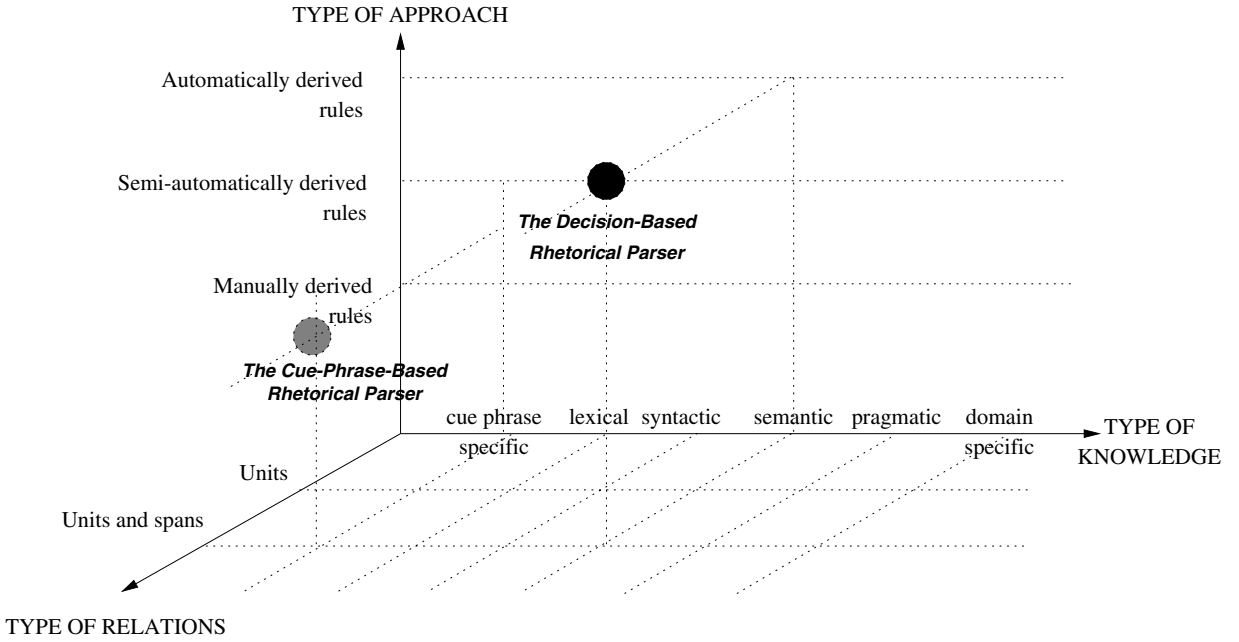

Figure II.1
The space of approaches that characterize the rhetorical parsing process

identify the elementary units of texts and hypothesize rhetorical relations that hold between elementary units and spans of text. All rules applied by this parser are automatically derived, using machine learning techniques.

Part II of the book is organized as follows. In Section 6.1, I argue that a shallow analysis of text that relies primarily on cohesion and knowledge about the way cue phrases like *because, however,* and *in addition* are used can indicate the underlying structure of text. In the rest of Chapter 6, I present an exploratory corpus study of cue phrases and explain how the information extracted from the corpus is used by a cue-phrase-based rhetorical parser that takes as input unrestricted text and produces the rhetorical trees of that text.

In Chapter 7, I present a decision-based rhetorical parser that uses automatically derived rules in order to identify the elementary units of texts and the rhetorical relations that hold between units and spans of texts. The rules are learned from a corpus of discourse trees, whose development is discussed in Section 7.2. The rhetorical parser, which is presented in Section 7.3, employs learning techniques and adopts a shift-reduce parsing model that is well suited for learning.

Part II of the book ends with a discussion of related work on discourse parsing of free texts and empirical research on discourse.

6 Rhetorical Parsing by Means of Manually Derived Rules

6.1 Arguments for a Shallow, Cue-Phrase-Based Approach to Rhetorical Parsing

The results in Part I suggest that one can derive the discourse structure of texts even though one cannot determine unambiguously the rhetorical relations that hold between elementary units and spans of text. In this chapter, I investigate whether one can build correct discourse trees by using only knowledge of connectives and cohesion. Hence, I hypothesize that cue phrases, cohesion, and shallow processing can be used to implement algorithms that determine:

- The elementary units of a text, i.e., the units that constitute the leaves of the discourse 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 intuitively each facet of this hypothesis 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. For the rest of the book, I use the phrases *discourse* and *rhetorical parsing* interchangeably. I also use interchangeably the phrases *cue phrase*, *connective*, and *potential discourse marker*. And I use the phrase *discourse marker* to refer to a cue phrase/connective that has a discourse function, i.e., a cue phrase that signals a rhetorical relation that holds between two text spans.

6.1.1 Determining the Elementary Units of Text Using Cue Phrases and Shallow Processing

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 that range from syntactic parsing to information packaging. For example, if a sentence consists of three arguments that are 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.

Also, in a corpus study that is to be described in Section 6.2, I have 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 6.1 are those enclosed within square brackets.

[*Although* Brooklyn College does not yet have a junior-year-abroad program,]
[a good number of students spend summers in Europe.] (6.1)

Difficulties Obviously, by relying only on orthography, connectives, and shallow processing it is unlikely that one will be capable of determining correctly 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 in order 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 6.2, the first *and* is sentential, i.e., it makes a semantic contribution to the interpretation of the complex nounphrase "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.

[John *and* Mary went to the theatre] [*and* saw a nice play.] (6.2)

If we are to use connectives to determine the elementary units of texts, we need to figure out that we have to insert a boundary before the second occurrence of *and* (the occurrence that has a discourse function), but that we have to insert no boundary before the first occurrence. Obviously, it seems that shallow processing is insufficient to properly solve this problem. But still, it is an open question to what degree shallow processing and knowledge about connectives can be used successfuly in order to determine the elementary units of texts. Using only such lean knowledge resources, should one expect to determine elementary unit boundaries with 10% or 80% accurracy? As our results show (see Section 6.3.3), the latter tends to apply.

6.1.2 Using Cohesion in Order to Determine Rhetorical Relations

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 of Harabagiu, Moldovan, and Maiorano [1996, 1999] show that cohesion can be used to determine rhetorical relations that hold between smaller discourse constituents as well. For example, if a sentence talks about "vegetables" and another sentence talks about "carrots" and "beats," it is possible that a rhetorical relation of ELABORATION holds between the two sentences because "carrots" and "beats" are kinds of "vegetables."

Difficulties In this chapter, 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., that 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 6.1.2, but does not follow from them. Section 6.3.7 evaluates empirically the impact that this assumption has on the problem of rhetorical structure derivation.

6.1.3 Using Cue Phrases/Connectives in Order to Determine Rhetorical Relations

Pro Arguments According to Crystal, the term "connective" is used "to characterize words or morphemes whose function is primarily to link linguistic units at any level" [Crystal, 1991, p. 74]. In other words, the primary function of connectives is to structure the discourse. There are two main reasons to use connectives in order to determine rhetorical relations that hold between spans of texts.

1. First of all, 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 in order to determine rhetorical relations that hold both between elementary units and between large spans of text.

2. Second, 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.[1] More importantly, the absence of markers correlates with a preference by readers to interpret the unmarked textual units as continuations of the topics of the units that precede them [Segal et al., 1991]. Hence, when there is no connective between two sentences, for example, it is likely that the second sentence elaborates on the first.

The facet of connectives that I explore in this part of the book is consistent with the position of Caron, who advocates that "rather than conveying information about states of things, connectives can be conceived as procedural instructions for constructing a semantic representation" [Caron, 1997, p. 70]. Among the three procedural functions of segmentation, integration, and inference that are used by Noordman and Vonk [1997] in order to study the role of connectives, I will concentrate primarily on the first two. That is, I will investigate how one can use connectives to determine the elementary units of texts (the segmentation part) and to determine the rhetorical relations among them (the integration part). The derivation of a valid discourse structure can be interpreted as an inferential process that is structural in nature.

Difficulties The arguments enumerated above support the idea that connectives can be used in order to determine the rhetorical structure of text. More precisely, these 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 *have the potential* of being useful for the purpose of determining the rhetorical structure of texts. Unfortunately, they cannot be used straightforwardly, because they are three-way ambiguous.

- In some cases, connectives have a sentential function, while, in other cases, they have a discourse function. And unless we can determine when a connective has a discourse function, we cannot use connectives to hypothesize rhetorical relations.
- Connectives do not signal explicitly the size of the textual spans that they relate.

1. A corpus of instructional texts that was studied by Moser and Moore [2000] 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.

• 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 now these three problems in turn.

SENTENTIAL AND DISCOURSE USAGES 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 usages in about 80% of cases and that these results can be improved if one uses 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 coverage. The corpus analysis is discussed in Section 6.2. The algorithm that determines elementary unit boundaries and identifies discourse usages of cue phrases is discussed in Section 6.3.3.

DISCOURSE MARKERS ARE AMBIGUOUS WITH RESPECT TO THE SIZE OF THE SPANS THAT THEY CONNECT AND THE RHETORICAL RELATIONS THEY SIGNAL. As I discussed in Section 2.4, connectives cannot be used to determine exactly the types of rhetorical relations and the elementary units and discourse segments that are connected by such relations. However, we have seen that this is not a problem. The mathematical model that constitutes the foundation of this work is sufficiently constrained to handle exclusively disjunctive hypotheses; and as it will become apparent in Section 6.3.4. connectives and cohesion can be used for generating such hypotheses. For example, although the hypotheses 2.11 that can be generated for text 2.10 on the basis of cue phrase occurrences and cohesion are ambiguous, there is only one valid rhetorical structure that can be built for text 2.10, that shown in Figure 6.1.

Note, for example, that the CONTRAST relation that holds between spans [A_8, B_8] and [C_8, D_8] is explained/determined by the simple rhetorical relation *rhet_rel*(CONTRAST, A_8, C_8), which is one of the exclusive disjuncts shown in 2.11; hence, the rhetorical structure in Figure 6.1 is consistent with the compositionality criterion. Note also that the hypothesis *rhet_rel*(ELABORATION, D_8, B_8), for example, cannot be used instead of the CONTRAST relation to link spans [A_8, B_8] and [C_8, D_8], because the relation *rhet_rel*(ELABORATION, D_8, C_8) was used in order to link units C_8 and D_8 and because relations *rhet_rel*(ELABORATION, D_8, B_8) and *rhet_rel*(ELABORATION, D_8, C_8) are exclusively disjunctive.

In fact, even though one could have hypothesized a different relation R to hold, let's say, between the satellite D_8 and the nucleus B_8, such a hypothesis would not yield other valid trees because such trees would violate the compositionality criterion. Relation R cannot be used to link spans [A_8, B_8] and [C_8, D_8], for example, because units B_8 and D_8 are not in

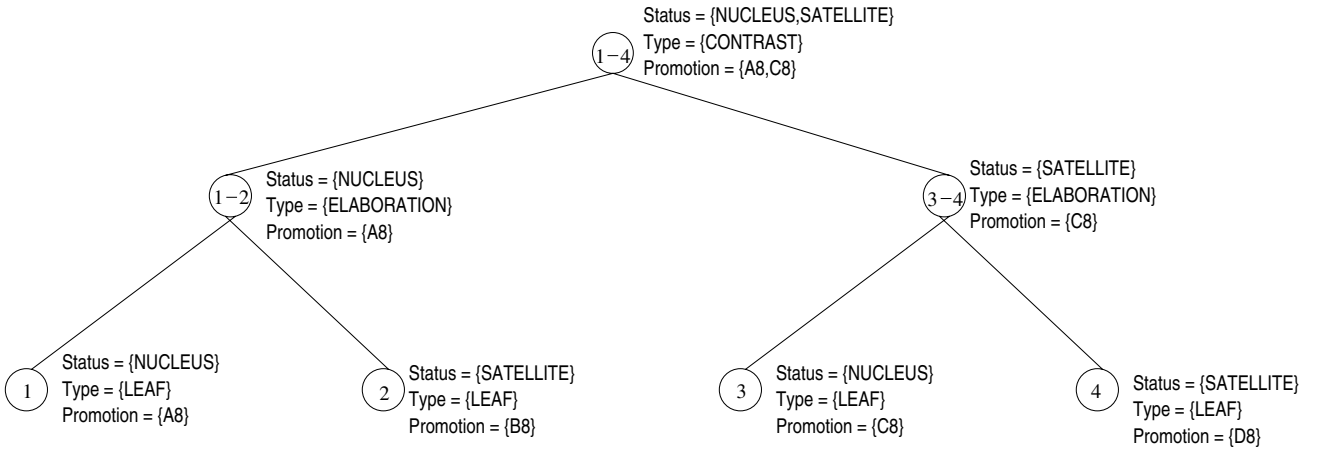

Figure 6.1
A valid rhetorical structure representation of text 2.10, which makes explicit the status, type, and promotion units that characterize each node

the promotion sets of spans [A_8, B_8] and [C_8, D_8], respectively. And there is no combination of rhetorical relations that would promote units B_8 and D_8 as salient in spans [A_8, B_8] and [C_8, D_8] respectively.

Hence, although we were not able to hypothesize precisely the spans and units between which the CONTRAST relation signaled by *In constrast* 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.

6.2 A Corpus Analysis of Cue Phrases

6.2.1 Motivation

The discussion in Section 6.1 suggests that in spite of their ambiguity, cue phrases may be used as a sufficiently accurate indicator of the boundaries between elementary textual units and of the rhetorical relations that hold between them. Unfortunately, although cue phrases have been studied extensively in the linguistic and computational linguistic literature, previous empirical studies did not provide enough data concerning the way cue phrases can be used to determine the elementary textual units that are found in their vicinity and to hypothesize rhetorical relations that hold among them. To overcome this lack of data, I designed an exploratory, empirical study of my own. In the rest of this section, I describe the annotation schema that I used in the study. In Section 6.3, I will explain how the annotated data was used in order to derive algorithms that identify connective occurrences (Section 6.3.2), that determine elementary units of discourse and determine which connectives have a discourse function (Section 6.3.3), and that hypothesize rhetorical relations that hold between elementary units and spans of texts (Section 6.3.4).

6.2.2 Materials

Many researchers have published lists of potential markers and cue phrases [Halliday and Hasan, 1976, Grosz and Sidner, 1986, Martin, 1992, Hirschberg and Litman, 1993, Knott, 1995, Fraser, 1990, Fraser, 1996]. I took the union of their lists and created a set of more than 450 cue phrases. For each cue phrase, I then used an automatic procedure that extracted from the Brown corpus a random set of text fragments that each contained that cue. My initial goal was to select ten text fragments for each occurrence of a cue phrase that was found at the beginning of a paragraph or sentence, and twenty fragments for the occurrences found in the middle and at the end of sentences. The rationale for this choice was the observation that the cue phrases located at the beginning of sentences and paragraphs seemed to exhibit more regular patterns of usage than those found in the middle or at the end of sentences.

On average, I selected approximately seventeen text fragments per cue phrase, having few texts for the cue phrases that do not occur very often in the corpus and up to sixty for cue phrases such as *and*, which I considered to be highly ambiguous. Overall, I randomly selected more than 7600 texts. Marcu [1998a] provides a complete list of the cue phrases that were used to extract text fragments from the Brown corpus, the number of occurrences of each cue phrase in the corpus, and the number of text fragments that were randomly extracted for each cue phrase.

Each text fragment that was extracted from the corpus contained a "window" of approximately 300 words and an occurrence of the cue phrase that was explicitly marked. The cue phrase occurrence was located approximately 200 words from the beginning of the text fragment. The text fragments that were extracted from the corpus were exported into a relational database. In addition to the text fragments, which were stored in a field having the name "Example," the database also contained a number of fields that codified two types of information.

DISCOURSE-RELATED INFORMATION. This information concerned the cue phrase under scrutiny; the rhetorical relations that were marked by the cue phrase; the statuses of the related spans (nucleus or satellite); the textual types of the related spans (from clause-like units to multiple paragraphs); the distance in clause-like units and sentences between the related spans, etc. Section 6.2.3 will describe in detail the semantics of each of the fields in this category: "Marker," "Usage," "Position," "Right boundary," "Where to link," "Rhetorical relation," "Statuses," "Types of textual units," "Clause distance," "Sentence distance," and "Distance to salient unit."

In the examples I annotated, a discourse marker always signaled one rhetorical relation. However, in some cases, I had to annotate complex markers, such as *and although*, which are obtained by concatenating a set of simple markers. Complex markers may signal multiple rhetorical relations, each relation involving different textual units with different rhetorical statuses. In order to account for these cases, the fields "Where to link$_i$", "Rhetorical relation$_i$", "Statuses$_i$", "Types of textual units$_i$", "Clause distance$_i$", "Sentence distance$_i$", and "Distance to salient unit$_i$" were indexed. Because the largest number of relations that were explicitly signalled in our corpus was four, we used field names in which $1 \leq i \leq 4$.

In the cases in which a cue phrase signalled a rhetorical relation that held between the textual unit that contained the cue phrase and a textual unit that came after, I considered it useful to also encode explicitly information pertaining to the rhetorical relation that holds between the textual unit that contains the cue phrase and the text that precedes it. The purpose of this enterprise was to investigate whether there exists a correlation between the

Table 6.1
The fields from the corpus that were used in developing the algorithms discussed in the rest of Chapter 6

Algorithm	Field names
The clause-boundary and discourse-marker identification algorithm (Section 6.3.3)	"Marker," "Usage," "Position," "Right boundary," "Break action"
The discourse-marker-based algorithm for hypothesizing rhetorical relations (Section 6.3.4)	"Marker," "Usage," "Where to link," "Rhetorical relation," "Statuses," "Types of textual units," "Clause distance," "Sentence distance," "Distance to salient unit"

markers that "link forward" and the preceding text. For example, in the text in Figure 6.3, which is discussed below, the marker *Although* signals a rhetorical relation of CONCESSION that holds between the clauses "*Although* faculties insist on governing themselves," and "they grant little prestige to a member who actively participates in college or university government." Obviously, the marker does not signal explicitly any relation between the sentence that contains it and the previous text. Nevertheless, in addition to fully describing the CONCESSION relation, I also described the relation between the sentence that contained the marker *Although*, and the text that precedes it. In the case of the text in Figure 6.3, this relation is one of ELABORATION on the rhetorical question "How well do faculty members govern themselves?"

ALGORITHMIC INFORMATION. In contrast to the discourse related information, which has a general linguistic interpretation, the algorithmic information was specifically tailored to the surface analysis that aimed at determining the elementary textual units of a text. This information involved only one field, called "Break action."
Hence, the initial database contained more than 7600 records, each corresponding to a text fragment. The field "Example" was the only field that was automatically generated. All the other fields were initially empty.

The information in the fields associated with each text fragment and cue phrase constitutes the empirical foundation of two algorithms: an algorithm that identifies elementary unit boundaries and discourse usages of cue phrases; and an algorithm that hypothesizes rhetorical relations that hold among textual units. Table 6.1 enumerates explicitly the fields that were used in developing each of these algorithms.

6.2.3 Requirements for the Corpus Analysis

Once the database was created, each field of each record in the database was updated according to the requirements described below.

One of the early strikes called by the AWOC was at the DiGiorgio pear orchards in Yuba County. We found that a labor dispute existed, and that the workers had left their jobs, which were then vacant because of the dispute. Accordingly, under clause (1) of the Secretary's Regulation, we suspended referrals to the employer. (Incidentally, no Mexican nationals were involved.) The employer, seeking to continue his harvest, challenged our right to cease referrals to him, and sought relief in the Superior Court of Yuba County. The court issued a temporary restraining order, directing us to resume referrals. We, of course, obeyed the court order. However, the Attorney General of California, at the request of the Secretary of Labor, sought to have the jurisdiction over the issue removed to the Federal District Court, on grounds that it was predominantly a Federal issue since the validity of the Secretary's Regulation was being challenged. [However, the Federal Court held] [that since the State had accepted the provisions of the Wagner-Peyser Act into its own Code,] [and presumably therefore also the regulations,] [it was now a State matter.] [It {\em accordingly} refused to assume jurisdiction,] [whereupon the California Superior Court made the restraining order permanent.] Under that order, we have continued referring workers to the ranch. A similar case arose at the Bowers ranch in Butte County, and the Superior Court of that county issued similar restraining orders.

The growers have strenuously argued that I should have accepted the Superior Court decisions as conclusive and issued statewide instructions to our staff to ignore this provision in the Secretary's Regulation.

Figure 6.2
A text fragment containing the cue phrase *accordingly*

Example The field "Example" contains one text fragment that was randomly extracted from the Brown corpus for a given cue phrase. The cue phrase under consideration is explicitly marked using the LaTeX macro for emphasizing text, {\em }, as shown, for example, in the text in Figure 6.2.

In the cases in which the cue phrase under scrutiny has a discourse function, the elementary textual units that are found in the neighborhood of the cue phrase are enclosed within square brackets. The number of textual units that are enclosed within square brackets depends on the kind of relation that the cue phrase marks. If it marks a relation between two clauses of the same sentence, only those clauses will be enclosed within square brackets. However, if it marks a relation between two elementary textual units that are a couple of sentences apart, then all the elementary units in between are each enclosed within square brackets. And if it marks a relation between two textual spans that are not elementary, then all the elementary units that are contained in the nonelementary units are each enclosed within square brackets as well. For example, if the marker under scrutiny in the text in Figure 6.2 is *accordingly*, we will bracket 6 elementary units, because *accordingly*, marks a VOLITIONAL-CAUSE relation between the units "[However, the Federal Court held] [it was now a State matter.]" and "[It *accordingly* refused to assume jurisdiction]".

The president expects faculty members to remember, in exercising their autonomy, that they share no collective responsibility for the university's income nor are they personally accountable for top-level decisions. He may welcome their appropriate participation in the determination of high policy, but he has a right to expect, in return, that they will leave administrative matters to the administration.

[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.]

[{\em Although} faculties insist on governing themselves,] [they grant little prestige to a member who actively participates in college or university government.] There are, nevertheless, several things that the president can do to stimulate participation and to enhance the prestige of those who are willing to exercise their privilege. He can, for example, present significant university-wide issues to the senate. He can encourage quality in faculty committee work in various ways: by seeing to it that the membership of each committee represents the thoughtful as well as the action-oriented faculty; by making certain that no faculty member has too many committee assignments; by assuring good liaison between the committees and the administration; by minimizing the number of committees.

Figure 6.3
A text fragment containing the cue phrase *Although*

The field "Example" in Figure 6.3 depicts all the elementary units between the question "How well do faculty members govern themselves?" and the sentence that contains the cue phrase under scrutiny *Although*, because the sentence containing the cue phrase elaborates on the question.

The elementary textual units enclosed within square brackets are not necessarily clauses in the traditional, grammatical sense. Rather, they are contiguous spans of text that can be smaller than a clause and that can provide grounds for deriving rhetorical inferences. For example, although the text in 6.3 corresponds to a single clause, I decided to break it into two elementary textual units because the first part, "Because of light leakage from one ultraviolet source to another," is explicitly marked by a cue phrase that enables one to infer that the explanation for the fact that the lights are switched is the light leakage.

[*Because of* light leakage from one ultraviolet source to another,] [the lights are switched by a commutator-like assembly rotated by a synchronous motor.] (6.3)

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), I assigned those spans an elementary unit status, although in some cases they were not fully fleshed clauses. In the rest of the book, I use the term *clause-like unit* in order to refer to such elementary units.[2]

Marker The field "Marker" encodes the orthographic environment of the cue phrase. That is, it contains the marker under consideration and all the punctuation marks that precede or follow it. If more than one cue phrase is used, the "Marker" field contains the adjacent markers as well. For example, for the text in Figure 6.2, the "Marker" environment will contain the string "␣accordingly␣" because no punctuation marks or cue phrases surround the cue phrase under scrutiny.[3] However, if the cue under scrutiny had been the phrase "However", from the sentence that precedes the one that contains the string "accordingly", the "Marker" field would have been ".␣However,␣", because the phrase is preceded by a period and followed by a comma. The beginning of a paragraph is conventionally labelled with a # character. Hence, the "Marker" field associated with text in Figure 6.3 is "#␣Although␣".

Usage The field "Usage" encodes the functional role of the cue phrase. The role can be one or more of the following:

• SENTENTIAL (S), when the cue phrase has no function in structuring the discourse. For example, in text 6.4, *above all* is used purely sententially: *above* is a preposition and *all* is a quantifier.

And finally, the best part of all, simply sit at the plank table in the kitchen with a bottle of wine and the newspapers, reading the ads as well as the news, registering nothing on her mind but letting her soul suspend itself *above all* wishing and desire. (6.4)

• DISCOURSE (D), when the cue phrase signals a discourse relation between two textual units. For example, in text 6.1, *Although* signals a concession relation between two clauses of the same sentence; the clauses are enclosed within square brackets.

• PRAGMATIC (P), when the cue phrase signals a relationship between some linguistic or nonlinguistic construct that pertains to the unit in which the cue phrase occurs and the beliefs, plans, intentions, and/or communicative goals of the speaker, hearer, or some character depicted in the text. In this case, the beliefs, plans, etc., might not be explicitly

2. See [Marcu, 1999c] for a refined notion of elementary discourse unit.

3. The symbol ␣ denotes a blank character.

stated in discourse; rather, it is the role of the cue phrase to help the reader infer them.[4] For example, in text 6.5, *again* presupposes that James was caught by the police before, but that event is not explicitly mentioned in the discourse. In this sense, one can say that there exists a relationship between sentence 6.5 and the speaker's knowledge and that *again* provides the means through which the hearer can infer that knowledge.

James was caught by the police *again*. (6.5)

In text 6.6, *already* is used to express an element of unexpectedness with respect to the events that are described. Because of this, we say that *already* plays a pragmatic role as well.

When May came the Caravan had *already* crossed the Equator. (6.6)

Right Boundary The field "Right boundary" contains a period, question mark, or exclamation mark if the cue phrase under scrutiny occurs in the last elementary unit of a sentence. If it does not occur in the last elementary unit, it contains the cue phrase and orthographic marker found at the beginning of the elementary unit that follows it. If there is no cue phrase or orthographic marker found at the boundary between the two units, the "Right boundary" field contains the first word of the unit that follows the one that contains the marker. For example, the content of the field "Right boundary" for the text in Figure 6.2 is ",␣*whereupon*␣" because "," and *whereupon* are the lexemes found at the boundary between the unit that contains the marker under scrutiny and the next unit in the text. The content of the field "Right boundary" associated with the text in Figure 6.3, text 6.1, and cue phrase *Although* is "," because the first lexeme in the second elementary unit of each text is not a cue phrase.

Where to link$_i$ The field "Where to link$_i$" describes whether the textual unit that contains the discourse marker under scrutiny is related to a textual unit found BEFORE (B) or AFTER (A) it. For example, the textual unit that contains the marker *accordingly* in the text in Figure 6.2 is rhetorically related to a textual unit that goes before it (B). In contrast, the clause that contains the discourse marker *Although* in text 6.1 is rhetorically related to the clause that comes immediately after it (A).

4. The definition of pragmatic connective that I use here is that proposed by Fraser [1996]. It should not be confused with the definition proposed by van Dijk [1979], who calls a connective "pragmatic" if it relates two speech acts and not two semantic units.

Types of textual units$_i$ The field "Types of textual units$_i$" describes the types of textual units that are connected through a rhetorical relation that is signalled by the marker under scrutiny. The types of the textual units range over the set {CLAUSE-LIKE UNIT (C), MULTICLAUSE-LIKE UNIT (MC), SENTENCE (S), MULTISENTENCE (MS), PARAGRAPH (P), MULTIPARAGRAPH (MP)}. The field contains two types that are separated by a semicolon: the first type corresponds to the first textual unit, and the second type corresponds to the second textual unit. For example, the "Types of textual units$_1$" field that corresponds to the marker *accordingly* in the text in Figure 6.2 is MC;C because it relates the multiclause-like unit "[However the Federal Court held] [it was now a State matter]" with the clause "[It *accordingly* refused to assume jurisdiction]". The "Types of textual units$_1$" field that corresponds to the marker *Although* in text 6.1 is C;C because it relates two clauses: "[*Although* Brooklyn College does not yet have a junior-year-abroad program,]" and "[a good number of students spend summers in Europe.]".

Clause distance$_i$ The field "Clause distance$_i$" contains a count of the clause-like units that separate the units that are related by the discourse marker. The count is 0 when the related units are adjacent. For example, the fields "Clause distance$_1$" for both the text in Figure 6.2 and text 6.1 have value 0.

Sentence distance$_i$ The field "Sentence distance$_i$" contains a count of the sentences that are found between the units that are related by the discourse markers. The count is -1 when the related units belong to the same sentence. For example, the field "Sentence distance$_1$" for the text in Figure 6.2 has value 0. However, the field for text 6.1 has value -1.

Distance to salient unit$_i$ The field "Distance to salient unit$_i$" 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 6.7 below, with respect to the cue phrase *for example*.

[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]] (6.7)

A rhetorical analysis of text 6.7 is shown in Figure 6.4. 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$_i$" reflects this state of affairs. For text 6.7 and marker *for example*, its value is 0.

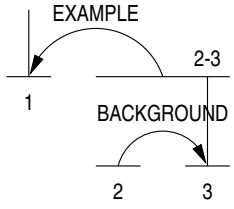

Figure 6.4
The discourse tree of text 6.7

Position$_i$ The field "Position$_i$" specifies the position of the discourse marker under scrutiny in the textual unit to which it belongs. The possible values taken by this field are: BEGINNING (B), when the cue phrase occurs at the beginning of the textual unit to which it belongs; MIDDLE (M), when it is in the middle of the unit; and END (E), when it is at the end. For example, the content of the field "Position$_1$" for the text in Figure 6.2 is M. However, the content of the field "Position$_1$" for text 6.1 is B.

Statuses$_i$ The field "Statuses$_i$" specifies the rhetorical statuses of the textual units that are related through a rhetorical relation that is signalled by the cue phrase under scrutiny. The status of a textual unit can be NUCLEUS (N) or SATELLITE (S). The field contains two rhetorical statuses that are separated by a semicolon: the first status corresponds to the first textual unit, and the second to the second. For example, the "Statuses$_1$" field for the marker *accordingly* in the text in Figure 6.2 is S;N because the multiclause-like units "[However the Federal Court held] [it was now a State matter]" are the SATELLITE and the clause "[It *accordingly* refused to assume jurisdiction]" is the NUCLEUS of a rhetorical relation of VOLITIONAL-CAUSE. The "Statuses$_1$" field for the marker *Although* in text 6.1 is S;N because it relates two clauses: "[*Although* Brooklyn College does not yet have a junior-year-abroad program,]" is the SATELLITE and "[a good number of students spend summers in Europe]" is the NUCLEUS of a rhetorical relation of CONCESSION.

Rhetorical relation$_i$ The field "Rhetorical relation$_i$" specifies one or more rhetorical relations that are signalled by the cue phrase under scrutiny. The list of relations that is used was derived from the list of relations initially proposed by Mann and Thompson [1988]. A new relation was added to Mann and Thompson's list whenever I came across an example for which none of the relations held. In the case in which more than one rhetorical relation definition seemed to adequately characterize the example under consideration, the field "Rhetorical relations$_i$" enumerated all these relations. For example, the contents of the "Rhetorical relation$_1$" field for the text in Figure 6.2 and text 6.1 are VOLITIONAL-CAUSE and CONCESSION, respectively. Overall, I used a set of fifty-four rhetorical relations.

Break action The field "Break action" contains one member of a set of instructions for a shallow analyzer that determines the elementary units of a text. The shallow analyzer assumes that text is processed in a left-to-right fashion and that a set of flags monitors the segmentation process. Whenever a cue phrase is encountered, the shallow analyzer executes an action from the set {NOTHING, NORMAL, COMMA, NORMAL_THEN_COMMA, END, MATCH_PAREN, COMMA_PAREN, MATCH_DASH, SET_AND, SET_OR, DUAL}. The effect of these actions can be one of the following:

• Create an elementary textual unit boundary in the input text stream. Such a boundary corresponds to the square brackets used in the examples that were discussed so far.

• Set a flag. Later, if certain conditions are satisfied, this may lead to the creation of a textual unit boundary.

Since a discussion of the semantics of the actions is meaningless in isolation, I will provide it below in Section 6.3.3, in conjunction with the clause-like unit boundary and 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 in order 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. However, this encoding does not preclude the employment of more sophisticated methods that derive algorithmic information automatically.

6.2.4 Method and Results

Once the database was created, I analyzed each record in it and updated its fields according to the requirements described in Section 6.2.3. Tables 6.2 and 6.3 show the information that I associated with the fields when I analyzed the text fragments shown in Figures 6.2 and 6.3 respectively.

Overall, I have manually analyzed 2100 of the text fragments in the corpus. Of the 2100 instances of cue phrases that I considered, 1197 had a discourse function, 773 were sentential, and 244 were pragmatic.[5] I annotated only 2100 fragments because the task was too time consuming to complete.

The taxonomy of relations that I used to label the 1197 discourse usages in the corpus contained 54 relations. This taxonomy is larger than the taxonomy of 24 relations which

5. The three numbers add up to more than 2100 because some cue phrases had multiple roles in some text fragments.

Table 6.2
A corpus analysis of the segmentation and integration function of the cue phrase *accordingly* from the text in Figure 6.2.

Field	Content
Example	See Figure 6.2
Marker	␣accordingly␣
Usage	D
Right boundary	,␣whereupon␣
Where to link$_1$	B
Types of textual units$_1$	MC;C
Clause distance$_1$	0
Sentence distance$_1$	0
Distance to salient unit$_1$	−1
Position$_1$	M
Statuses$_1$	S;N
Rhetorical relation$_1$	VOLITIONAL-CAUSE
Break action	NOTHING

was proposed initially by Mann and Thompson [1988]. The reason for this is that, during the corpus analysis, it often happened that none of the relations proposed by Mann and Thompson seemed to capture well enough 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. In order to manage the new relations, I did not provide for them definitions similar to those proposed by Mann and Thompson [1988]; I only kept a list of text examples that I considered to reflect the meaning of each new rhetorical relation that I introduced. Table 6.4 lists the most frequent relations that were used during the annotation. For a complete list, see Marcu [1998a].

In addition to the information above, I have extracted from the corpus, for each cue phrase, information that enables:

- Its recognition in text.
- The determination of the boundaries of the elementary textual units found in its vicinity.
- The hypothesizing of rhetorical relations that hold among textual units found in its vicinity.

These results are discussed in Section 6.3, where I establish the connection between the corpus analysis and the algorithms that derive text structures for unrestricted texts.

Table 6.3
A corpus analysis of the segmentation and integration function of the cue phrase *Although* from the text in Figure 6.3

Field	Content
Example	See Figure 6.3
Marker	#␣Although␣
Usage	D
Right boundary	,
Where to $link_1$	A
Types of $textual\ units_1$	C;C
$Clause\ distance_1$	0
$Sentence\ distance_1$	−1
$Distance\ to\ salient\ unit_1$	−1
$Position_1$	B
$Statuses_1$	S;N
$Rhetorical\ relation_1$	CONCESSION
Where to $link_2$	B
$Types\ of\ textual\ units_2$	S;S
$Clause\ distance_2$	5
$Sentence\ distance_2$	4
$Distance\ to\ salient\ unit_2$	−1
$Position_2$	B
$Statuses_2$	N;S
$Rhetorical\ relation_2$	ELABORATION
Break action	COMMA

Because the corpus analysis has not been fully completed, it would be premature to draw any conclusions with respect to the taxonomy of rhetorical relations. In fact, this problem is beyond the scope of this book. For the moment, I prefer to make no claims with respect to the size and nature of an appropriate taxonomy of rhetorical relations.

6.2.5 Discussion

The main advantage of the corpus analysis described in this section consists in the empirical grounding that it provides to two algorithms that determine the elementary units and derive the valid text structures of unrestricted texts. These algorithms are grounded partially in the empirical data derived from the corpus and partially in the intuitions that I developed during the discourse analysis of the 2100 fragments of text.

Since I was the only analyst of 2100 of the 7600 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

Table 6.4
Distribution of the most frequent fifteen rhetorical relations in the corpus of cue phrases

Relation	Percent
ELABORATION	12.76
JOINT	11.57
SEQUENCE	8.65
CIRCUMSTANCE	8.43
CONTRAST	6.49
EVIDENCE	5.84
CONCESSION	4.11
NONVOLITIONAL-CAUSE	4.00
BACKGROUND	3.78
INTERPRETATION	3.78
ANTITHESIS	3.62
PARENTHETICAL	3.24
EXPLANATION	2.59
CONDITION	2.22
ARGUMENTATION	1.95

of the algorithms was done against data that did not occur in the corpus and that was analyzed independently by other judges.

6.3 A Cue-Phrase-Based Approach to Rhetorical Parsing

6.3.1 Introduction

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 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. And in my own

experiments (see Section 6.3.7), I have noticed that human judges manually built, in nine out of ten cases, 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 an efficiency-related one. Just as sentences are ambiguous and syntactic parsers can derive thousands of syntactic trees, so are texts ambiguous and so can rhetorical parsers derive thousands of rhetorical trees. By assuming that the rhetorical structure of text correlates with sentence, paragraph, and section boundaries, one can reduce significantly the search space of possible rhetorical interpretations and increase the speed of a rhetorical parser.

The Cue-Phrase-Based Rhetorical Parsing Algorithm—A Bird's-Eye View The rhetorical parsing algorithm that we propose in this chapter is outlined in Figure 6.5. The rhetorical parser first determines the set of all 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 6.5), the rhetorical parser retraverses the input and, by using information derived from the corpus study in Section 6.2, it determines the elementary textual units of the text 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 6.5). The 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 in order 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 using a simple cohesion-based method, 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).

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 6.3.5 (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 6.5).

Input: A text T.

Output : The valid text structures of T.

I. Determine the set D of all cue phrases (potential discourse markers) in T.
II. Use information derived from the corpus analysis in order to determine recursively all the sections, paragraphs, sentences, and clause-like units of the text and the set $D_d \in D$ of cue phrases that have a discourse function.
III. For each of the three highest levels of granularity (sentences, paragraphs, and sections):
    III.1 Use information derived from the corpus analysis about the discourse markers D_d in order to hypothesize rhetorical relations among the elementary units that correspond to that level.
    III.2 Use cohesion in order to hypothesize rhetorical relations among the units for which no hypotheses were made in step III.1.
    III.3 Apply one of the algorithms discussed in Chapter 4 in order to determine all the valid text trees that correspond to that level.
    III.4 Assign a weight to each of the text trees and determine the tree with maximal weight.
IV. Merge the best trees that correspond to each level into a discourse tree that spans the whole text and that has clause-like units as its elementary units.

Figure 6.5
Outline of the cue-phrase-based rhetorical parsing algorithm

In the rest of this section, 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 analysis in Section 6.2, 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 as an example the text in Figure 1.1.

6.3.2 Determining the Potential Discourse Markers of a Text

From the Corpus Analysis to the Potential Discourse Markers of a Text The corpus analysis discussed in Section 6.2 provides information about the orthographic environment of cue phrases and the function that 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 6.8, it usually signals a rhetorical relation that holds between the clause-like unit that contains it and the clause or clauses coming after it. However, if

the same cue phrase occurs at the beginning of a sentence and is immediately followed by a comma, as in text 6.9, it usually signals a rhetorical relation that holds between the sentence to which *Besides* belongs and a textual units that precedes it.

[*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.] (6.8)

[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.] (6.9)

I have taken each of the cue phrases 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 and the orthographic environments that characterized the cue phrases that had a discourse role in most of the text fragments in order 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 the cue phrases that had a sentential role in a majority of the instances in the corpus and the cue phrases that were too ambiguous to be exploited in the context of a surface-based approach. In general, I preferred to be conservative and to consider only potential cue phrases whose discourse role could be determined with a relatively high level of confidence. Table 6.5 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.

The regular expressions shown in Table 6.5 obey the conventions used by the Unix tool *lex*. Table 6.6 describes the semantics of the symbols used in Table 6.5. For example, the regular expressions associated with *Although, With* and *Yet* match occurrences that are enclosed by space, tab, or newline characters. The regular expression associated with *for example* matches occurrences that are optionally preceded and optionally followed by a comma. The end of a sentence matches the occurrence of a period, question mark, or exclamation mark; or any of these followed by quotation marks. The beginning of a paragraph is associated with zero or more spaces followed by one of the following:

Table 6.5
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,}))

- A newline and a tab character, followed by zero or more occurrences of spaces and tabs.
- A newline followed by at least two occurrences of space, tab, or newline characters.

By considering only the cue phrases that have 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 less units than humans, hoping that, in this way, most of the identified units will be correct.

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 6.5). 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 the text in Figure 1.1, the algorithm recognizes all punctuation marks and the cue phrases shown in italics in Figure 6.6.

6.3.3 Determining the Elementary Units of a Text

From the Corpus Analysis to the Elementary Textual Units of a Text As I discussed in Section 6.2, 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 eleven actions that constitutes the foundation of an algorithm to determine automatically

Table 6.6
The semantics of the symbols used in Table 6.5

Symbol	Semantics
␣	blank character
\t	tab character
\n	newline character
[*e*]	optional occurrence of expression *e*
( )	grouping
a | *b*	alternative (*a* or *b*)
e+	one or more occurrences of expression *e*
e⋆	zero or more occurrences of expression *e*
e{*n*, }	at least *n* occurrences of expression *e*
" "	enclose special symbols

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.

Figure 6.6
The cue phrases that are automatically identified in the text in Figure 1.1

the elementary units of a text. The algorithm processes a 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 the text. 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 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 instead. 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 usually used for determining the boundaries of parenthetical units. These units, such as those shown in italics in 6.10, 6.11, 6.12, and 6.13 below, are related only to the larger units that they belong to or to the units that immediately precede them.

With his anvillike upper body, McRae might have been tapped for the National Football League instead of the U.S. national weight lifting team if he had not stopped growing at 160 centimeters *(five feet three inches).* (6.10)

With its distant orbit *—50 percent farther from the sun than the Earth—* and slim atmospheric blanket, Mars experiences frigid weather conditions. (6.11)

Yet, even on the summer pole, *where the sun remains in the sky all day long,* temperatures never warm enough to melt frozen water. (6.12)

They serve cracked wheat, oats, or cornmeal. Occasionally, the children find steamed, whole-wheat grains for cereal, *which they call "buckshot."* (6.13)

Because the deletion of parenthetical units does not affect the readability of a text, in the algorithms that I present here I do not assign them an elementary unit status. Such an assignment would only create problems at the formal level, because then discourse trees could no longer be represented as binary trees. Instead, I will only determine the boundaries of parenthetical units and record, for each elementary unit, the set of parenthetical units that belong to it.

• Action SET_AND instructs the analyzer to store the information that the input stream contains the lexeme *and*.

• Action SET_OR instructs the analyzer to store the information that the input stream contains the lexeme *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 6.14 signals a rhetorical relation of CONCESSION between the clause to which it belongs and the previous clause. However, in text 6.15, 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.

[I went to the theater] [*although* I had a terrible headache.] (6.14)

[The trip was fun,] [*and although* we were badly bitten by black flies,] [I do not regret it.] (6.15)

In addition to the algorithmic information that is explicitly encoded in the field "Break action," the shallow analyzer also 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 in the corpus that could play a discourse function is assigned a structure with two features:

Table 6.7
The list of actions that correspond to the potential discourse markers and punctuation marks shown in Table 6.5

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

- 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 6.7 lists the actions and the positions in the elementary units of the cue phrases and orthographic markers shown in Table 6.5.

The Section, Paragraph, and Sentence Identification Algorithm As I discussed in Section 6.3.1, 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. It uses the set of regular expressions that are associated with the potential discourse markers END_SENTENCE and BEGIN_PARAGRAPH found in Table 6.5, and a list of abbreviations, such as *Mr., Mrs.,* and *Inc.*, that prevent the setting of sentence and paragraph boundaries at places that are inappropriate. This simple algorithm located correctly 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 6.3.3. Other texts and semistructured HTML/SGML documents may need more sophisticated algorithms to solve this segmentation problem.

A machine-learning algorithm for identifying the sentence and paragraph breaks, which reflects the approach taken by Palmer and Hearst [1997], is presented in Section 7.3.2.

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 6.7 shows only its skeleton and focuses on the variables and steps that are used in order to determine the elementary units. To avoid overloading the figure, the steps that assert the discourse function of a marker are not shown; however, these steps are mentioned in the discussion of the algorithm that is given below. Marcu [1998a] provides a full description of the algorithm.

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 ∈ status ∨ MATCH_DASH ∈ status ∨ COMMA_PAREN ∈ status
4.          ⟨deal with parenthetical information⟩
5.      if COMMA ∈ status ∧ markerTextEqual(i,",") ∧
6.          NextAdjacentMarkerIsNotAnd() ∧ NextAdjacentMarkerIsNotOr()
7.          ⟨insert textual boundary after comma⟩
8.      if (SET_AND ∈ status ∨ SET_OR ∈ status) ∧ 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 ∪ {COMMA};
14.         case NORMAL_THEN_COMMA: ⟨insert textual boundary before marker⟩
15                                              status := status ∪ {COMMA};
16.         case NOTHING: ⟨assign discourse usage⟩⋆
17.         case MATCH_PAREN, COMMA_PAREN, MATCH_DASH: status := status ∪ {getActionType(i)};
18.         case SET_AND, SET_OR: status := status ∪ {getActionType(i)};
19.     }
20. end for
21. finishUpParentheticalsAndClauses();
```

Figure 6.7
The skeleton of the clause-like unit and discourse-marker identification 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 6.3.2). 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.

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 variable "status," thus influencing the processing of further markers. Lines 3–9 concern actions that are executed when the "status" variable is not NIL. We discuss now in turn each of these actions.

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 6.3.3, to assign no elementary unit status to parenthetical information. Because of this, the text shown in italics in text 6.16, 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 *and*. As example 6.16 shows, taking the first comma as boundary of the parenthetical unit would be inappropriate.

[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.] (6.16)

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 6.17 below.

[I gave John a boat,] [which he liked, and a duck,] [which he didn't.] (6.17)

Nevertheless, the evaluation results discussed in Section 6.3.3 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 *or* marker determines the identification of a new elementary unit (see lines 5–7 in Figure 6.7).

Usually, the discourse role of the cue phrases *and* and *or* is ignored because the surface-form algorithm that we propose in this chapter 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 6.18, *and* acts as an indicator of a LIST relation between the first two clauses of the text.

[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.] (6.18)

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 6.7 is satisfied, the algorithm not only inserts textual boundaries as discussed above, but it 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 variable "status" to COMMA when processing the marker *although* in example 6.19, but

only insert a textual unit boundary when processing the same marker in example 6.20. The final textual unit boundaries that are assigned by the algorithm are shown using square brackets.

[John is a nice guy,] [*but although* his colleagues do not pick on him,] [they do not invite him to go camping with them.] (6.19)

[John is a nice guy,] [*although* he made a couple of nasty remarks last night.] (6.20)

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 variable "status" such that a textual unit boundary will be identified at the first occurrence of a comma. In the case when a marker of type NOTHING is processed, the only action that might be executed is that of assigning that marker a discourse usage.

Lines 7–8 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 6.7 flushes this text into the last clause-like unit that is under consideration.

The clause-like unit boundary and discourse marker identification algorithm has been fully implemented in C++. When it processes the text in Figure 6.6, it determines that the text has ten elementary units and that seven cue phrases have a discourse function. Figure 6.8 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.

Evaluation of the Clause-Like Unit and Discourse-Marker Identification Algorithm

The algorithm shown in Figure 6.7 determines clause-like unit boundaries and identifies discourse usages of cue phrases using methods based on surface form. The algorithm relies heavily on the corpus analysis discussed in Section 6.2.

[*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]]

Figure 6.8
The elementary units determined by the clause-like unit identification algorithm

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) is used as a discourse marker in the majority of the examples in the corpus. The enforcement of this criterion on one hand reduces the recall of the discourse markers that can be detected, but, on the other hand, 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 that is 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 potential 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 5036 words from *Scientific American*.
2. A magazine article of 1588 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

Table 6.8
Evaluation of the marker identification procedure

Text	No. of discourse markers identified manually	No. of discourse markers identified by the algorithm	No. 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%

units. The judges were given no instructions about the criteria that they were to apply in order to determine the clause-like unit boundaries; rather, they were supposed to rely on their intuition and preferred definition of clause. The locations in texts that were labelled 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 determined manually 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 usage. Otherwise, I assigned it a sentential usage. I applied this procedure with respect to instances of all 450 cue phrases in the corpus; and not only with respect 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 on the same texts. The algorithm found 80.8% of the discourse markers with a precision of 89.5% (see Table 6.8), a result that seems to outperform Hirschberg and Litman's [1993].[6] 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, there is a large number of occurrences of the discourse marker *and*. And as we discussed above, the clause-like unit and discourse-marker identification algorithm labels correctly only a small percentage of these occurrences.

6. 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 in the case of Hirschberg and Litman.

Table 6.9
Evaluation of the clause-like unit boundary identification procedure

Text	No. of sentence boundaries	No. of clause-like unit boundaries identified manually	No. of clause-like unit boundaries identified by the algorithm	No. 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%

The algorithm correctly identified 81.3% of the clause-like unit boundaries, with a precision of 90.3% (see Table 6.9). I am not aware of any surface-form algorithms that achieve similar results. Still, the clause-like unit and discourse-marker identification algorithm has its limitations. These are primarily due to the fact that the algorithm relies entirely on cue phrases and orthographic features that can be detected by shallow methods. For example, such methods are unable to classify correctly the sentential usage of *but* in example 6.21; as a consequence, the algorithm incorrectly inserts a textual unit boundary before it.

[The U.S. has] [*but* a slight chance to win a medal in Atlanta,] [because the championship eastern European weight lifting programs have endured in the newly independent countries that survived the fracturing of the Soviet bloc.] (6.21)

6.3.4 Hypothesizing Rhetorical Relations between Textual Units of Various Granularities

From Discourse Markers to Rhetorical Relations In Sections 6.3.2 and 6.3.3, we have seen how the data in the corpus enabled the development of algorithms that determine the elementary units of a text and the cue phrases that have discourse functions. I now explain how the data in the corpus enables the development of algorithms that hypothesize rhetorical relations that hold among textual units.

In order 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 6.3.2) a set of features for each of the discourse functions that a regular expression can signal. 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 database field **Statuses**. Hence, the accepted values are SATELLITE_NUCLEUS, NUCLEUS_SATELLITE, and NUCLEUS_NUCLEUS.

• The feature "Where to link" specifies whether the rhetorical relations signalled 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 database field **Where to link**.

• 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 database field **Types of textual units**. The accepted values are CLAUSE, SENTENCE, and PARAGRAPH.

• The feature "Rhetorical relation" specifies the names of rhetorical relations that may be signalled by the cue phrase under consideration. Its value is given by the names listed in the database field **Rhetorical relation**.

• 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** 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 are 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**.

Table 6.10 lists the feature sets associated with the cue phrases that were initially listed in Table 6.5. Table 6.10 uses the following abbreviations: Max. dist. stands for "Maximal distance"; Dist. sal. for "Distance to salient unit"; N_S for NUCLEUS_SATELLITE; N_N for NUCLEUS_NUCLEUS; S_N for SATELLITE_NUCLEUS; B for BEFORE; A for AFTER; C for CLAUSE-LIKE UNIT; S for SENTENCE; and P for PARAGRAPH.

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* signalled 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 $\vee$ PARAGRAPH, ELABORATION, 5, 0} specifies that the marker also signals an ELABORATION relation that holds between two sentences or two paragraphs. The first sentence or

Table 6.10
The list of features sets that are used to hypothesize rhetorical relations for the discourse markers and punctuation marks shown in Table 6.5

Marker	Statuses	Where to link	Types of textual units	Rhetorical relations	Max. dist.	Dist. sal.
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		

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 5. 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).

A Discourse-Marker-Based Algorithm for Hypothesizing Rhetorical Relations At the end of step II of the rhetorical parsing algorithm (see Figure 6.5), 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 are hypothesized, on the basis of

information extracted from the corpus. The algorithm that generates these hypotheses is shown in Figure 6.9.

At each level of granularity (sentence, paragraph, and section), the discourse-marker-based hypothesizing algorithm 6.9 iterates over all textual units of that level and over all discourse markers that are relevant to them (see lines 2–4 in Figure 6.9). For each discourse marker, the algorithm constructs an exclusively disjunctive hypothesis concerning the rhetorical relation that the marker under scrutiny may signal. Assume, for example, that the algorithm is processing the i-th unit of the sequence of n units and assume that unit i contains a discourse marker that signals a rhetorical relation that links the unit under scrutiny with one that went before, and whose satellite goes after the nucleus. Given the data derived from the corpus analysis shown in Table 6.10, an appropriate disjunctive hypothesis is that shown in 6.22 below, where NAME is the name of the rhetorical relation that can be signalled by the marker, 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 Dist_sal(m) is the maximum distance to the salient unit found in the rightmost position.

$$\begin{aligned}
&\mathit{rhet_rel}(\text{NAME}, i, i-1) \oplus \cdots \oplus \mathit{rhet_rel}(\text{NAME}, i, i - \mathit{Max}(m)) \\
&\quad \oplus \mathit{rhet_rel}(\text{NAME}, i+1, i-1) \oplus \cdots \oplus \mathit{rhet_rel}(\text{NAME}, i+1, i - \mathit{Max}(m)) \\
&\quad \vdots \\
&\quad \oplus \mathit{rhet_rel}(\text{NAME}, i + \mathit{Dist_sal}(m) + 1, i-1) \oplus \cdots \\
&\quad \oplus \mathit{rhet_rel}(\text{NAME}, i + \mathit{Dist_sal}(m) + 1, i - \mathit{Max}(m))
\end{aligned} \tag{6.22}$$

Essentially, the disjunctive hypothesis enumerates relations of type NAME over members of the Cartesian product $\{i, i+1, \ldots, i + \text{Dist\_sal}(m) + 1\} \times \{i - \text{Max}(m), i - \text{Max}(m) + 1, \ldots, i-1\}$, i.e., all the pairs of units that are separated by an imaginary line drawn between units $i-1$ and i (see Figure 6.10). The disjunctive hypotheses that are generated by the algorithm are exclusive ($\oplus$), because a rhetorical relation that is signalled by a discourse marker cannot be used more than once in building a valid text structure for a text.

The discourse-marker-based hypothesizing algorithm shown in Figure 6.9 automatically builds exclusively disjunctive hypotheses of the kind shown in 6.22 by iterating over all pairs of the Cartesian product. Lines 6–16 concern the case in which the marker m of unit i signals a rhetorical relation that holds between a span that contains unit i and a unit that precedes it. Figure 6.10 illustrates the relations that are generated by these lines in the subcase that is dealt with in line 14 of the algorithm, in which the satellite of the relation comes after the nucleus. In contrast, lines 18–28 concern the case in which the marker m of unit i signals a rhetorical relation that holds between a span that contains unit i and a unit that comes after it.

Input: A sequence $U[n]$ of textual units.
The set D_d of discourse markers that occur in U.

Output: A list RR_d of disjunctive hypotheses of relations that hold among the units in U.

```
1.  RR_d := NULL;
2.  for i from 1 to n
3.      for each marker m ∈ D_d that belongs to U[i] and that
4.              relates units having the same type as those in U
5.          if Where_to_link(m) = BEFORE
6.              rr := NULL;
7.              l := i − 1;
8.              while (l ≥ 0 ∧ i − l ≤ Maximal_distance(m))
9.                  r := i;
10.                 while (r ≤ n ∧ r − i ≤ Distance_to_salient_unit(m) + 1)
11.                     if (Statuses(m) = SATELLITE_NUCLEUS)
12.                         rr := rr ⊕ rhet_rhel(name(d), l, r);
13.                     else
14.                         rr := rr ⊕ rhet_rhel(name(d), r, l);
15.                     r := r + 1;
16.                 l := l − 1;
17.         else
18.             rr := NULL;
19.             r := i + 1;
20.             while (r ≤ n ∧ r − i ≤ Maximal_distance(m))
21.                 l := i;
22.                 while (l ≥ 0 ∧ i − l ≤ Distance_to_salient_unit(m) + 1)
23.                         if (Statuses(m) = SATELLITE_NUCLEUS)
24.                           rr := rr ⊕ rhet_rhel(name(d), l, r);
25.                         else
26.                           rr := rr ⊕ rhet_rhel(name(d), r, l);
27.                         l := l − 1;
28.                 r := r + 1;
29.         endif
30.     RR_d := RR_d ∪ {rr};
31.     endfor
32. endfor
```

Figure 6.9
The discourse-marker-based hypothesizing algorithm

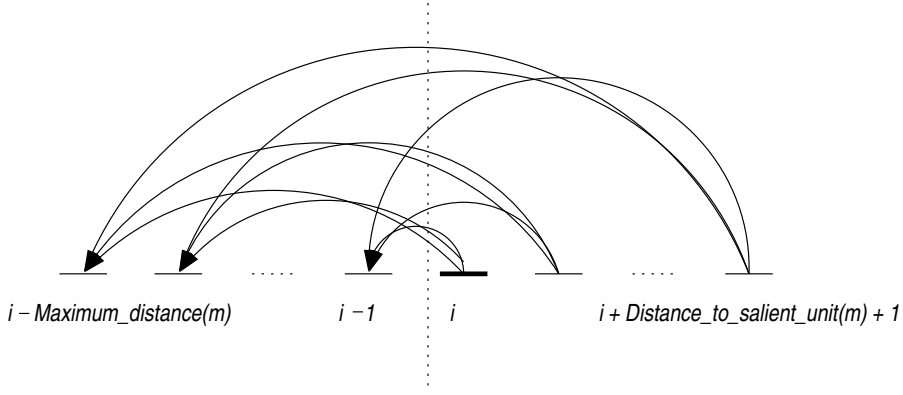

Figure 6.10
A graphical representation of the disjunctive hypothesis that is generated by the discourse-marker-based hypothesizing algorithm for a discourse marker *m* that belongs to unit *i* and that signals a rhetorical relation whose nucleus comes before the satellite

A Word Cooccurrence-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 valid text structures for 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. Unfortunately, this might not be the case when we consider the paragraph and section levels, 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 marker at all; or that use only markers that link clause-like units within sentences.

Given our commitment to surface-form methods, there are two ways we can deal with this problem. One is to construct text trees using only the information provided by the discourse markers. If we adopt this strategy, given a text, we can obtain a sequence of unconnected valid text structures that span across all the units of that text. Once this sequence of unconnected trees is obtained, we can then use various methods for joining the members of the sequence into a connected structure that spans across all the units of the text. The second way is to hypothesize additional rhetorical relations by using other indicators that can be exploited by surface-form methods, such as word cooccurrences or lexical chains [Morris and Hirst, 1991].

In step III.2, the rhetorical parser employs the second choice: it relies on a facet of cohesion [Halliday and Hasan, 1976] that has been shown to be adequate for determining topic shifts [Hearst, 1997] and clusters of sentences and paragraphs that have a unique theme [Hoey, 1991, Salton et al., 1995, Salton and Allan, 1995]. The algorithm that hypothesizes new, additional rhetorical relations assumes that if two sentences or paragraphs "talk about" the same thing, it is likely 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 likely that a topic shift occurs at the boundary between the two units. The decision as to whether two sentences or paragraphs talk about the same thing is taken by measuring the similarity between the sentences, i.e., by counting the number of words that cooccur in both textual units. If this similarity is above a certain threshold, the textual units are considered to be related. Otherwise, a topic shift is assumed to occur at the boundary between the two.

The steps taken by the word cooccurrence-based hypothesizing algorithm are shown in Figure 6.11. The algorithm generates a disjunctive hypothesis for every pair of adjacent textual units that were not already hypothesized to be related by the discourse-marker-based hypothesizing algorithm. As in the case of the discourse-marker-based algorithm, each hypothesis is a 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.

In order to assess the similarity between two units $l \in \{i - LD, \ldots, i\}$ and $r \in \{i + 1, \ldots, i + RD\}$, stop words 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 (see function "cleanedUp" on line 9 in Figure 6.11), so that words that have the same root could be accounted for by the similarity measurement even in the cases in which 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 units l and h. Otherwise, a JOINT relation is hypothesized to hold between the two units (see lines 12, 14 of the algorithm). 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.

During my experiments, I have noticed that whenever 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.

Input: A sequence $U[n]$ of textual units.
The set RR_d of all rhetorical relations that have been hypothesized to hold among the units is U by the discourse-marker-based algorithm.

Output: The complete set RR of disjunctive rhetorical relations that hold among the units in U.

```
1.  RR_c := NULL;
2.  for every pair of adjacent units (i, i + 1)
3.      if there is no relation in RR_U that is hypothesized
4.          to hold between units i and i + 1
5.          rr := NULL;
6.          l = i;
7.          while (l ≥ 0 ∧ i − l ≤ LD)
8.              r := i + 1;
9.              while (r ≤ n ∧ r − i ≤ RD)
10.                 if (similarity(cleanedUp(l),cleanedUp(r)) >
11.                     SimilarityThreshHold)
12.                     rr := rr ⊕ rhet_rel(ELABORATION, r, l) ⊕ rhet_rel(BACKGROUND, l, r);
13.                 else
14.                     rr := rr ⊕ rhet_rel(JOINT, l, r);
15.             r = r + 1;
16.         l = l − 1;
17.     endif
18.     RR_c = RR_c ∪ {rr};
19. endfor
20. RR := RR_d ∪ RR_c;
```

Figure 6.11
The word cooccurrence-based hypothesizing algorithm

Hypothesizing Rhetorical Relations—An Example Let us consider, again, the text in Figure 1.1. Given the textual units and the discourse markers that were identified by the clause-like unit and discourse-marker identification algorithm (see Figure 6.8), we now examine the relations that are hypothesized by the discourse-marker- and word cooccurrence-based hypothesizing algorithms at the sentence, paragraph, and section levels. The text in Figure 6.8 has three sentences that have more than one elementary unit. For the sentence shown in 6.23, the discourse-marker-based algorithm hypothesizes the disjunction shown in 6.24. This hypothesis is consistent with the information given in Table 6.10, which shows that, in the corpus, the marker "With" consistently signalled BACKGROUND and JUSTIFICA-

TION relations between a satellite, the unit that contained the marker, and a nucleus, the unit that followed it.

[*With* its distant orbit {—50 percent farther from the sun than Earth—} and slim atmospheric blanket,[1]] [Mars experiences frigid weather conditions.[2]] (6.23)

rhet_rel(BACKGROUND, 1, 2) ⊕ *rhet_rel*(JUSTIFICATION, 1, 2) (6.24)

For the sentence shown in 6.25, the discourse-marker-based algorithm hypothesizes the two disjunctions shown in 6.26.

[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]] (6.25)

$$\begin{cases} \textit{rhet\_rel}(\text{CONTRAST}, 4, 5) \oplus \textit{rhet\_rel}(\text{CONTRAST}, 4, 6) \\ \textit{rhet\_rel}(\text{CAUSE}, 6, 4) \oplus \textit{rhet\_rel}(\text{EVIDENCE}, 6, 4) \\ \quad \oplus \textit{rhet\_rel}(\text{CAUSE}, 6, 5) \oplus \textit{rhet\_rel}(\text{EVIDENCE}, 6, 5) \end{cases} \quad (6.26)$$

This hypothesis is consistent with the information given in Table 6.10 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 affects the clause-like unit that comes after the one that contains the marker *but* (the **Distance to salient unit** feature has value 0), so *rhet_rel*(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 6.27, there is only one rhetorical relation that is hypothesized, that shown in 6.28.

[*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]] (6.27)

rhet_rel(CONCESSION, 7, 8) (6.28)

The text in Figure 6.8 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. The word co-occurrence-based algorithm deletes first the stop words from the three sentences of the paragraph and removes the suffixes of the remaining words, thus obtaining a list of the root words. When the boundary between the first two sentences is examined by the word cooccurrence-based algorithm, no stemmed words are found to cooccur in the first two sentences, but the stem *sun* is found to

cooccur in the first and third sentences. Therefore, the algorithm hypothesizes the first disjunct in 6.29. When the boundary between the last two sentences is examined, a disjunct having the same form is hypothesized. To distinguish between the two different sources that generated the disjuncts, I assign different subscripts to the rhetorical relations shown in 6.29.

$$\begin{cases} \textit{rhet\_rel}(\text{JOINT}_1, [1, 2], 3) \oplus \textit{rhet\_rel}(\text{ELABORATION}_1, [4, 6], [1, 2]) \\ \quad \oplus \textit{rhet\_rel}(\text{BACKGROUND}_1, [1, 2], [4, 6]) \\ \textit{rhet\_rel}(\text{ELABORATION}_2, [4, 6], [1, 2]) \\ \quad \oplus \textit{rhet\_rel}(\text{BACKGROUND}_2, [1, 2], [4, 6]) \oplus \textit{rhet\_rel}(\text{JOINT}_2, 3, [4, 6]) \end{cases} \tag{6.29}$$

If we apply the heuristic that assumes that the relations between textual units are of type ELABORATION in the cases in which the number of units is small, the rhetorical relations that are hypothesized by the word cooccurrence-based algorithm are those shown in 6.30.

$$\begin{cases} \textit{rhet\_rel}(\text{ELABORATION}, 3, [1, 2]) \\ \textit{rhet\_rel}(\text{ELABORATION}, [4, 6], 3) \end{cases} \tag{6.30}$$

In contrast with the situation discussed with respect to the first paragraph of the text in Figure 6.8, the second paragraph contains markers that provide enough information to link the sentences that belong to it. 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 6.10, 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 can be the sentence that follows the sentence that contains the discourse marker as well (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 6.31 represents all the hypotheses that are made by the algorithm. Because at least one rhetorical relation has been hypothesized for each pair of adjacent sentences in the second paragraph, the word cooccurrence-based algorithm makes no further predictions.

$$\begin{cases} \textit{rhet\_rel}(\text{EXAMPLE}, 9, [7, 8]) \oplus \textit{rhet\_rel}(\text{EXAMPLE}, 10, [7, 8]) \\ \textit{rhet\_rel}(\text{ANTITHESIS}, 9, 10) \oplus \textit{rhet\_rel}(\text{ANTITHESIS}, [7, 8], 10) \end{cases} \tag{6.31}$$

During the corpus analysis, I was not able to draw a line between the discourse markers that could signal rhetorical relations that hold between sentences and relations that hold

between sequences of sentences, paragraphs, and multiparagraphs. However, I have noticed that a discourse marker signals a rhetorical relation that holds between two paragraphs when the marker under scrutiny is located either at the beginning of the second paragraph, or at the end of the first paragraph. The rhetorical parser implements this observation by assuming that rhetorical relations between paragraphs can be signalled 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 comes after 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 the existence of 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 the text in Figure 6.8, that shown in 6.32, below.

rhet_rel(ELABORATION, [7, 10], [1, 6]) (6.32)

The current implementation of the rhetorical parser does not hypothesize any relations among the sections of a text.

Implementing the Proof-Theoretic Account In step III.3, the cue-phrase-based rhetorical parsing algorithm uses the exclusively disjunctive hypotheses and the proof-theoretic account described in Chapter 3 in order to derive the valid trees of each sentence, paragraph, and section. The proof-theoretic account is implemented as a simple chart parser.

The main idea of chart parsing is to store in a data structure the partial results of the parsing process in such a way that no operations are performed more than once. The chart-parsing algorithm takes as input a sequence of units, which are labeled from 1 to N, and a set of simple, extended, and disjunctive rhetorical relations that hold among these units. Parsing the sequence of N units consists in building a chart with N + 1 vertices and adding edges to it, one at a time, in an attempt to create an edge that spans all the units of the input. Each edge of the chart parser has the form [*start*, *end*, *grammar_rule*, *valid_node*, *rhet_rels*], where *start* and *end* represent the first and last node of the span that is covered by the edge, *grammar_rule* represents the grammar rule that accounts for the parse, *valid_node* is a data structure that describes the status, type, and promotion units of a valid tree structure that spans over the units of the interval [*start*, *end*], and *rhet_rels* is the set of rhetorical relations that can be used to extend the given edge. The rhetorical parser uses only two types of grammar rules, which are shown in 6.33, below.

$S \rightarrow i$ *For each elementary unit i in the text* (6.33)

$S \rightarrow S\ S$

The grammar rules that are associated with the chart might be only partially completed. We use the traditional bullet symbol "•" in order to separate the units that have been processed from the units that are still to be processed. For example, an edge of the form $[0, 3, S \rightarrow S \bullet S, vn_1, r_1]$ describes the situation that corresponds to a valid text structure vn_1 that spans over units 1 to 3; if we could build a valid text structure that spans the remaining symbols of the input, then we would have a complete parse of the text. This would correspond to an edge of the form $[0, \text{N}, S \rightarrow SS\bullet, vn_2, r_2]$.

Traditionally, the chart-parsing method provides four different ways for adding an edge to a chart: INITIATE, SCAN, PREDICT, and COMPLETE (see [Russell and Norvig, 1995, Maxwell and Kaplan, 1993] for a discussion of the general method). Because the grammar that we use is very simple, we can compile into the chart-parsing algorithm the choices that pertain to each of the four possible ways of adding an edge to the chart. To do this, we consider the following labels, which describe all the possible levels of completion that could characterize the partial and complete parses of each grammar rule:

Grammar rule	Label
$S \rightarrow \bullet i$	STARTUNIT
$S \rightarrow i\bullet$	ENDUNIT
$S \rightarrow \bullet SS$	STARTCOMPOUND
$S \rightarrow S \bullet S$	MIDDLECOMPOUND
$S \rightarrow SS\bullet$	ENDCOMPOUND

The chart-parsing algorithm that implements the disjunctive proof-theoretic account for deriving text structures in given in Figure 6.12. Initially, the chart is set to *nil*. The INITIALIZER adds an edge to the chart that indicates that the parser is attempting to derive a valid tree starting at position 0 using any of the rhetorical relations in the initial set. The only grammar rule that can be used to do this corresponds to the type STARTCOMPOUND.[7] The PREDICTOR takes an incomplete edge ($grammar\_rule_p \in$ {STARTCOMPOUND, MIDDLECOMPOUND}) and adds new edges that, if completed, would account for the first nonterminal that follows the bullet. There are only two possible types of edges that can be predicted: they correspond to the types STARTUNIT and STARTCOMPOUND. The COMPLETER is looking for an incomplete edge that ends at vertex j (STARTCOMPOUND or MIDDLECOMPOUND) and that is looking for a new nonterminal of type S that starts at vertex j and has S as its left side. In other words, the COMPLETER is trying to join an existing valid text structure, which spans over units $i + 1$ to j, with another text structure that spans over units $j + 1$ to k. The function "canPutTogether" checks to see whether the valid structures and the sets of rhetorical relations that can be used to extend them match one of the axioms given in 3.25–

7. The rhetorical parser assumes that the input has at least two units.

Input: A sequence $U = 1, 2, \ldots,$ N of elementary textual units.
A set *RR* of rhetorical relations that hold among these units.

Output: A chart that subsumes all valid text structures of U.

```
1.  function CHART-PARSER(N, RR)
2.      chart := nil;
3.      INITIALIZER(RR);
4.      for i from 1 to N
5.          SCANNER(i);
6.      return chart;

7.  procedure ADD-EDGE(edge)
8.      if edge ∉ chart[EndOf(edge)]
9.          push edge in chart[EndOf(edge)];
10.         if GrammarRuleOf(edge) ∈ {ENDUNIT,ENDCOMPOUND})
11.             COMPLETER(edge);
12.         else
13.             PREDICTOR(edge);

14. procedure INITIALIZER(RR)
15.     ADD_EDGE([0, 0, STARTCOMPOUND, NULL, RR]);

16. procedure SCANNER(j)
17.     for each [i, j, STARTUNIT, valid_node_c, rr_c] in chart[j] do
18.         ADD-EDGE([i, j + 1, ENDUNIT, new_valid_node, RR]);

19. procedure PREDICTOR([i, j, grammar_rule_p, valid_node_p, rr_p])
20.     ADD-EDGE([j, j, STARTCOMPOUND, NULL, RR]);
21.     ADD-EDGE([j, j, STARTUNIT, NULL, RR]);

22. procedure COMPLETER([j, k, grammar_rule_c, valid_node_c, rr_c])
23.     for each [i, j, STARTCOMPOUND, valid_node, rr] in chart[j] do
24.         if (r=canPutTogether(valid_node_c, valid_node, rr_c, rr)) ≠ nil
25.           ADD-EDGE([i, k, MIDDLECOMPOUND, new_valid_node, rr ∩ rr_c \⊕ {r}]);
26.     for each [i, j, MIDDLECOMPOUND, valid_node, rr] in chart[j] do
27.         if (r=canPutTogether(valid_node_c, valid_node, rr_c, rr)) ≠ nil
28.           ADD-EDGE([i, k, ENDCOMPOUND, new_valid_node, rr ∩ rr_c \⊕ {r}]);
```

Figure 6.12
A chart-parsing algorithm that implements the proof-theoretic account of building valid text structures

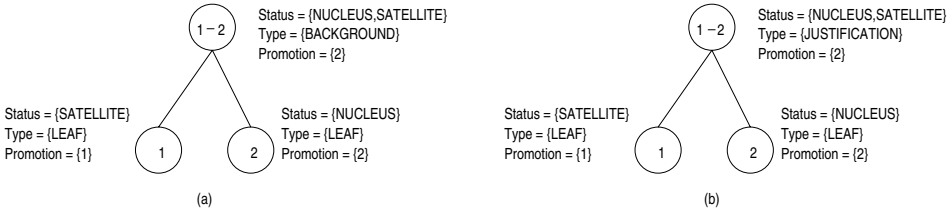

Figure 6.13
The valid text structures of sentence 6.23

3.42. If the two structures can be used to create a valid structure that has relation r in its top node and that spans over units $i + 1$ to k, a new edge is added to the chart. The text structure *new_valid_node* that characterizes the new edge enforces the constraints specified in one of the axioms 3.25–3.42. The SCANNER is like the COMPLETER, except that it uses the input units rather than completed edges in order to generate new edges. In the final text structure, the valid nodes that correspond to these edges will have the type LEAF.

The chart-parsing algorithm produces a chart that subsumes all valid text structures of the text given as input. A simple traversal of the chart can recover any of the valid structures in polynomial time.

When the chart-parsing implementation uses as input the rhetorical relations that were hypothesized by the discourse-marker- and word cooccurrence-based algorithms at the sentence, paragraph, and section levels of the text in Figure 6.8, it derives the valid text structures shown in Figures 6.13–6.18.

6.3.5 The Ambiguity of Discourse

A Weight Function for Text 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 text structures for sentence 6.25 (see Figure 6.14). In my experiments, I noticed, at least for English, that the "best" discourse trees are often those that are skewed to the right. I believe that the explanation of 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 to interpret unmarked textual units as continuations of the topics of the units

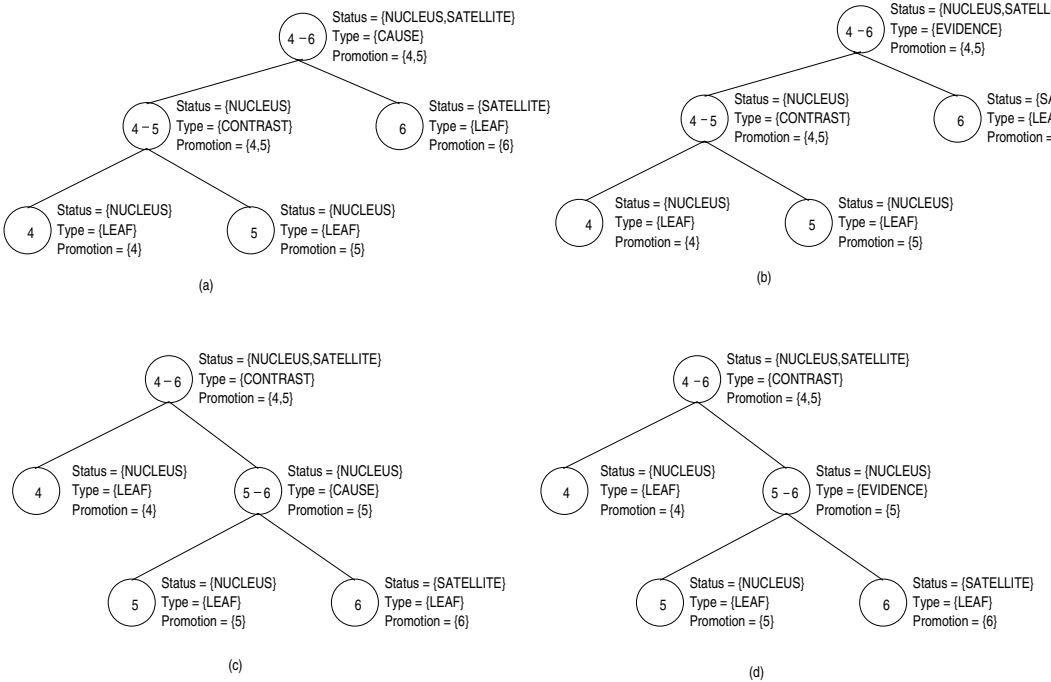

Figure 6.14
The valid text structures of sentence 6.25

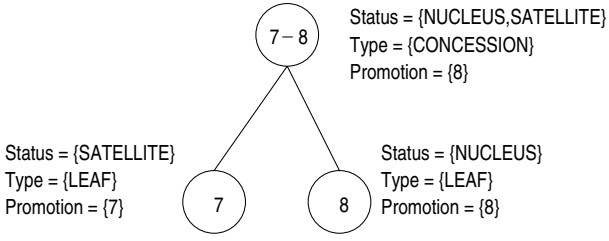

Figure 6.15
The valid text structure of sentence 6.27

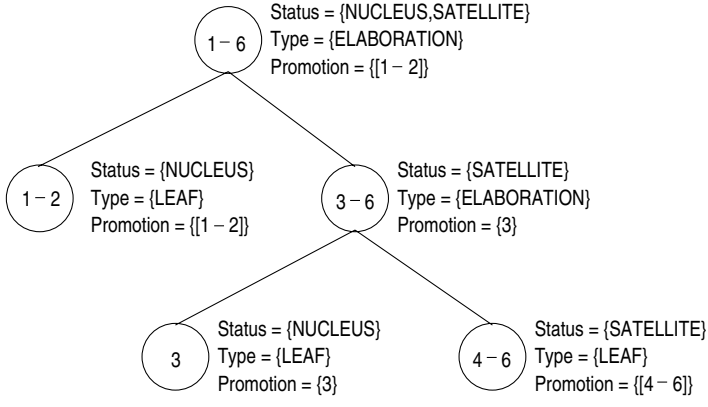

Figure 6.16
The valid text structure of the first paragraph of the text in Figure 6.8 (see relations 6.30)

that precede them [Segal et al., 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 6.34, is computed recursively by summing up the weights of the left and right branches of a text 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.

$$w(tree) = \begin{cases} 0 & \text{if } isLeaf(tree), \\ w(leftOf(tree)) + w(rightOf(tree)) & \text{otherwise} \\ \quad + depth(rightOf(tree)) - depth(leftOf(tree)) & \end{cases} \tag{6.34}$$

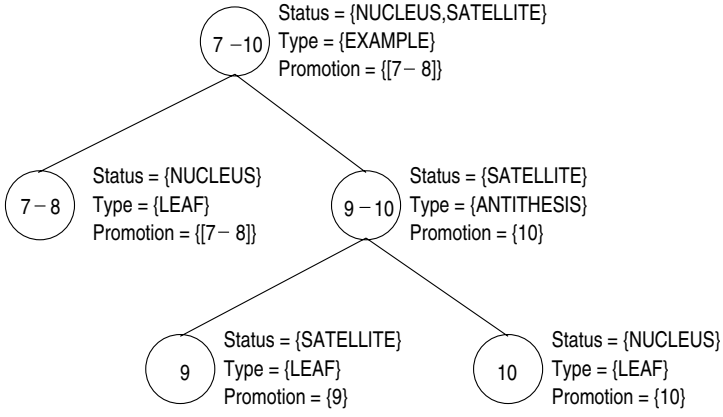

Figure 6.17
The valid text structure of the second paragraph of the text in Figure 6.8 (see relations 6.31)

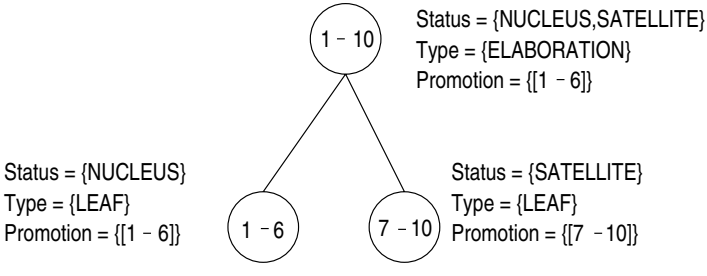

Figure 6.18
The valid text structure of the text in Figure 6.8 (see relation 6.32)

For example, when applied to the valid text structures of sentence 6.25, the weight function will assign the value -1 to the trees shown in Figures 6.14a and 6.14b, and the value $+1$ to the trees shown in Figures 6.14c and 6.14d.

The Ambiguity of Discourse—An Implementation Perspective There are two ways in which one can disambiguate discourse. One way is to consider, during the parsing process, all of the valid text 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 6.5 implements the second approach. Hence, instead of keeping all the partial structures that characterize sentence 6.25, it will keep only the partial structures of maximal weight, i.e., the structures shown in Figures 6.14c and 6.14d. In this way, the overall efficiency of the system is increased.

When the rhetorical parser selects the trees of maximal weight for the text in Figure 6.8 at each of the three levels of abstraction, it selects the trees shown in Figures 6.13a, 6.14c, 6.15, 6.16, 6.17, and 6.18. If no weight function were used, the rhetorical parser would generate eight distinct valid text structures for the whole text.

6.3.6 Deriving the Final Text Structure

In the last step (lines 16–17 in Figure 6.5), 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. 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. In this way, the rhetorical parser builds one tree for each of the sections of a given document. The rhetorical parser has a back-end process that uses "dot," a preprocessor for drawing oriented graphs, in order to automatically generate PostScript representations of the text structures of maximal weight.

When applied to the text in Figure 1.1, the rhetorical parser builds the text structure shown in Figure 6.19. The conventions that I use are are solid boxes to surround nuclei and dotted boxes to surround satellites; the links between a node and the subordinate nucleus or nuclei are represented by solid arrows, and the links between a node and the subordinate satellites by dotted lines. The occurrences of parenthetical information are enclosed in the text with 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.

All the algorithms described in this book have been implemented in C++.

6.3.7 Evaluation of the Cue-Phrase-Based Rhetorical Parser

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 these trees can have on solving other natural language processing problems, such as anaphora resolution, intention recognition, or text summarization. The rhetorical parser discussed in this chapter was evaluated in both ways.

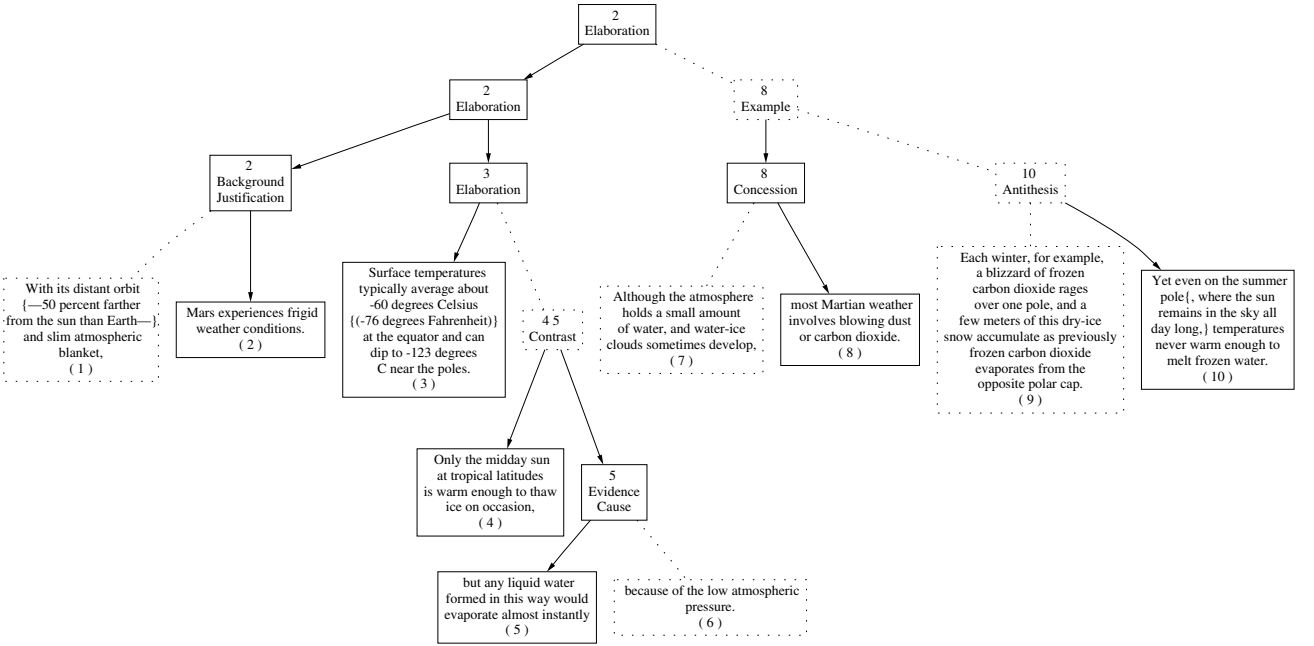

Figure 6.19
The discourse tree of maximal weight that is built by the rhetorical-parsing algorithm for the text in Figure 1.1

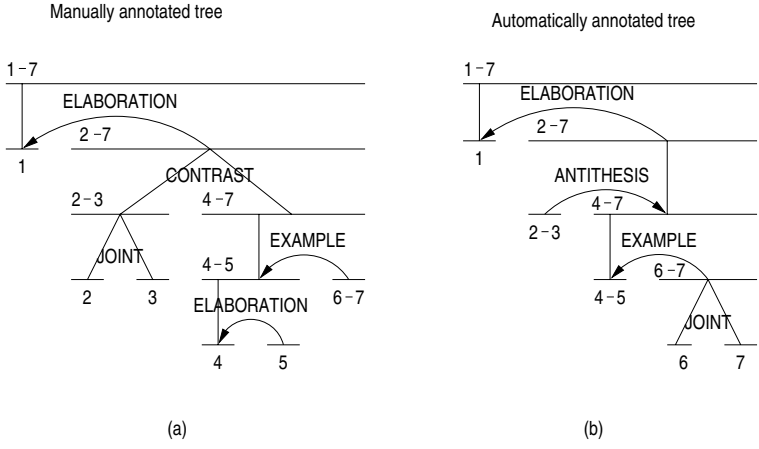

Figure 6.20
Computing the performance of a rhetorical parser

Evaluating the Correctness of the Trees In order 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 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 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 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. We computed labeled recall and precision figures with respect to the ability of the cue-phrase-based rhetorical parser to identify elementary units, hierarchical text spans, text span nuclei and satellites, and rhetorical relations as described below.

Assume for example that an analyst identified six elementary units in a text and built the discourse structure in Figure 6.20a and that the program identified five elementary units and built the discourse structure in Figure 6.20b. When we align the two structures, we obtain the labels in Figure 6.20, 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 that

Table 6.11
Computing the performance of a rhetorical parser (P = Program; A = Analyst)

	Units		Spans		Nuclearity		Relations	
Constit	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		R=6/10		R=5/10		R=4/10	
	P=1/5		P=6/8		P=5/8		P=4/8	

the unit labeled 6–7 in the analyst's annotation is made of two units. Table 6.11 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 6.11 shows, the program 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.

The reader can note that during the evaluation process I did not assign the rhetorical labels to the father nodes, as in the formalization, but rather to the children nodes. For example, the EXAMPLE relation that holds between spans [4,5] and [6,7] in the tree in Figure 6.20a 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 annotators did not construct only binary trees; some of the nodes in their manually built representations had multiple children. In order to represent in a binary form a tree such as that shown in Figure 6.21a, for example, I would have needed to introduce an additional hierarchical level on the spans, as shown in Figures 6.21b and c, that was not part of the original analysis. In order 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.

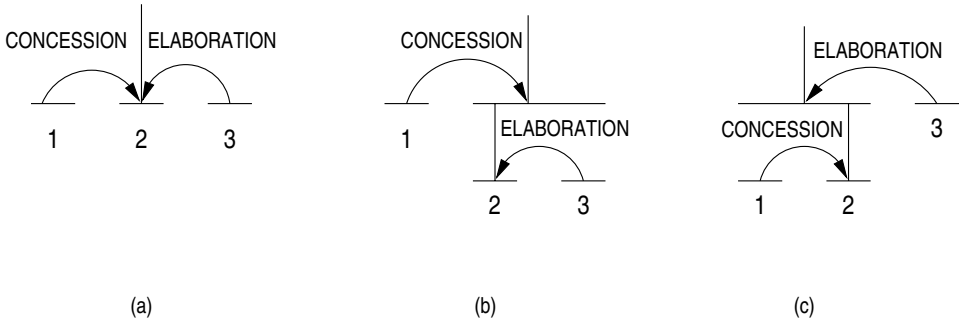

Figure 6.21
Evaluating nonbinary discourse trees

Table 6.12
Performance of the cue-phrase-based 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

Table 6.12 shows the results of evaluating the rhetorical parser on the five *Scientific American* texts. In addition to the recall and precision figures specific to the program, Table 6.12 also displays the average recall and precision figures obtained for the trees built only by the judges. 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 would be the same as the figures that correspond to the analysts, the discourse trees built by the rhetorical parser would be indistinguishable from those built by a human.

As the results in Table 6.12 show, the cue-phrase-based rhetorical parser fails to identify a fair number of elementary units. As a consequence, the overall performance is affected. With respect to identifying hierarchical spans, the recall is about 25% lower than the average human performance; with respect to labeling the nuclear status of spans, the recall is about 30% below human performance; and with respect to labeling the rhetorical relations that hold between spans, the recall is about 40% below human performance. In

general, the precision of the rhetorical parser comes very 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 both of them as a single elementary unit. As a consequence, the recall figures at all levels were significantly lower than those specific to the humans. For a qualitative evaluation of the rhetorical parser see [Marcu, 2000].

Although these performance levels seem low, as we will see in Chapter 9, they seem to be sufficient for the task of summarizing text.

Evaluating the Utility 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 6.19, 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 in order to induce a partial ordering on the importance of these units. The intuition behind this approach is that the textual units that are 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 6.19, such an approach induces the partial ordering in 6.35, 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 Chapter 9 for a mathematical formulation of this method that uses rhetorical structures for deriving a partial ordering of the important units in texts.)

$$2 > 8 > 3, 10 > 1, 4, 5, 7, 9 > 6 \tag{6.35}$$

If we are interested in generating a very short summary of the text in Figure 6.8, 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 (see Chapter 9). The algorithm uses the cue-phrase-based rhetorical parser in order 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: a collection of five *Scientific American* texts, and a collection of forty short newspaper articles taken from the TREC corpus [Jing et al., 1998]. Both corpora were labeled for textual saliency by a panel of independent judges: thirteen judges labeled clause-like units as being important, somewhat important, and nonimportant in the texts of the *Scientific American* corpus; and five judges labeled sentences as worthy to be included in 10% and 20% summaries of the texts in the TREC corpus. The clauses/sentences that human judges agreed were important were taken as gold standard for summarization. To assess the performance of the rhetorical-based summarizer, I use recall and precision 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.

When a summarization program used the discourse structures derived by the rhetorical parsing algorithm to extract important clauses and sentences from the same corpora it performed as follows. For the *Scientific American* corpus, where the recall and precision associated with the human judges were 72.66% and 69.63% respectively, our program recalled 51.35% of the clauses in the gold standard with a precision of 63.33%. In contrast, Microsoft's Office97 summarizer recalled 27.77% of the important clauses with a precision of 25.54%. For the TREC corpus, where the recall and precision figures associated with humans were 82.83% and 64.93% respectively, our program recalled 46.54% of the sentences in the 20% gold standard, with a precision of 49.73%.

In Chapter 9, I describe in detail these experiments and in Chapter 10 I present a learning algorithm that uses six heuristics, besides the skewedness of the discourse trees, for the task of rhetorical parsing disambiguation. When all these heuristics are used, the performance of the program improves to 67.57% recall and 73.53% precision for the *Scientific American* corpus and 61.79% recall and 60.83% precision for the TREC corpus.

7 Rhetorical Parsing by Means of Automatically Derived Rules

7.1 Introduction

The application of decision-tree-based learning techniques over rich sets of linguistic features has improved significantly the coverage and performance of syntactic (and to various degrees semantic) parsers [Simmons and Yu, 1992, Magerman, 1995, Hermjakob and Mooney, 1997]. In this chapter, I apply a similar paradigm to developing a rhetorical parser that derives the discourse structure of unrestricted texts.

Crucial to this approach is the reliance on a corpus of manually built discourse trees and the adoption of a shift-reduce parsing model that is well suited for learning. Both the corpus and the parsing model are used to generate learning cases of how texts should be partitioned into elementary discourse units (*edu*s) and how discourse units and segments should be assembled into discourse trees.

In this chapter, I first describe the corpus that I used for learning (Section 7.2). Then, I present the parsing model and I discuss in detail the decision tree-based rhetorical parser (Section 7.3). I conclude with an evaluation section.

7.2 A Corpus Analysis of Discourse Trees

I used a corpus of ninety rhetorical structure trees, which were built manually using rhetorical relations that were defined informally in the style of Mann and Thompson [1988]: thirty texts were taken from the MUC7 coreference corpus, thirty texts from the Brown-Learned corpus, and thirty texts from the *Wall Street Journal* (WSJ) corpus. The MUC corpus contained news stories about changes in corporate executive management personnel; the Brown corpus contained long, highly elaborate scientific articles; and the WSJ corpus contained mostly editorials. The average number of words for each text was 405 in the MUC corpus, 2029 in the Brown corpus, and 878 in the WSJ corpus. The average number of elementary discourse units *edu*s in each text was 52 in the MUC corpus, 170 in the Brown corpus, and 95 in the WSJ corpus. Each of the MUC texts was tagged by three annotators; each of the Brown and WSJ texts was tagged by two annotators.

The rhetorical structure assigned to each text obeyed the constraints discussed in Chapter 3, with the exception of some trees being nonbinary. In order to represent nonbinary trees that are characterized by one nucleus node to which multiple satellites are attached, I follow the same conventions I used in Section 6.3.7. That is, I associate rhetorical relation names with the children nodes and, by convention, I assign to the nuclei of mononuclear relations the rhetorical name SPAN.

As in the corpus work described in Section 6.2, elementary discourse units were defined functionally as clauses or clause-like units that are unequivocally the NUCLEUS or SATELLITE of a rhetorical relation that holds between two adjacent spans of text. *Parenthetical units*, i.e., embedded units whose deletion does not affect the understanding of the *edu* to which they belong, were also marked. In contrast with the corpus analysis discussed in Section 6.2, where *edu*s were marked intuitively by only one annotator, in this study, the assignment of *edu* and parenthetical unit boundaries was done by multiple annotators using a better-defined procedure: the annotators were given a tagging manual that listed rules for determining the discourse units and for building the discourse structure of texts.

The annotation process was carried out using a rhetorical tagging tool. The process consisted in assigning *edu* and parenthetical unit boundaries, in assembling *edu*s and spans into discourse trees, and in labeling the relations between *edu*s and spans with rhetorical relation names from a taxonomy of seventy relations. The annotators were trained. During the annotation, they used an annotation manual [Marcu, 1998b] that listed the seventy relations partitioned into clusters, each cluster containing a subset of relations that shared some rhetorical meaning. For example, one cluster contained the contrast-like rhetorical relations of ANTITHESIS, CONTRAST, and CONCESSION. Another cluster contained REASON, EVIDENCE, and EXPLANATION. Each relation was paired with an informal definition given in the style of Mann and Thompson [1988] and Moser and Moore [2000] and one or more examples. No explicit distinction was made between intentional and informational relations. In addition, annotators also marked two constituency relations that were ubiquitous in the corpora and that often subsumed complex rhetorical constituents, and one textual relation. The constituency relations were ATTRIBUTION, which was used to label the relation between a reporting and a reported clause, and APPOSITION. The textual relation was TEXTUALORGANIZATION; it was used to connect in an RST-like manner the textual spans that corresponded to the title, author, and textual body of each document in the corpus. The annotators were also allowed to use the label OTHERRELATION whenever they felt that no relation in the manual captured sufficiently well the meaning of a rhetorical relation that held between two text spans. Table 7.1 shows the fifteen relations that were used most frequently by annotators in each of the three corpora; the associated percentages reflect averages computed over all annotators. The table also shows the percentage of cases in which the annotators used the label OTHERRELATION.

In an attempt to manage the inherent rhetorical ambiguity of texts, annotators were given a protocol that listed the clusters of relations in decreasing order of specificity. Hence, the relations at the beginning of the protocol were more specific than the relations at the end of the protocol. The protocol specified that in assigning rhetorical relations annotators should choose the first relation in the protocol whose definition was consistent with the case under consideration. For example, it is often the case that when an EVIDENCE relation holds

Table 7.1
Distribution of the most frequent fifteen rhetorical relations in the three corpora of discourse trees

MUC Corpus		WSJ Corpus		Brown Corpus	
Relation	Percentage	Relation	Percentage	Relation	Percentage
ELABORATION-ADDITIONAL	13.80	ELABORATION-ADDITIONAL	17.41	ELABORATION-ADDITIONAL	21.64
ATTRIBUTION	12.07	ATTRIBUTION	14.78	LIST	16.29
LIST	9.99	LIST	11.25	JOINT	6.58
TEXTUAL-ORGANIZATION	6.23	CONTRAST	6.84	CONTRAST	5.60
APPOSITION	5.02	JOINT	4.35	TEXTUAL-ORGANIZATION	3.22
TOPIC-SHIFT	4.76	EVIDENCE	3.82	PURPOSE	2.88
JOINT	4.19	APPOSITION	3.31	EXPLANATION-ARGUMENTATIVE	2.68
CONTRAST	3.99	TOPIC-SHIFT	2.96	SEQUENCE	2.57
ELABORATION-OBJECT-ATTRIBUTE	2.88	BACKGROUND	2.41	ELABORATION-GENERAL-SPECIFIC	2.23
EVIDENCE	2.54	ELABORATION-OBJECT-ATTRIBUTE	2.37	TOPIC-SHIFT	2.12
BACKGROUND	2.42	PURPOSE	2.21	BACKGROUND	1.96
PURPOSE	2.26	ELABORATION-GENERAL-SPECIFIC	2.19	CONCESSION	1.84
ELABORATION-GENERAL-SPECIFIC	2.21	TOPIC-DRIFT	1.88	ELABORATION-OBJECT-ATTRIBUTE	1.76
TOPIC-DRIFT	1.85	CONDITION	1.77	CONDITION	1.76
SEQUENCE	1.59	SEQUENCE	1.31	EVIDENCE	1.62
. . .		. . .		. . .	
OTHER-RELATION	0.38	OTHER-RELATION	0.32	OTHER-RELATION	0.19

between two segments, an ELABORATION relation holds as well. Because EVIDENCE is more specific than ELABORATION, it comes first in the protocol, and hence, whenever both of these relations hold, only EVIDENCE is supposed to be used for tagging. Hence, in contrast with the corpus analysis of cue phrases discussed in Section 6.2, where I used multiple relations in the ambiguous cases, in the empirical work described here the annotators used only one relation name, that that was most specific.

Marcu [1998b] provides the *edu* and rhetorical relation definitions, and the annotation protocol that the annotators were supposed to follow. Marcu et al. [1999a] assess the inter-judge agreement and the reliability of the annotation.

In contrast with the empirical research described in Section 6.2, the corpus of annotated discourse trees was exploited only by means of automated learning techniques.

7.3 A Decision Tree-Based Approach to Rhetorical Parsing

7.3.1 The Parsing Model

In order to learn to build discourse trees similar to those in the three corpora, I model the discourse parsing process as a sequence of shift-reduce operations. As a front-end, the parser uses a *discourse segmenter*, i.e., an algorithm that partitions the input text into *edu*s. The discourse segmenter, which is decision-based, is presented and evaluated in Section 7.3.2.

The input to the parser is an empty stack and an input list that contains a sequence of elementary discourse trees, *edt*s, one *edt* for each *edu* produced by the discourse segmenter. The status and rhetorical relation associated with each *edt* is UNDEFINED, and the promotion set is given by the corresponding *edu*. At each step, the parser applies a SHIFT or a REDUCE operation. Shift operations transfer the first *edt* of the input list to the top of the stack. Reduce operations pop the two discourse trees located on the top of the stack; combine them into a new tree updating the statuses, rhetorical relation names, and promotion sets associated with the trees involved in the operation; and push the new tree to the top of the stack.

Assume, for example, that the discourse segmenter partitions a text given as input as shown in Figure 7.1, where only the *edu*s numbered from 12 to 19 are shown. Figure 7.2 shows the actions taken by a shift-reduce discourse parser starting with step i. At step i, the stack contains four partial discourse trees, which span units [1,11], [12,15], [16,17], and [18], and the input list contains the *edt*s that correspond to units whose numbers are higher than or equal to 19. At step i the parser decides to perform a SHIFT operation. As a result, the *edt* corresponding to unit 19 becomes the top of the stack. At step $i+1$, the parser performs a REDUCE-APPOSITION-NS operation, that combines *edt*s 18 and 19 into a

. . . [Close parallels between tests and practice tests are common,[12]] [some educators and researchers say.[13]] [Test-preparation booklets, software and worksheets are a booming publishing subindustry.[14]] [But some practice products are so similar to the tests themselves that critics say they represent a form of school-sponsored cheating.[15]]

["If I took these preparation booklets into my classroom,[16]] [I'd have a hard time justifying to my students and parents that it wasn't cheating,"[17]] [says John Kaminsky,[18]] [a Traverse City, Mich., teacher who has studied test coaching.[19]] . . .

Figure 7.1
Example of text whose elementary units are identified

discourse tree whose nucleus is unit 18 and whose satellite is unit 19. The rhetorical relation that holds between units 18 and 19 is APPOSITION. At step $i + 2$, the trees that span over units [16,17] and [18,19] are combined into a larger tree, using a REDUCE-ATTRIBUTION-NS operation. As a result, the status of the tree [16,17] becomes NUCLEUS and the status of the tree [18,19] becomes SATELLITE. The rhetorical relation between the two trees is ATTRIBUTION. At step $i + 3$, the trees at the top of the stack are combined using a REDUCE-ELABORATION-NS operation. The effect of the operation is shown at the bottom of Figure 7.2 (continued).

In order to enable a shift-reduce discourse parser to derive any discourse tree (including nonbinary trees), it is sufficient to implement one SHIFT operation and six types of REDUCE operations, whose operational semantics is shown in Figure 7.3. For each possible pair of nuclearity assignments NUCLEUS-SATELLITE (NS), SATELLITE-NUCLEUS (SN), and NUCLEUS-NUCLEUS (NN) there are two possible ways to attach the tree located at position *top* in the stack to the tree located at position $top - 1$. If one wants to create a binary tree whose immediate children are the trees at *top* and $top - 1$, an operation of type REDUCE-NS, REDUCE-SN, or REDUCE-NN needs to be employed. If one wants to attach the tree at *top* as an extrachild of the tree at $top - 1$, thus creating or modifying a nonbinary tree, an operation of type REDUCE-BELOW-NS, REDUCE-BELOW-SN, or REDUCE-BELOW-NN needs to be employed. Figures 7.3a,b illustrate how the statuses and promotion sets associated with the trees involved in the reduce operations are affected in each case. As one can easily notice, the trees that are derived by each of the six reduce operations obey the constraints discussed in Chapter 3. In particular, the promotion units are determined using the rules that concern the strong compositionality criterion of valid text structures.

Since the labeled data that I relied upon was sparse, I did not attempt to automatically derive trees using all seventy-two rhetorical relation names, but rather the names of the clusters of rhetorical similarity to which the relations belonged. For example, each rhetorical relation of EVALUATION and INTERPRETATION was replaced with the name of the cluster

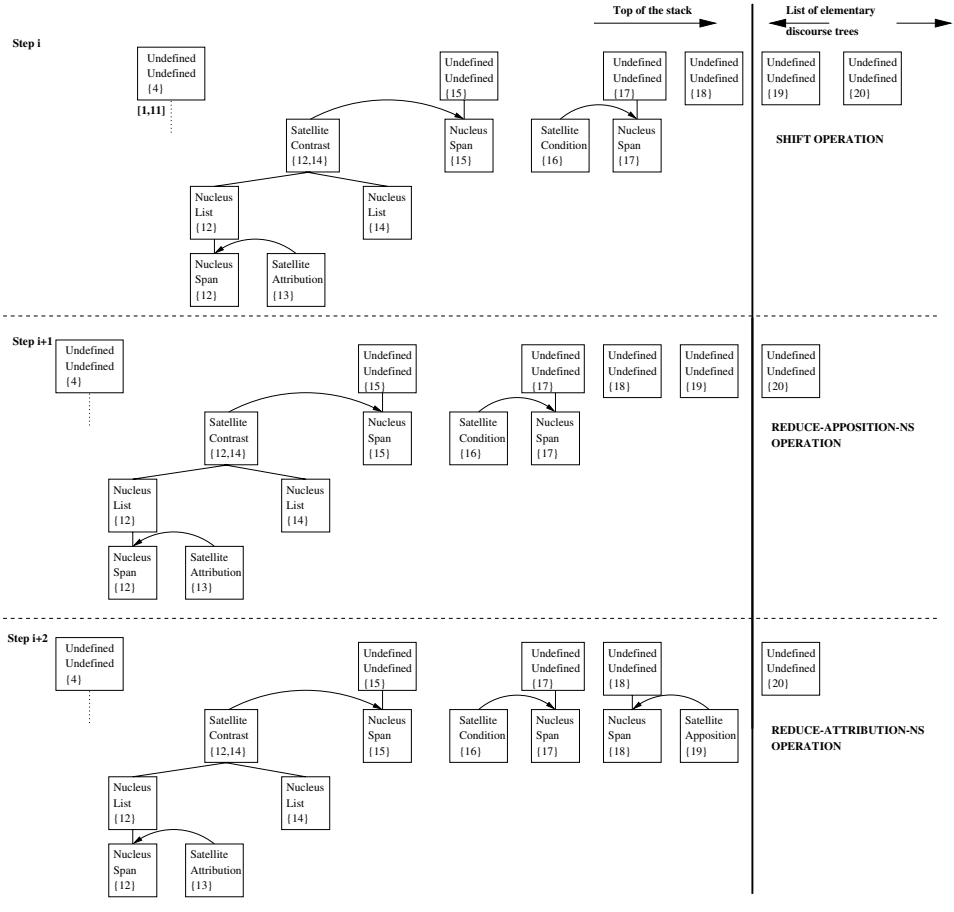

Figure 7.2
Example of a sequence of shift-reduce operations that concern the discourse parsing of the text in Figure 7.1

of rhetorical similarity that contained these two relations, EVALUATION-INTERPRETATION. Overall, I used seventeen clusters, each characterized by a generalized rhetorical relation name. These names were: APPOSITION-PARENTHETICAL, ATTRIBUTION, CONTRAST, BACKGROUND-CIRCUMSTANCE, CAUSE-REASON-EXPLANATION, CONDITION, ELABORATION, EVALUATION-INTERPRETATION, EVIDENCE, EXAMPLE, MANNER-MEANS, ALTERNATIVE, PURPOSE, TEMPORAL, LIST, TEXTUAL, and OTHER. The name OTHER subsumed relations such as QUESTIONANSWER, PROPORTION, RESTATEMENT, and COMPARISON, which were rarely used in the annotations.

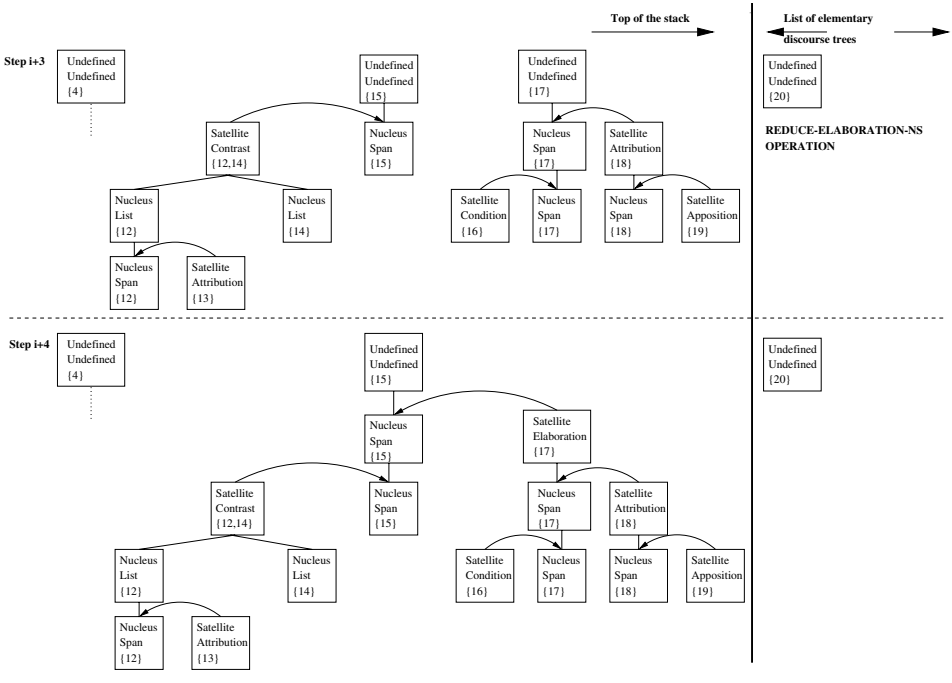

Figure 7.2 *(continued)*

Hence, I attempted to automatically derive rhetorical structure trees that were labeled with relations names that corresponded to the seventeen clusters of rhetorical similarity. Since there are six types of reduce operations and since each discourse tree in the three corpora uses relation names that correspond to the seventeen clusters of rhetorical similarity, it follows that the shift-reduce rhetorical parser needs to learn what operation to choose from a set of $6 \times 17 + 1 = 103$ operations (the 1 corresponds to the SHIFT operation).

7.3.2 The Discourse Segmenter

Generation of Learning Examples The discourse segmenter I implemented processes an input text one lexeme (word or punctuation mark) at a time and recognizes sentence and *edu* boundaries, and beginnings and ends of parenthetical units. I used the leaves of the discourse trees in the corpus in order to derive the training examples (learning cases). To

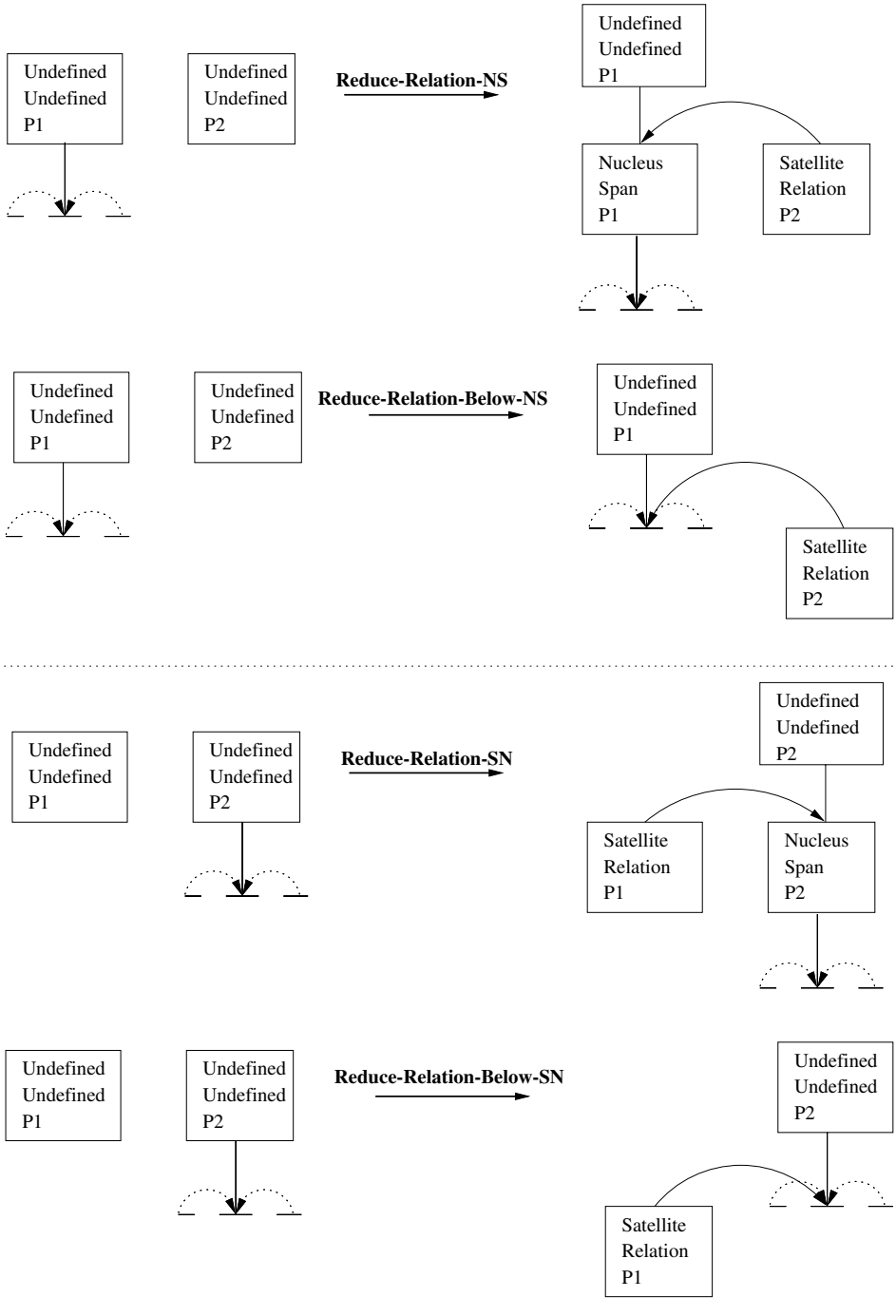

Figure 7.3
The reduce operations supported by the shift-reduce parsing model

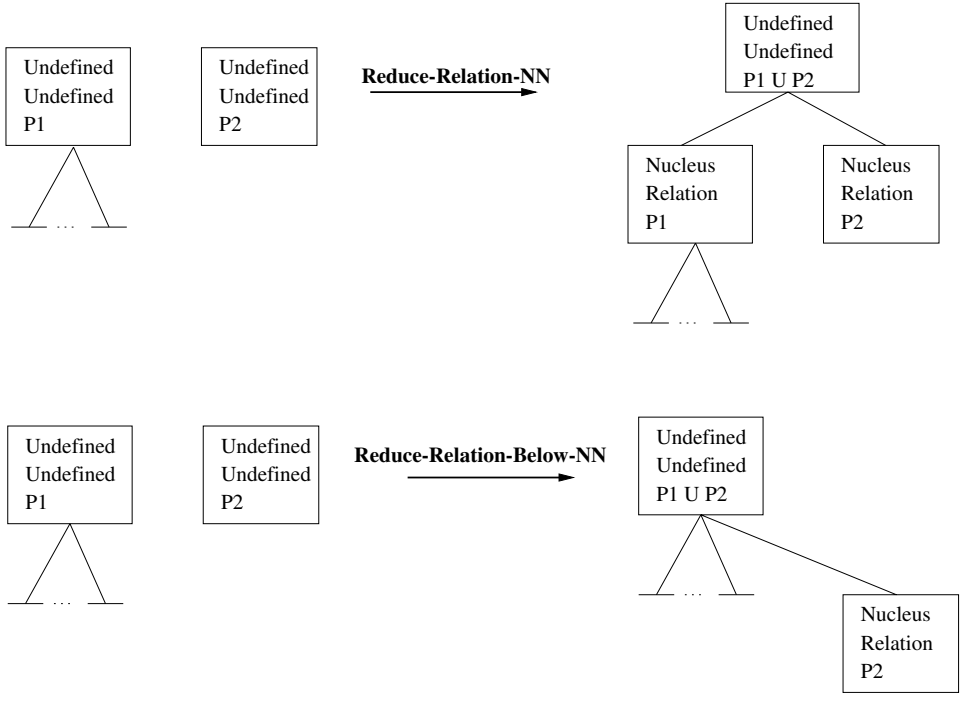

Figure 7.3 *(continued)*

each lexeme in a text, I associated one learning case, using the features described below. The classes to be learned, which are associated with each lexeme, are as follows:

• *sentence-break* corresponds to the situation in which the discourse segmenter inserts a sentence break after the lexeme under scrutiny.

• *edu-break* corresponds to the situation in which the discourse segmenter inserts an elementary (clause-like) unit break after the lexeme under scrutiny.

• *start-paren* corresponds to the situation in which the discourse segmenter identifies the beginning of a parenthetical unit in the input stream.

• *end-paren* corresponds to the situation in which the discourse segmenter identifies the end of a parenthetical unit in the input stream.

• *none* corresponds to the situation in which the discourse segmenter identifies no elementary or parenthetical unit boundary.

Features Used for Learning To partition a text into *edu*s and to detect parenthetical unit boundaries, I relied on features that model both the local and global contexts.

The local context consists of a window of size $5(1 + 2 + 2)$ that enumerates the Part-Of-Speech (POS) tags of the lexeme under scrutiny (1) and the two lexemes found immediately before (2) and after it (2). The POS tags are determined automatically, using the Brill tagger [1995]. Since discourse markers play a major role in rhetorical parsing (see Chapter 6), I also consider a list of features that specify whether a lexeme found within the local contextual window is a potential discourse marker; hence, I specify for each lexeme under scrutiny whether it is a special orthographic marker, such as comma, dash, and parenthesis, or whether it is a potential discourse marker, such as "accordingly," "afterwards," and "and." The local context also contains features that estimate whether the lexemes within the window are potential abbreviations; the system uses a hard-coded list of 250 potential abbreviations.

The global context reflects features that pertain to the boundary identification process. These features specify whether there are any commas, closed parentheses, and dashes before the estimated end of the sentence, whether there are any verbs in the unit under consideration, and whether any discourse marker that introduces expectations was used in the sentence under consideration. These markers are phrases such as *Although* and *With*, which were associated with COMMA and DUAL action types in the corpus analysis in Section 6.2; they were also studied by Cristea and Webber [1997] in the context of incremental discourse parsing.

The decision-based segmenter uses a total of twenty-five features, some of which can take as many as 400 values. When we represent these features in a binary format, we obtain learning examples with 2417 binary features/example.

Examples of Rules Specific to the Segmenter Figure 7.4 shows some of the rules that were learned by the C4.5 program [Quinlan, 1993] using a binary representation of the features and learning cases extracted from the MUC corpus. Rule 1 specifies that if the POS tag of the lexeme that immediately precedes the lexeme under scrutiny is a closed parenthesis and the previous marker recognized during the processing of the current sentence was an open parenthesis, then the action to be taken is to insert an end of parenthetic unit. Rule 1 can correctly identify the end of the parenthetic unit at the location marked with the symbol ↑ in sentence 7.1.

Surface temperatures typically average about −60 degrees Celsius (−76 degrees Fahrenheit)$_{\uparrow}$ at the equator. (7.1)

Rule 2 can correctly identify the beginning of the parenthetic unit 44 years old in sentence 7.2, because the unit is preceded by a comma and starts with a numeral (CD) followed by a plural noun (NNS).

Rule 1: **if** pos(−1) = ")" ∧ previous-marker = "("
then *end-paren*
Rule 2: **if** pos(−1) = "," ∧ pos(0) = CD
pos(+1) = NNS
then *start-paren*
Rule 3: **if** pos(−1) ≠ DOT ∧ pos(0) = DOT
pos(+1) ≠ DOT ∧ pos(+1) ≠ DOUBLEQUOTE
then *end-sentence*
Rule 4: **if** pos(+2) = VBD ∧ word(+1) = "and"
then *edu-break*
Rule 5: **if** word(+1) = "until" ∧ isThereAnyVerbBeforeNextPotentialBreak
then *edu-break*
Rule 6: **if** pos(0) = "," ∧ previous-marker = "while"
then *edu-break*
Rule 7: **if** pos(1) = DOT
then *nothing*

Figure 7.4
Examples of automatically derived segmenting rules

Ms. Washington,↑ 44 years old, would be the first woman and the first black to head the five-member commission that oversees the securities markets. (7.2)

Rule 3 identifies the end of a sentence after the occurrence of a DOT (period, question mark, or exclamation mark) that is not preceded or followed by another DOT and that is not followed by a DOUBLEQUOTE. This rule will correctly identify the sentence end after the period in example 7.3, but will not insert a sentence end after the period in example 7.4. However, another rule that is derived automatically will insert a sentence break after the double quote that follows the ↑ mark in example 7.4.

The meeting went far beyond Mr. Clinton's normal weekly gathering of business leaders.↑ Economic adviser Gene Sperling described it as "a true full-court press" to pass the deficit-reduction bill, the final version of which is now being hammered out by House and Senate negotiators. (7.3)

The executives "are here, just as I am, not because anyone agrees with every last line and jot and tittle of this economic program," Mr. Clinton acknowledged, but "because it does far more good than harm.↑" Despite resistance from some lawmakers in his own party, the president predicted the bill would pass. (7.4)

Rule 4 identifies an *edu* boundary before the occurrence of an "and" followed by a verb in the past tense (VBD). This rule will correctly identify the marked *edu* boundary in sentence 7.5.

Ashley Boone ran marketing and distribution$_{\uparrow}$ and left the company late last year. (7.5)

Rule 5 inserts *edu* boundaries before the occurrence of the word "until", provided that "until" is followed not necessarily immediately by a verb. This rule will correctly insert an *edu* boundary in example 7.6.

Several appointees of President Bush are likely to stay in office at least temporarily,$_{\uparrow}$ until permanent successors can be named. (7.6)

Rule 6 is an automatically derived rule that mirrors the manually derived rule specific to COMMA-like actions in the surface-based unit identification algorithm presented in Section 6.3.3. Rule 6 will correctly insert an *edu* boundary after the comma marked in example 7.7, because the marker "While" was used at the beginning of the sentence.

While the company hasn't commented on the probe,$_{\uparrow}$ persons close to the board said that Messrs. Lavin and Young, along with some other top Woolworth executives, were under investigation by the special committee for their possible involvement in the alleged irregularities. (7.7)

Rule 7 specifies that no elementary or parenthetical unit boundary should be inserted immediately before a DOT.

As one can easily notice, the rules in Figure 7.4 are more complex than those derived manually in Chapter 6. The automatically derived rules make use not only of orthographic and cue-phrase-specific information, but also of syntactic information, which is encoded as part of speech tags.

Evaluation I used the C4.5 program [Quinlan, 1993] in order to learn decision trees and rules that classify lexemes as boundaries of sentences, *edu*s, or parenthetical units, or as nonboundaries. I learned both from binary (when I could) and nonbinary representations of the cases.[1] In general the binary representations yielded slightly better results than the nonbinary representations and the tree classifiers were slightly better than the rule-based ones. Table 7.2 displays accuracy results that concern the nonbinary, decision-tree classifiers and the corresponding standard deviations of the accuracy results. The accuracy figures were computed using a tenfold *stratified* cross-validation procedure. That is, the dataset was randomly split into ten mutually exclusive subsets, each containing approximately the same proportion of labels as the original dataset. The classifier was trained and tested ten times; each time it was tested on a subset, it was trained on the other nine subsets. The cross-

1. Learning from binary representations of features in the Brown corpus was too computationally expensive to terminate—the Brown data file had about 0.5GBytes.

Table 7.2
Performance of a discourse segmenter that uses a decision-tree, nonbinary classifier. B1 corresponds to a majority-based baseline classifier that assigns *none* to all lexemes. B2 corresponds to a baseline classifier that assigns a sentence boundary to every DOT lexeme and a nonboundary to all other lexemes.

Corpus	# cases	B1(%)	B2(%)	Acc(%)
MUC	14362	91.28	93.1	96.24±0.06
WSJ	31309	92.39	94.6	97.14±0.10
Brown	72092	93.84	96.8	97.87±0.04

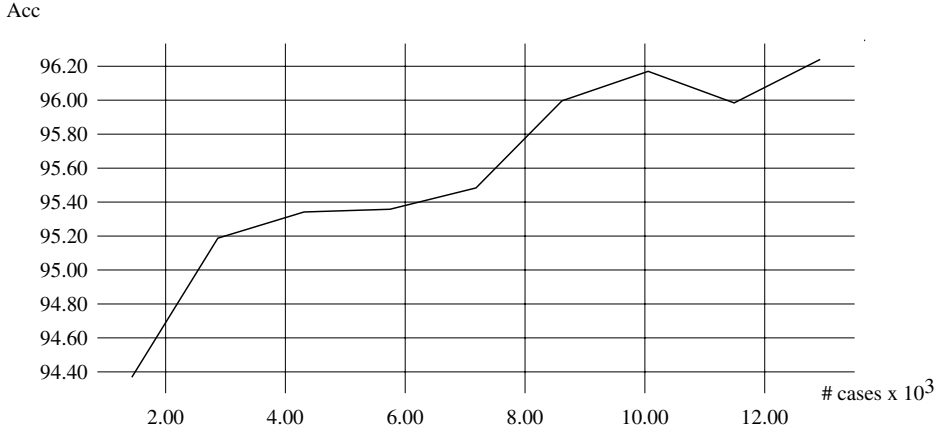

Figure 7.5
Learning curve for discourse segmenter (the MUC corpus)

validation experiments described in this chapter, the accuracy estimation, and the standard deviation were all computed using the MLC software package [Kohavi et al., 1996].

Figure 7.5 shows the learning curve that corresponds to the MUC corpus. It suggests that more data can increase the accuracy of the classifier.

The confusion matrix shown in Table 7.2 corresponds to a nonbinary-based tree classifier that was trained on cases derived from twenty-seven Brown texts and that was tested on cases derived from three different Brown texts, which were selected randomly. The matrix shows that the segmenter has problems mostly with identifying the beginning of parenthetical units and the intra-sentential *edu* boundaries; for example, it correctly identifies only 133 of the 220 $(= 133 + 3 + 84)$ *edu* boundaries and only 4 of the $30(= 26 + 4)$ beginnings of parenthetical units. The performance is high with respect to recognizing sentence boundaries and ends of parenthetical units. The performance with respect to identifying

Table 7.3
Confusion matrix for the decision-tree, nonbinary classifier (the Brown corpus)

Action		(a)	(b)	(c)	(d)	(e)
sentence-break	(a)	272				4
edu-break	(b)		133		3	84
start-paren	(c)			4		26
end-paren	(d)				20	6
none	(e)	2	38	1	4	7555

sentence boundaries ($272/276 = 98.55\%$) appears to be close to that of systems aimed at identifying *only* sentence boundaries [Palmer and Hearst, 1997], whose accuracy is in the range of 98 to 99%.

7.3.3 The Shift-Reduce Action Identifier

Generation of Learning Examples All the steps taken by the annotators during the construction of the corpus of discourse trees were automatically logged. Initially, I intended to use the logs of the annotation process in order to construct a rhetorical parser that would mirror the way humans parse discourse. However, during the development of the corpus, I noticed that the annotation patterns used by human judges were too complicated to be managed by a simple learning paradigm. Frequently, annotators used complicated operations that are specific to Tree Adjoining Grammars [Joshi, 1987], they undid previous parsing decisions, and performed operations that modified globally the discourse structures they built (see [Marcu et al., 1999a] for details). To avoid complications, I chose the parsing paradigm discussed in Section 7.3.1.

The learning cases were generated automatically, in the style of Magerman [1995], by traversing in-order the final rhetorical structures built by annotators and by generating a sequence of discourse parse actions that used only SHIFT and REDUCE operations of the kinds discussed in Section 7.3.1. When a derived sequence is applied as described in the parsing model, it produces a rhetorical tree that is a one-to-one copy of the original tree that was used to generate the sequence. For example, the tree at the bottom of Figure 7.2—the tree found at the top of the stack at step $i + 4$—can be built if the sequence of operations given in 7.8 is performed.

SHIFT 12; SHIFT 13; REDUCE-ATTRIBUTION-NS; SHIFT 14; REDUCE-JOINT-NN; SHIFT 15; REDUCE-CONTRAST-SN; SHIFT 16; SHIFT 17; REDUCE-CONDITION-SN; SHIFT 18; SHIFT 19; REDUCE-APPOSITION-NS; REDUCE-ATTRIBUTION-NS; REDUCE-ELABORATION-NS. (7.8)

I associated one learning case to each shift-reduce action.

Features Used for Learning To make decisions with respect to parsing actions, the shift-reduce action identifier focuses on the three topmost trees in the stack and the first *edt* in the input list. I refer to these trees as the trees in focus. The identifier relies on the following classes of features.

Structural features

• Features that reflect the number of trees in the stack and the number of *edt*s in the input list

• Features that describe the structure of the trees in focus in terms of the type of textual units that they subsume (sentences, paragraphs, titles); the number of immediate children of the root nodes; the rhetorical relations that link the immediate children of the root nodes, etc.[2]

Lexical (cue-phrase-like) and syntactic features

• Features that denote the actual words and POS tags of the first and last two lexemes of the text spans subsumed by the trees in focus

• Features that denote whether the first and last units of the trees in focus contain potential discourse markers and the position of these markers in the corresponding textual units (beginning, middle, or end)

Operational features

• Features that specify what the last five parsing operations performed by the parser were. I could generate these features because, for learning, I used sequences of shift-reduce operations and not discourse trees.

Semantic-similarity-based features

• Features that denote the semantic similarity between the textual segments subsumed by the trees in focus. This similarity is computed by applying, in the style of Hearst [1997], a cosine-based metric on the morphed segments. If we represent two segments S_1 and S_2 as sequences of $\langle t, w(t)\rangle$ pairs, where t is a token and $w(t)$ is its weight, we can compute the similarity between the segments using the formula shown in 7.9 below, where $w(t)_{S_1}$ and

2. The identifier assumes that each sentence break that ends in a period and is followed by two '\n' characters, for example, is a paragraph break; and that a sentence break that does not end in a punctuation mark and is followed by two '\n' characters is a title.

$w(t)_{S_2}$ represent the weights of token t in segments S_1 and S_2 respectively.

$$sim(S_1, S_2) = \frac{\sum_{t \in S_1 \cup S_2} w(t)_{S_1} w(t)_{S_2}}{\sqrt{\sum_{t \in S_1} w(t)_{S_1}^2 \sum_{t \in S_2} w(t)_{S_2}^2}} \tag{7.9}$$

The weights of tokens are given by their frequencies in the segments.

• Features that denote Wordnet-based measures of similarity between the bags of words in the promotion sets of the trees in focus. I use fourteen Wordnet-based measures of similarity, one for each Wordnet relation [Fellbaum, 1998]. Wordnet-based similarities reflect the degree of synonymy, antonymy, meronymy, hyponymy, etc. between the textual segments subsumed by the trees in focus. The Wordnet-based similarities are computed over the tokens that are found in the promotion units associated with each segment. If we represent the words in the promotion units of two segments S_1 and S_2 as two sequences W_1 and W_2, we can compute the Wordnet-based similarities between the two segments using the formula shown in 7.10 below, where the function $\sigma_R(w_1, w_2)$ returns 1 if there exists a Wordnet relation of type R between the words w_1 and w_2, and 0 otherwise.

$$sim_{wordnetRelation}(W_1, W_2) = \frac{\sum_{w_1 \in W_1, w_2 \in W_2} \sigma_{wordnetRelation}(w_1, w_2)}{|W_1| \times |W_2|} \tag{7.10}$$

The Wordnet-based similarity function takes values in the interval [0,1]: the larger the value, the more similar with respect to a given Wordnet relation the two segments are.

In addition to these features that modeled the Wordnet-based similarities of the trees in focus, I also use $14 \times 13/2 = 91$ *relative* Wordnet-based measures of similarity, one for each possible pair of Wordnet-based relations. For each pair of Wordnet-based measures of similarity w_{r_1} and w_{r_2}, each relative measure (feature) takes the value $<$, $=$, or $>$, depending on whether the Wordnet-based similarity w_{r_1} between the bags of words in the promotion sets of the trees in focus is lower than, equal to, or higher than the Wordnet-based similarity w_{r_2} between the same bags of words. For example, if both the synonymy- and meronymy-based measures of similarity are 0, the relative similarity between the synonymy and meronymy of the trees in focus will have the value $=$.

A binary representation of these features yields learning examples with 2789 features/example.

Examples of Rules Specific to the Action Identifier Figure 7.6 shows some of the rules that were learned by the C4.5 program [Quinlan, 1993] using a binary representation of the features and learning cases extracted from the MUC corpus. Rule 1, which is very similar to a rule derived manually using the corpus in Section 6.2, specifies that if the last

Rule 1: **if** lastTag(Top−1) = “,” ∧
position(firstUnit(Top−1),”if”) = ’b’
then REDUCE-CONDITION-SN
Rule 2: **if** firstTag(Top) = WRB ∧
secondTag(Top) ≠ VBG ∧
position(firstUnit(Top),”when”) = ’b’ ∧
sim(Top,Unit) > 0.0793052
then REDUCE-BACKGROUND-CIRCUMSTANCE-NS
Rule 3: **if** lastTag(Top−1) = NNS ∧
firstTag(Top) = IN ∧
hyponymy(Top−1,Top) = synonymy(Top−1,Top)
then REDUCE-BACKGROUND-CIRCUMSTANCE-NS
Rule 4: **if** isParagraphEnd(Top) ∧
position(firstUnit(Top),”but”) = ’b’
then REDUCE-CONTRAST-NN
Rule 5: **if** noTreesInStack ≤ 2 ∧
noUnitsInList = 0 ∧
topRelation(Top−1) ≠ TEXTUAL
then REDUCE-TEXTUAL-NN

Figure 7.6
Examples of automatically derived shift-reduce rules

lexeme in the tree at position *top* − 1 in the stack is a comma and there is a marker “if” that occurs at the beginning of the text that corresponds to the same tree, then the trees at position *top* − 1 and *top* should be reduced using a REDUCE-CONDITION-SN operation. This operation will make the tree at position *top* − 1 the satellite of the tree at position *top*. If the *edt* at position *top* − 1 in the stack subsumes unit 1 in example 7.11 and the *edt* at position *top* subsumes unit 2, this reduce action will correctly replace the two *edt*s with a new rhetorical tree, that shown in Figure 7.7.

[If you refer to someone as a butt-head,[1]] [ordinarily speaking, no one is going to take that as any specific charge of any improper conduct or insinuation of any character trait.[2]] (7.11)

Rule 2 makes the tree at the *top* of the stack the BACKGROUND-CIRCUMSTANCE satellite of the tree at position *top* − 1 when the first word in the text subsumed by the *top* tree is “when”, which is a while-adverb (WRB), when the second word in the same text is not a gerund or past participle verb (VBG), and when the cosine-based similarity between the text subsumed by the top node in the stack and the first unit in the list of elementary discourse units that have not been shifted to the stack is greater than 0.0793052. If the *edt* at position *top* − 1 in the stack subsumes unit 1 in example 7.12 and the *edt* at position *top*

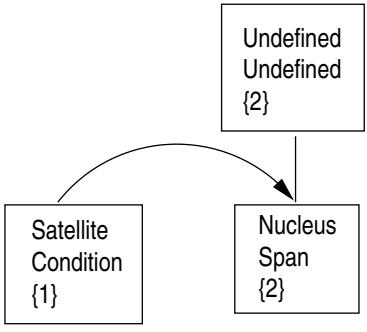

Figure 7.7
Result of applying rule 1 in Figure 7.6 on the *edt*s that correspond to the units in example 7.11

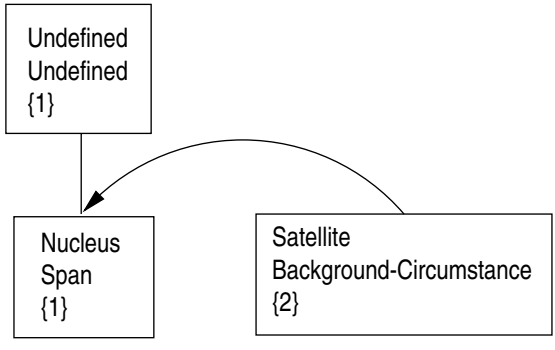

Figure 7.8
Result of applying rule 2 in Figure 7.6 on the *edt*s that correspond to the units in example 7.12

subsumes unit 2, rule 2 will correctly replace the two *edt*s with the rhetorical tree shown in Figure 7.8.

[Mrs. Graham, 76 years old, has not been involved in day-to-day operations at the company since May 1991,[1]] [when Mr. Graham assumed the chief executive officer's title.[2]] (7.12)

In case the last word in the text subsumed by the tree at position $top - 1$ in the stack is a plural noun (NNS), the first word in the text subsumed by the tree at the *top* of the stack is a preposition or subordinating conjunction (IN), and the hyponymy-based similarity between the two trees at the top of the stack is equal with their synonymy-based similarity, then the

[Some of the executives who attended yesterday's session weren't a surprise. Tenneco Inc. Chairman Michael Walsh, for instance, is a staunch Democrat who provided an early endorsement for Mr. Clinton during the presidential campaign. Xerox Corp.'s Chairman Paul Allaire was one of the few top corporate chief executive officers who contributed money to the Clinton campaign. And others, such as Atlantic Richfield Co. Chairman Lodwrick M. Cook and Zenith Electronics Corp. Chairman Jerry Pearlman, have also previously voiced their approval of Mr. Clinton's economic strategy.[1]]

[But some faces were fresh. Norman Augustine, the chairman of defense contractor Martin Marietta Corp., is a registered Republican who has never stood behind Mr. Clinton. It was also the first formal show of support by Rand Araskog, the chairman of ITT Corp.[2]]

Figure 7.9
Example of CONTRAST relation that holds between two paragraphs

action to be applied is REDUCE-BACKGROUND-CIRCUMSTANCE-NS. When this rule is applied in conjunction with the *edt*s that correspond to the units marked in 7.13, the resulting tree has the same shape as the tree shown in Figure 7.8.

[In an April 7 *Wall Street Journal* article, several experts suggested that IBM's accounting grew much more liberal since the mid-1980s[1]] [as its business turned sour.[2]] (7.13)

When the tree at the top of the stack subsumes a paragraph and starts with the marker "but", the action to be applied is REDUCE-CONTRAST-NN. For example, if the trees at the top of the stack subsume the paragraphs shown in Figure 7.9 and are characterized by promotion sets P1 and P2, as a result of applying rule 4 in Figure 7.6, one would obtain a new tree, whose shape is shown in Figure 7.10; the promotion units of the root node of this tree are given by the union of the promotion units of the child nodes.

The last rule in Figure 7.6 reflects the fact that each text in the MUC corpus is characterized by a title. When there are no units left in the input list (noUnitsInList = 0) and a tree that subsumes the whole text has been built (noTreesInStack <= 2), the two trees that are left in the tree—the one that corresponds to the title and the one that corresponds to the text—are reduced using a REDUCE-TEXTUAL-NN operation.

Evaluation The shift-reduce action identifier uses the C4.5 program in order to learn decision trees and rules that specify how discourse segments should be assembled into trees. In general, the tree-based classifiers performed slightly better than the rule-based classifiers. Table 7.4 displays the accuracy of the shift-reduce action identifiers, determined for each of the three corpora by means of a tenfold stratified cross-validation procedure. In Table 7.4, the B3 column gives the accuracy of a majority-based classifier, which chooses

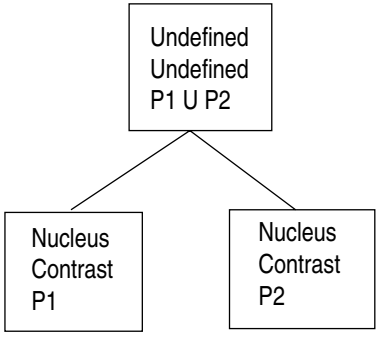

Figure 7.10
Result of applying rule 4 in Figure 7.6 on the trees that subsume the two paragraphs in Figure 7.9

Table 7.4
Performance of the tree-based, shift-reduce action classifiers. B3 corresponds to a baseline majority-based classifier that chooses action SHIFT in all cases. B4 corresponds to a baseline classifier that chooses shift-reduce operations randomly, with probabilities that reflect the probability distribution of the operations in each corpus.

Corpus	# cases	B3(%)	B4(%)	Acc(%)
MUC	1996	50.75	26.9	61.12±1.61
WSJ	4360	50.34	27.3	61.65±0.41
Brown	8242	50.18	28.1	61.81±0.48

action SHIFT in all cases. Since choosing only the action SHIFT never produces a discourse tree, in column B4, I also provide the accuracy of a baseline classifier that chooses shift-reduce operations randomly, with probabilities that reflect the probability distribution of the operations in each corpus.

Figure 7.11 shows the learning curve that corresponds to the MUC corpus. As in the case of the discourse segmenter, this learning curve suggests that more data can increase the accuracy of the shift-reduce action identifier.

7.3.4 Evaluation of the Decision-Based Rhetorical Parser

Obviously, by applying the two classifiers sequentially, one can derive the rhetorical structure of any text. Unfortunately, the performance results presented in Sections 7.3.2 and 7.3.3 only suggest how well the discourse segmenter and the shift-reduce action identifier perform with respect to individual cases. They say nothing about the performance of a rhetorical parser that relies on these classifiers.

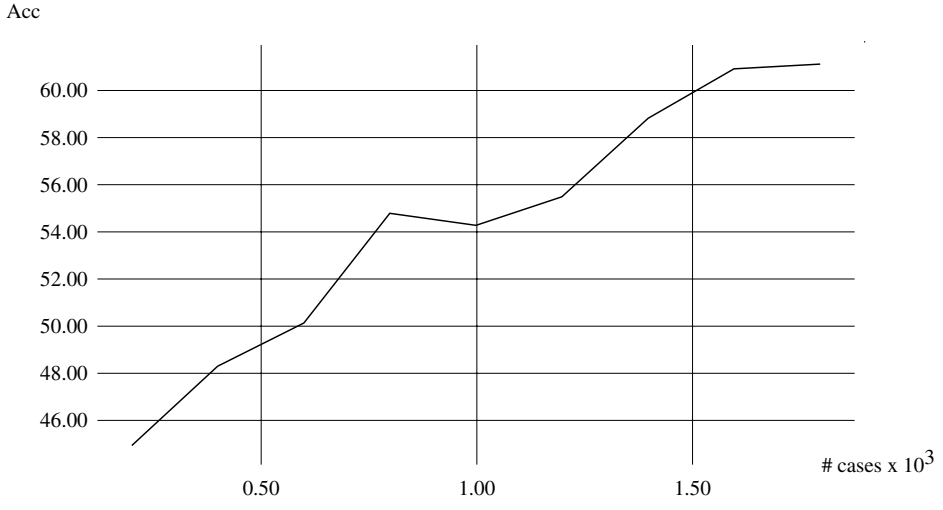

Figure 7.11
Learning curve for the shift-reduce action identifier (the MUC corpus)

In order to evaluate the rhetorical parser as a whole, I partitioned randomly each corpus into two sets of texts: twenty-seven texts were used for training and the last three texts were used for testing. The evaluation employs labeled recall and precision measures, as described in Section 6.3.7. Table 7.5 displays results obtained using segmenters and shift-reduce action identifiers that were trained either on twenty texts from each corpus (the MUC, WSJ, and Brown raws) and tested on three unseen texts from the same corpus; or that were trained on 27×3 texts from all corpora (the All raws) and tested on three unseen texts from each corpus. The training and test texts were chosen randomly. Table 7.5 also displays results obtained using a manual discourse segmenter (the M raws), which identified correctly all *edu*s. Since all texts in the three corpora were manually annotated by multiple judges, I could also compute an upper bound of the performance of the rhetorical parser by calculating, for each text in the test corpus and for each judge, the average labeled recall and precision figures with respect to the discourse trees built by the other judges. Table 7.5 displays these upper-bound figures as well. The upper-bound figures indicate the degree of agreement between judges on the task of manually constructing the corpus of discourse trees.

The results in Table 7.5 primarily show that errors in the discourse segmentation stage affect significantly the quality of the trees our parser builds. When a segmenter is trained only on twenty-seven texts (especially for the MUC and WSJ corpora, which have shorter

Table 7.5
Performance of the decision-based rhetorical parser: labeled (R)ecall and (P)recision. The segmenter is either Decision-Tree-Based (DT) or Manual (M).

Corpus	Segmenter	Train-	Elementary units				Hierarchical spans				Span nuclearity				Rhetorical relations			
		ing	Judges		Parser		Judges		Parser		Judges		Parser		Judges		Parser	
		corpus	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P
MUC	DT	MUC	88.0	88.0	37.1	100.0	84.4	84.4	38.2	61.0	79.1	83.5	25.5	51.5	78.6	78.6	14.9	28.7
	DT	All			75.4	96.9			70.9	72.8			58.3	68.9			38.4	45.3
	M	MUC			100.0	100.0			87.5	82.3			68.8	78.2			72.4	62.8
	M	All			100.0	100.0			84.8	73.5			71.0	69.3			66.5	53.9
WSJ	DT	WSJ	85.1	86.8	18.1	95.8	79.9	80.1	34.0	65.8	67.6	77.1	21.6	54.0	73.1	73.3	13.0	34.3
	DT	All			25.1	79.6			40.1	66.3			30.3	58.5			17.3	36.0
	M	WSJ			100.0	100.0			83.4	84.2			63.7	79.9			56.3	57.9
	M	All			100.0	100.0			83.0	85.0			69.0	82.4			59.8	63.2
Brown	DT	Brown	89.5	88.5	60.5	79.4	80.6	79.5	57.3	63.3	67.6	75.8	44.6	57.3	69.7	68.3	26.7	35.3
	DT	All			44.2	80.3			44.7	59.1			33.2	51.8			15.7	25.7
	M	Brown			100.0	100.0			81.1	73.4			60.1	67.0			59.5	45.5
	M	All			100.0	100.0			80.8	77.5			60.0	72.0			51.8	44.7

texts than the Brown corpus), it has very low performance. Many of the intrasentential *edu* boundaries are not identified, and as a consequence, the overall performance of the parser is low. When the segmenter is trained on 27 × 3 texts (the All raws), its performance increases significantly with respect to the MUC and WSJ corpora, but decreases with respect to the Brown corpus. This can be explained by the significant differences in style and discourse marker usage between the three corpora. When a perfect segmenter is used, the rhetorical parser determines hierarchical constituents and assigns them a nuclearity status at levels of performance that are not far from those of humans. However, the rhetorical labeling of discourse spans is even, in this case, to about 15–20% below human performance.

These results suggest that the features that I use may be sufficient for determining the hierarchical structure of texts and the nuclearity statuses of discourse segments. However, they are insufficient for determining correctly the elementary units of discourse and the rhetorical relations that hold between discourse segments.

8 Discussion

8.1 Related Work

The results presented in Part II of the book owe much to inspiration from recent developments in empirical discourse analysis. Particularly relevant is the work on segmenting discourse, distinguishing between discourse and sentential usages of cue phrases, and determining the correlation between cue phrases and discourse structure. I review briefly this work here, as well as relevant work on the discourse parsing of unrestricted texts.

8.1.1 Empirical Research on Discourse Segmentation

Empirical studies on discourse segmentation can be divided into two categories. In the first category, I include the studies that investigate the ability of human judges to agree on discourse segment boundaries. In the second, I include the studies aimed at deriving algorithms that would identify these boundaries.

Research on discourse segmentation has relied on various definitions of discourse segments. Discourse segments were defined in terms of Grosz and Sidner's discourse theory [1986]; in terms of an informal notion of topic [Hearst, 1997]; in terms of *transactions* [Carletta et al., 1997], i.e., subdialogues that accomplish one major step in the participants' plan for achieving a task; in terms of arguments [Burstein et al., 1998], i.e., rational presentations of points whose purpose is to persuade a reader; and in terms of intentional- and informational-based accounts that reflect the functional role of segments in text [Moser and Moore, 2000]. Studies performed on both text and speech [Grosz and Hirschberg, 1992, Nakatani et al., 1995, Hirschberg and Nakatani, 1996, Passonneau and Litman, 1993, Passonneau and Litman, 1996, Passonneau and Litman, 1997] have shown that humans agree consistently and reliably on segment boundaries when they use the intention-based definition proposed by Grosz and Sidner. Consistent and reliable agreement figures are obtained when the notions of transaction [Carletta et al., 1997] and topic [Hearst, 1997], and when the Relational Discourse Analysis methodology [Moser and Moore, 2000] are applied as well.

The studies aimed at deriving algorithms for the automatic identification of segment boundaries [Grosz and Hirschberg, 1992, Hirschberg and Litman, 1993, Passonneau and Litman, 1997, Moser and Moore, 2000, Di Eugenio et al., 1997] used sets of manually encoded linguistic and nonlinguistic features that pertained to prosody, cue phrases, referential links, intentional and informational structure of segments, types of relations, level of embedding, etc. The best algorithm that determines intention-based discourse segments recalled 53% of the discourse segments identified by humans, with a precision of 95% [Passonneau and Litman, 1997]. The algorithm was derived automatically using machine learning techniques. When instead of "intention" Hearst [1997] used "topic" as the

main criterion for assigning discourse segment boundaries, she showed that by exploiting word repetitions one can automatically find boundaries identified by humans with a recall of 59% and a precision of 71%. On different corpora, by applying statistical techniques and the noisy-channel model, Beeferman et al. [1999] have reported recall and precision figures of 57% and 60% respectively. By applying an entropy-based algorithm, Reynar [1999] has reported recall and precision figures of 59% and 60% respectively. Both Beeferman's and Reynar's approaches outperformed Hearst's algorithm on the corpora considered for their experiments.

In contrast with this previous work, the corpus studies discussed in Sections 6.2 and 7.2 were designed to enable the development of an algorithmic approach to identifying the *elementary* units of discourse and the *rhetorical relations* that held between discourse segments. Because the notions of intention [Grosz and Sidner, 1986], argument [Burstein et al., 1998], and topic [Hearst, 1997] yield discourse segments that are too coarse for this purpose, I could not use the algorithms described in this section in order to determine the elementary units of texts.

8.1.2 Empirical Research on Cue Phrase Disambiguation

Hirschberg and Litman [1993] showed that just by using the orthographic environment in which cue phrases occur, one can distinguish between sentential and discourse usages in about 80% of the cases and they suggested that cooccurrence data may provide useful information for cue phrase disambiguation. They also showed that POS tags can improve only slightly the disambiguation figures. In addition, Siegel and McKeown [1994] and Litman [1996] proved that Hirschberg and Litman's results [1993] can be improved up to figures in the range of 83% when genetic algorithms and machine learning techniques are used.

The corpus study and the algorithms presented in Chapter 6 have benefited extensively from the lessons learnt from Hirschberg and Litman's study. As we have seen, the orthographic environment and the neighboring cues played an important role in determining whether a given cue phrase had a discourse function in a text. The corpus analysis discussed in Chapter 6 was also meant to fill a coverage gap. Hirschberg, Litman, Siegel, and McKeown relied upon a corpus that had 953 occurrences of 34 cue phrases. In contrast, our corpus of cue phrases had 7600 occurrences of more than 450 cue phrases. The study discussed in Section 7.2 is the first attempt to develop a consistently annotated corpus of rhetorical structure trees. Marcu et al. [1999b] discusses at length the protocol used during the annotation, the agreement reached by the judges, and the annotation issues that still need to be provided adequate solutions.

8.1.3 Empirical Research on the Discourse Function of Cue Phrases

Most empirical research on cue phrases has focused on very specific facets. For example, Di Eugenio [1992, 1993] and Delin et al. [1994] studied the role of *by* and *to* in purpose clauses; Grote et al. [1997] studied the role of *but* and *although* in concessive relations; Anscombre and Ducrot [1983], Cohen [1983], and Elhadad and McKeown [1990] studied the role of *since* and *because* in argumentation; Hirschberg and Litman [1987] studied the relationship between the discourse usage of *now* and intonation; and Moens and Steedman [1988] studied the role of *before*, *after*, and *when* in temporal discourse. In an exploratory study of the relationship between discourse markers, pragmatics, and discourse, Schiffrin [1987] provided a careful sociolinguistic analysis of dialogue usages of *and, then, so, because,* and *but*. Sweetser [1990] studied the role of discourse markers in the context of etymology and pragmatics. A broad empirical investigation of cue phrases was carried out by Knott [1995], Knott and Dale [1996], and Knott and Mellish [1996] in order to motivate on psycholinguistic bases a taxonomy of coherence relations.

The corpus analysis that comes closest to ours is that of Moser and Moore [1995, 2000]. They collected a set of seventeen student-tutor interactions encompassing 144 question-answer exchanges that had 854 clauses. For each interaction in the corpus, the analysts determined the elementary and nonelementary discourse constituents and the discourse relations that hold between them. The analysts also labeled the functional status of the segments, i.e., they distinguished between segments that expressed what was essential to the writer's purpose—these were called *core* segments—and the segments that served the purpose manifested by the core—these were called *contributors*. They also labeled the syntactic relation between segments (independent sentences, coordinated clauses, subordinated clauses), the relative order of the core and contributors, the cue phrases associated with various segments, etc. The most important finding by Moser and Moore was that the placement of cue phrases correlates with both the functional status of the segment to which they belong and the linear order of the core and contributor segments.

As an extention to Moser and Moore's analysis, Di Eugenio, Moore, and Paolucci [1997] have investigated the possibility of using the same corpus data for deriving algorithms that would enable a natural language generation system to determine when and how to use cue phrases in explanatory texts. Decision trees that were derived using traditional machine learning techniques showed that the ordering of the core and contributor was crucial for determining whether a cue phrase needed to be used.

Although Moser and Moore's corpus analysis implemented many of the features that are present in my studies, it had a relatively narrow coverage. Because the motivation for their corpus analysis was given primarily by unsolved problems in the field of natural

language generation, it did not encode information that would enable the development of algorithms for determining the discourse segments of a text. In contrast, the corpus study in Section 6.2 was designed to permit the development of algorithms that can be used not only to determine the elementary units of discourse, but also the rhetorical relations that hold between units and spans of text. And the corpus study in Section 7.2 was designed to enable the application of learning algorithms in order to automatically generate rules for deriving the rhetorical structure of texts.

8.1.4 Research on Discourse Parsing of Free Texts

At the time the cue-phrase-based rhetorical parser was developed, there was no rhetorical parser for English. The research that came closest to that described in Chapter 6 was that of Sumita et al. [1992] and of Kurohashi and Nagao [1994].

Sumita et al. [1992] report on a discourse analyzer for Japanese. The key difference between the work described here and Sumita's comes from 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, the cue-phrase-based rhetorical parser uses a mathematical model in which this ambiguity is acknowledged and appropriately treated. In addition, the decision-based rhetorical parser presented in Chapter 7 can automatically learn rules for determining the elementary discourse units and for hypothesizing discourse relations that are dependent on the context encoded by the features.

Another key difference comes from the fact that the discourse trees built by the rhetorical parsers described here are very constrained structures (see Chapter 3). As a consequence, the rhetorical parsers do not overgenerate invalid trees as Sumita et al.'s does. Furthermore, the cue-phrase-based 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.

As in the case of Sumita's system, Kurohashi and Nagao's also takes as input a sequence of parse trees and assume that the elementary units of the discourse are sentences. However, in contrast to Sumita et al., 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], Asher and Lascarides [1993],

and Cristea and Webber [1997]. 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. Our empirical experiments on manual discourse structure derivation suggested that, in many cases, humans cannot construct a discourse tree incrementally, without applying any form of backtracking [Marcu et al., 1999a]. To circumvent this problem, the decision-based rhetorical parser described in Chapter 7 derives rhetorical trees using a shift-reduce parsing model. In this model, partial trees are put together into larger discourse trees only after sufficient information has been obtained. For example, rule 4 in Figure 7.6, which was learned automatically, can be applied only after the discourse tree that corresponds to the entire paragraph that starts with the conjunction "But" has been derived.

Corston-Oliver [1998] recently elaborated on the work described in Chapter 6 and investigated the possibility of using syntactic information in order to hypothesize relations. His system uses thirteen rhetorical relations and builds discourse trees for articles in Microsoft's *Encarta 96 Encyclopedia*. Corston uses the formal model of discourse that was presented in Chapter 3 and implements the procedural account of discourse structure derivation in a manner that is very similar to that implemented in the chart parser (see Section 6.3.4).

All these previous parsers, aimed at determining the rhetorical structure of unrestricted texts [Sumita et al., 1992, Kurohashi and Nagao, 1994, Corston-Oliver, 1998], employed manually written rules. Because of the lack of discourse corpora, their authors could not evaluate the correctness of the discourse trees they built. And because of this, we cannot compare the performance of others' parsers with the performance of the rhetorical parsers discussed in this book. As a matter of fact, we cannot even properly compare the performance of the cue-phrase- and decision-based parsers because the rhetorical relations that they use are different and because one parser derives binary trees, while the other derives nonbinary trees.

A parallel line of research has recently been investigated by Hahn and Strube [1997]. 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 parsers discussed in this book: their algorithms

rely primarily on coreference 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. It is likely that a deeper understanding of the relation between discourse structures and the use of referential expressions could improve the performance of the rhetorical parsers described here.

8.2 Open Problems

STUDYING THE RELATION BETWEEN SYNTAX AND DISCOURSE. Designing and carrying out corpus analyses are not trivial tasks. Due to their exploratory nature, each of the corpus analyses discussed in this part of the book was started over many times. Deciding what information to annotate and how to carry out an annotation could be an endless task, because once you start dealing with real texts, it is very likely that you will soon come across examples that cannot be represented properly using the annotation rules that you have designed. Improving iteratively an annotation protocol is, hence, more of a norm than an exception.

The two corpus analyses discussed in Part II confirm the condition just described. They reflect the evolution of my degree of understanding of discourse and of the expectations and goals that a corpus analysis is supposed to serve. The lack of objective definitions of elementary discourse units and rhetorical relations is the most significant shortcoming of the empirical work in this book. Unfortunately, this shortcoming reflects the general, current understanding of discourse phenomena: providing unambiguous definitions for these still constitute open research questions.

Just reconsider, for instance, the problem of determining the elementary units of discourse. In example 8.1, which is given below, it seems obvious that one should consider that the text consists of two units between which an ELABORATION relation holds.

[This is a nice book.] [It was written by Bertrand Russell.] (8.1)

[This is a nice book,] [which was written by Bertrand Russell.] (8.2)

[This is a nice book,] [written by Bertrand Russell.] (8.3)

[This is a nice [Bertrand Russell] book.] (8.4)

However, in example 8.2, the same ELABORATION is conveyed by means of a nonrestrictive relative; in 8.3, it is conveyed by means of a reduced relative; and in 8.4, it is conveyed by means of an adjectival phrase. Which of these units should we treat as ele-

mentary? When does a rhetorical analysis become syntactic analysis and vice versa? This book offers no adequate answers to such questions.[1]

IMPROVING THE PERFORMANCE OF THE DISCOURSE PARSERS. The rhetorical parsing algorithms presented in this book rely primarily on cue-phrases, POS tags, Wordnet relations, cohesion, and a well-constrained model of discourse trees in order to determine the rhetorical structure of texts. As the evaluation results show, the rhetorical structures that can be derived by relying only on these types of resources do not match very well those built by humans. A visual inspection of the trees in Figure 1.3 and Figure 6.19 helps one identify immediately some of the problems.

As one can see, the cue-phrase-based rhetorical parser is not able, for example, 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 identifies correctly 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. As we have seen in Chapter 7, the decision-tree-based rhetorical parser is capable of learning rules that identify correctly locations where an elementary unit boundary should be inserted before the occurrence of an *and* (see rule 4 in Figure 7.4); yet, the decision-based parser needs much more training data in order to approach human performance levels.

Also, by examining the trees in Figures 1.3 and 6.19, one can note that although the cue-phrase-based rhetorical parser identified correctly the hierarchical segments and the nuclearity statuses in the first paragraph, the cue-phrase-based parser 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.

The decision-tree-based parser was able to learn more sophisticated rules for determining the rhetorical relations that hold between various spans. And some of these rules exploited not only the occurrences of cue phrases, but also the lexical relations that held between the content words of the discourse segments under scrutiny. Yet, the decision-tree-based parser was unable to produce trees whose rhetorical labels matched closely those

1. In recent work Marcu [1999c] addresses this problem by making explicit the conventions that one should use in order to identify the elementary units of texts and by enabling annotators to build text structures that blend the syntactic and rhetorical levels of representation.

assigned by humans even when it took as input perfectly segmented data. This is not surprising. After all, some of the rhetorical relations are defined intentionally. So, without a deep semantic and intentional analysis of the texts given as input, it is unlikely that such relations can be correctly determined. 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. However, it is not very clear what forms of reasoning one should use in order to derive, for unrestricted texts, relations that are as difficult to infer as the EVIDENCE relation in Figure 1.3.

The discussion so far concerned mostly the bad news. Fortunately, there is also a good side to the story. The results in this book suggest that the rhetorical structures derived by our rhetorical parsers can be used successfully in the context of text summarization. Hence, although the rhetorical parsers do not get the RS-trees perfectly right, they still can be used to determine the important units of text at levels of performance that are not far from those of humans. One possible explanation can be found in the fact that the rhetorical-based summarizer that is to be described in Section 9 exploits only the difference between satellites and nuclei and the hierarchical structure of text in order to determine text units that are important.

The rhetorical summarizer is a niche application that shows that by understanding the hierarchical organization of text one can solve difficult natural language problems. It is possible that discourse structures of the kinds derived by our 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 of 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 answered better if one has access to rhetorical structure representations of the texts that contain the answers to the questions. However, how much of an impact the rhetorical parsers presented here can have on solving these problems remains an empirical open question.

EXTENSIONS. The rhetorical parsers presented in Part II can take advantage of other lexicogrammatical constructs that indicate rhetorical relations. In what follows, I review a few.

Grammatical morphemes. As argued, for example, by Talmy [1983] and Morrow [1986], grammatical morphemes often express notions that are more schematic than those expressed by content words. For instance, a combination of a shift from past to present tense and from third to first person correlates both with a shift from impersonal narration to di-

rect report or monologue and a shift in participant's perspective [Morrow, 1986, p. 434]. And psycholinguistic research shows that readers are more likely to consider a collection of sentences as being related if they contain the definite article "the", instead of the indefinite article "a" [de Villiers, 1974, Gernsbacher, 1997].

A rhetorical parser that encodes these findings is likely to achieve better results than the parsers presented in this book.

Tense and aspect. Decker [1985], Morrow [1986], Moens and Steedman [1988], Webber [1988b], Lascarides and Asher [1993], Barker and Szpakowicz [1995], and Hitzeman [1995] show that the tense and aspect of verbs provide clues to the discourse structure of a text. These clues may be genre dependent and may be applied in isolation or in conjunction with other features. For example, in narratives, the use of the present tense tends to express situations occurring at the time of narration [Kamp, 1979]. In the context of news reports, the use of simple past verbs in simple sentences usually corresponds to foreground material (see the use of the verb *meet* in example 8.5); but the use of simple past verbs in relative clauses usually corresponds to background material (see the use of the verb *engineer* in example 8.6) [Decker, 1985].

After weeks of maneuvering and frustration, presidential envoy Richard B. Stone *met* face to face yesterday for the first time with a key leader of the Salvadoran guerilla movement. [Decker, 1985, p. 317] (8.5)

"The ice has been broken," proclaimed President Belisario Betancur of Colombia, who *engineered* the meeting. [Decker, 1985, p. 317] (8.6)

The semantics of certain verbs also conveys information about discourse relations in the cases in which some tense constraints are enforced. For example, in Lascarides and Asher's [1993] formalization of discourse relations, the event of pushing associated with the second sentence in example 8.7 is normally assumed to have produced the event of falling associated with the first sentence, if the pushing event occurred before the falling event.

Max fell. John pushed him. [Lascarides and Asher, 1993] (8.7)

Hence, a causal relation is normally assumed to hold between the sentences in 8.7.

A rhetorical parser that exploits information specific to the tense and aspect of the verbs may achieve better results.

Syntactic constructs. Prince [1978] and Delin and Oberlander [1992] have observed that cleft constructions could serve a subordinating function in discourse. The information conveyed by a cleft sentence concerns some background material against which the related sentences have to be interpreted; a cause whose effect is given in the related sentences; or some background material that not only is subordinated to the related sentences but that also mentions events that occurred prior to those described in the related sentences. For example, the cleft sentence shown in italics in text 8.8 provides background information for the preceding text and must be interpreted as describing events that occurred prior to the events described in the preceding text [Delin and Oberlander, 1992, p. 282].

Mr. Butler, the Home Secretary, decided to meet the challenge of the 'Ban-the-Bomb' demonstrators head-on. Police leave was cancelled and secret plans were prepared. *It was Mr. Butler who authorized action which ended in 32 members of the Committee of 100 being imprisoned.* The Committee's president and his wife were each jailed for a week. (8.8)

In order to recognize rhetorical relations, Corston-Oliver [1998] also relies on syntactic constraints. For example, in his corpus analysis of Microsoft's *Encarta 96 Encyclopedia*, Corston-Oliver noticed that whenever a CONTRAST relation held between two clauses, none of the clauses was subordinated to the other; hence, he encoded this as a necessary criterion for recognizing CONTRAST relations. Also, he noticed that in many of the cases in which the syntactic subject of a clause was the pronoun *some* (or had the modifier *some*) and the subject of the other clause was the pronoun *other* (or had the modifier *other*), the relation between the two clauses was CONTRAST as well; since, this observation was violated in some cases, it was considered to be only a cue for recognizing CONTRAST relations.

It is very likely that a tight coupling of a discourse and syntactic parser can increase our ability to automatically derive the rhetorical structure of texts.

Pronominalization and anaphoric usages. Sidner [1981], Grosz and Sidner [1986], Fox [1987], Sumita [1992], and Grosz, Joshi, and Weinstein [1995], and Hoover [1997] have speculated that certain patterns of pronominalization and anaphoric usages correlate with the structure of discourse. Vonk's psycholinguistic work [1992] has confirmed that anaphoric expressions that are more specific than necessary for their identification function not only establish coreference links but also contribute to the signaling of thematic shifts. For example, in the sequence of sentences given in 8.9, which is taken from [Vonk et al., 1992, p. 303], the use of *She* in sentence 5 poses no referential problem. However, the use of *Sally*, which is more specific than necessary, would sound better because it suggests a topic shift.

1. Sally Jones got up early this morning.
2. She wanted to clean the house.
3. Her parents were coming to visit her.
4. She was looking forward to seeing them. (8.9)
5. *She/Sally* weighs 80 kilograms.
6. She had to lose weight on her doctor's advice.
7. So she planned to cook a nice but sober meal.

In fact, Vonk's experiments not only show that readers are typically led to infer a theme shift when encountering an overspecification, but also that overspecifications cause a decrease in the availability of words from the preceding text [Vonk et al., 1992, p. 326].

More recent empirical evidence collected by Passonneau [1998, 1997] also suggests that overly informative discourse anaphoric expressions occur at shifts in global discourse focus. More specifically, Passonneau's experiments suggest that there exists a correlation between the usage of overly informative anaphoric expressions and the intention-based, discourse segments that pertain to Grosz and Sidner's discourse theory [1986]. A parallel line of research is explored by Walker [1998], who proposes that the relationship between anaphoric usages and discourse structure can be best explained with a model of attention that distinguishes between the long-term and the short-term (working) memory [Walker, 1996]. The same concept is explored by Givón [1995], in a psycholinguistic setting.

Enabling the rhetorical parsing algorithms to take advantage of any of these rhetorical structure indicators may improve significantly their performance. Relying on different parsing models, such as probabilistic context free grammars, and finding better ways of exploiting the information encoded in the corpora can also improve the performance of the rhetorical parsers.

8.3 Summary

In this part of the book, I have presented a variety of linguistic constructs that can be used to detect the elementary textual units in a text and the rhetorical relations that hold among them. I then discussed the assumptions that constitute the foundations of a surface-based approach to text structure derivation, one that relies primarily on cue phrases and lexicogrammatical constructs that can be detected without a deep syntactic and semantic analysis. I described an exploratory corpus study of cue phrases and a cue-phrase-based rhetorical parser that is deeply rooted in this corpus study. This parser relies on the following algorithms:

- A surface-form algorithm that uses information derived from the corpus analysis of cue phrases in order to determine the elementary units of a text and the cue phrases that have a discourse structuring function.
- An algorithm that uses information that was derived from the corpus analysis in order to hypothesize exclusively disjunctive rhetorical relations that hold between elementary units and text spans.
- An algorithm that uses word cooccurrences in order to hypothesize exclusively disjunctive rhetorical relations that hold between text spans.
- A chart-parsing algorithm that implements the proof-theoretic account of the problem of text structure derivation.

In Chapter 7, I presented a shift-reduce rhetorical parsing algorithm that learns to construct rhetorical structures of texts from tagged data. The parser has two components: a discourse segmenter, which identifies the elementary discourse units in a text; and a shift-reduce action identifier, which determines how these units should be assembled into rhetorical structure trees.

The evaluation of the shift-reduce rhetorical parser, which was carried out on a collection of ninety discourse trees that were manually annotated, suggests that a high-performance discourse segmenter would need to rely on more training data and more elaborate features than the ones described in this book—the learning curves did not converge to performance limits. If one's goal is, however, to construct discourse trees whose leaves are sentences (or units that can be identified at high levels of performance), then the segmenter described here appears to be adequate. The evaluation results also suggest that the rich set of features that constitute the foundation of the action identifier may be sufficient for constructing discourse hierarchies and for assigning to discourse segments a rhetorical status of nucleus or satellite at levels of performance that are close to those of humans. However, more research is needed in order to approach human performance in the task of assigning to segments correct rhetorical relation labels.

III SUMMARIZATION

Preamble

In the last couple of years, the word "summary" has been associated with a variety of meanings and has been used in a variety of contexts (see [Sparck Jones, 1999, Hovy and Lin, 1999] for a list of abstract dimensions along which summaries can be classified). Depending on the input, one can have *single-* or *multiple-document* summaries. Depending on the output, one can have *extract-* or *abstract-like* summaries. An *extract* is a summary that lists the important clauses, sentences, and paragraphs of the input text; an *abstract* is a summary that fuses and rewrites the important clauses, sentences, and paragraphs of the input, so that the resulting text is coherent. Depending on the usage, a summary can be *indicative*, i.e., it can provide only an indication of the main topics in the input text, or *informative*, i.e., it can reflect to a certain extent the semantic content of the input text as well. Depending on the purpose, a summary can be *generic*, i.e., it can reflect the author's point of view with respect to all important topics in the input text, or it can be *query-oriented*, i.e., it can reflect only the topics in the input text that are specific to a given query. In this part of the book, I will focus on studying the relationship between rhetorical structure and single-document, informative, generic extracts.

The rhetorical parsers presented in Part II not only construct discourse structures that make explicit the rhetorical relations between different spans of text but also assign to each node in a discourse tree the elementary units of its promotion set. In what follows, I show how one can use the text structures and the promotion units associated with them in order to determine the most important parts of a text. In Section 9.1, I show how, starting from its text structure, one can induce a partial ordering on the importance of the units in a text and I propose a discourse-based summarization algorithm. I then discuss general issues concerning the evaluation of automatically generated summaries and propose that we should evaluate not only the results of the programs that we build, but also the assumptions that constitute their foundations. Hence, I design an experiment to test whether the assumption that text structures can be used effectively for text summarization is valid (Section 9.2.2). The experiment confirms that there exists a strong correlation between the nuclei of a text structure and what readers perceive as being important in the corresponding text.

In Section 9.2.3, I evaluate a straightforward implementation of the discourse-based summarization algorithm that uses the cue-phrase-based rhetorical parser in Chapter 6. In Chapter 10, I show how this rhetorical parser can be tuned by means of machine learning techniques in order to produce trees that reflect closely what humans consider to be important. The tuning process integrates within the discourse-based framework several other indicators of textual importance that have been proposed in the literature. I discuss strengths and weaknesses of the discourse-based summarizer and end this part of the book with a review of related work and with suggestions of rhetorical-parsing applications.

9 Summarizing Natural Language Texts

9.1 From Discourse Structures to Text Summaries

9.1.1 From Discourse Structures to Importance Scores

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. A simple inspection of the structure in Figure 9.2, for example, allows us to determine that, according to the formalization in Chapter 3, unit 2 is the most important textual unit in the text in Figure 9.1 because it is the only promotion unit associated with the root node. Similarly, we can determine that unit 3 is the most important unit of span [3,6] and that units 4 and 5 are the most important units of span [4,6]. (The tree in Figure 9.2 is the same as the tree in Figure 6.19; and the text in Figure 9.1 is the same as the text in Figure 6.8. The only difference between these two texts concerns the labeling of the parenthetical units. In Figure 9.1, they are labeled with strings having the form *Pn*, where *n* denotes the elementary unit to which the parenthetical unit is related. In Figure 6.8, the parenthetical units were not labeled. The text in Figure 6.8 has been replicated here only for convenience.)

A more general way of exploiting the promotion units that are associated with a discourse tree is from the perspective of text summarization. If we repeatedly apply the concept of salience to each of the nodes of a discourse structure, we can induce a partial ordering on the importance of all the units of a text. The intuition behind this approach is that the textual units that are 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. A very simple way to induce such an ordering is by computing a score for each elementary unit of a text on the basis of the depth in the tree structure of the node where the unit occurs first as a promotion unit. The larger the score of a unit, the more important that unit is considered to be in a text. Formula 9.1, which is given below, provides a recursive definition for computing the importance score of a unit u in a discourse structure D that has depth d.

$$score(u, D, d) = \begin{cases} 0 & \text{if } D \text{ is NIL,} \\ d & \text{if } u \in promotion(D), \\ d-1 & \text{if } u \in parentheticals(D), \\ \max(score(u, leftChild(D), d-1), & \text{otherwise} \\ \quad score(u, rightChild(D), d-1)) & \end{cases} \quad (9.1)$$

The formula assumes that the discourse structure is a binary tree and that the functions *promotion*(*D*), *parentheticals*(*D*), *leftChild*(*D*), and *rightChild*(*D*) return the promotion set, parenthetical units, and the left and right subtrees of each node respectively. If a unit is among the promotion set of a node, its score is given by the current value of d. If a

[*With* its distant orbit {—50 percent farther from the sun than Earth—^{P1}} 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,^{P10}} temperatures never warm enough to melt frozen water.[10]]

Figure 9.1
The *Mars* text

unit is among the parenthetical units of a node, which can happen only in the case of a leaf node, the score assigned to that unit is $d - 1$ because the parenthetical unit can be represented as a direct child of the elementary unit to which it is related. For example, when we apply formula 9.1 to the tree in Figure 9.2, which has depth 6, we obtain the scores in Table 6.1 for each of the elementary and parenthetical units of the text in Figure 9.1. Because unit 2 is among the promotion units of the root, it gets a score of 6. Unit 3 is among the promotion units of a node found two levels below the root, so it gets a score of 4. Unit 6 is among the promotion units of a leaf found 5 levels below the root, so it gets a score of 1. Unit $P1$ is a parenthetical unit of elementary unit 1, so its score is $score(1, D, 6) - 1 = 3 - 1 = 2$ because the elementary unit to which it belongs is found 3 levels below the root.

If we consider now the importance scores that are induced on the textual units by the discourse structure and formula 9.1, we can see that they correspond to a partial ordering on the importance of these units in a text. This ordering enables the construction of text summaries with various degrees of granularity. Consider, for example, the partial ordering shown in 9.2, which was induced on the textual units of the text in Figure 9.1 by the discourse structure in Figure 9.2 and formula 9.1.

$$2 > 8 > 3, 10 > 1, 4, 5, 7, 9 > P1, P10 > 6 \tag{9.2}$$

If we are interested in generating a very short summary of the text in Figure 9.1, we can create a text with only one unit, which is unit 2. A longer summary can contain units 2 and 8. A longer one, units 2, 8, 3, and 10. And so on.

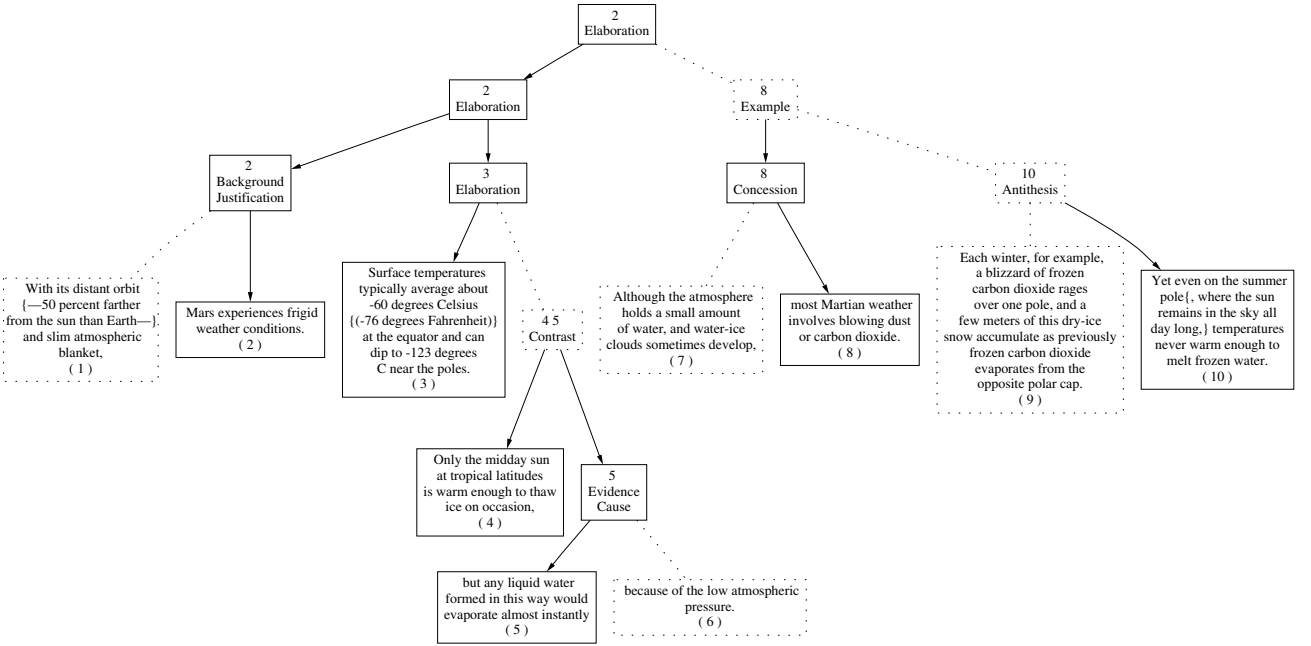

Figure 9.2
The discourse tree of maximal weight that is built by the rhetorical parsing algorithm for the text in Figure 9.1

Table 9.1
The importance scores of the textual units in the text in Figure 9.1

Unit	1	*P*1	2	3	4	5	6	7	8	9	10	*P*10
Score	3	2	6	4	3	3	1	3	5	3	4	2

Input: A text T
A number p, such that $1 \leq p \leq 100$.

Output: The most important $p\%$ of the elementary units of T.

1. I. Determine the discourse structure DS of T by means of the rhetorical
2. parsing algorithm in Figure 6.5.
3. II. Determine a partial ordering on the elementary and parenthetical
4. units of DS by means of formula (9.1).
5. III. Select the first $p\%$ units of the ordering.

Figure 9.3
The discourse-based summarization algorithm

The idea of using discourse structures for constructing text summaries is not new. Researchers in computational linguistics have long speculated that the nuclei of a rhetorical structure tree constitute an adequate summarization of the text for which that tree was built [Mann and Thompson, 1988, Matthiessen and Thompson, 1988, Hobbs, 1993, Polanyi, 1993, Sparck Jones, 1993a, Sparck Jones, 1993b]. Using the partial orderings induced by formula 9.1 on the text structures derived by the rhetorical parser is only a precise expression of the original intuition.

9.1.2 A Discourse-Based Summarizer

Given that we can use the cue-phrase-based rhetorical parser described in Chapter 6 to build the discourse structure of any text and that we can use formula 9.1 to determine the partial ordering that is consistent with the idea that the nuclei of a discourse structure constitute a good summary of a text, it is trivial now to implement a summarization program. Figure 9.3 presents a discourse-based summarization program that takes two arguments: a text and a number p between 1 and 100. The program first uses the cue-phrase-based rhetorical parser in order to determine the discourse structure of the text given as input. It then applies formula 9.1 and determines a partial ordering on the elementary and parenthetical units of the text. It then uses the partial ordering in order to select the $p\%$ most important textual units of the text.

9.2 Evaluation of the Rhetorical-Based Approach to Summarization

9.2.1 General Remarks

Since I am focusing on single-document, informative, generic extracts, I adopt an evaluation methodology that is aimed at determining how well a discourse-based summarizer can select the important units in a text.[1] Unfortunately, since the discourse trees that the cue-phrase-based rhetorical parser derives are not perfect (see Section 6.3.7), a direct evaluation of the program would say nothing about the adequacy of the discourse-based method for text summarization. The position that I advocate here (and elsewhere [Marcu, 1999a]) is that in evaluating various summarization techniques, one should distinguish between the evaluation of a method and the evaluation of a particular implementation of it. If one does so, one is able to distinguish the problems that pertain to a particular implementation from those that pertain to the underlying theoretical framework and explore new ways to improve each.

The experiment described in this section is aimed at accounting for the difference between evaluating a method and evaluating an implementation of it. As I have mentioned already, researchers in computational linguistics have long speculated that the nuclei of a rhetorical structure tree constitute an adequate summarization of the text for which that tree was built [Mann and Thompson, 1988, Matthiessen and Thompson, 1988, Sparck Jones, 1993b]. However, to my knowledge, there has been no experiment to confirm how valid this speculation really is. In what follows, I describe an experiment that shows that there exists a strong correlation between the nuclei of the discourse tree of a text and what readers perceive to be the most important units in a text. The experiment shows that the concepts of discourse structure and nuclearity *can* be used effectively for determining the most important units in a text. I then also evaluate the performance of the discourse-based summarization program in Figure 9.3.

9.2.2 From Discourse Structure to Extracts—an Empirical View

Materials and Methods of the Experiment We know from the results reported in the psychological literature on summarization [Johnson, 1970, Chou Hare and Borchardt, 1984, Sherrard, 1989] that there exists a certain degree of disagreement between readers with respect to the importance that they assign to various textual units and that the disagreement is dependent on the quality of the text and the comprehension and summarization skills of the readers [Winograd, 1984]. In my experiment, I used the same five *Scientific*

1. For a more extensive discussion of summary evaluation see [Hovy and Lin, 1999].

[With its distant orbit[1]] [—50 percent farther from the sun than Earth—[2]] [and slim atmospheric blanket,[3]] [Mars experiences frigid weather conditions.[4]] [Surface temperatures typically average about −60 degrees Celsius (−76 degrees Fahrenheit) at the equator[5]] [and can dip to −123 degrees C near the poles.[6]] [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,[7]] [but any liquid water formed in this way would evaporate almost instantly[8]] [because of the low atmospheric pressure.[9]]

[Although the atmosphere holds a small amount of water,[10]] [and water-ice clouds sometimes develop,[11]] [most Martian weather involves blowing dust or carbon dioxide.[12]] [Each winter, for example, a blizzard of frozen carbon dioxide rages over one pole,[13]] [and a few meters of this dry-ice snow accumulate[14]] [as previously frozen carbon dioxide evaporates from the opposite polar cap.[15]] [Yet even on the summer pole,[16]] [where the sun remains in the sky all day long,[17]] [temperatures never warm enough to melt frozen water.[18]]

Figure 9.4
The *Mars* text, as was given to the subjects

American texts referred to in Section 6.3.7. The shortest text was the text on Mars that I have used as an example throughout the book.

Because my intention was to evaluate the adequacy for summarizing text not only of the program that I implemented but also of the discourse theory that I developed, I first determined manually the minimal textual units of each text. Overall, I broke the five texts into 160 textual units with the shortest text being broken into eighteen textual units, and the longest into seventy. Each textual unit was enclosed within square brackets and labeled in increasing order with a natural number from 1 to N, where N was the number of units in each text. For example, when the text on Mars was manually broken into elementary units, I obtained not ten units, as is the case when the discourse-marker and clause-like unit identification algorithm was applied (see Figure 9.1), but eighteen. The text whose minimal units were obtained manually is given in Figure 9.4.

I followed Johnson's [1970] and Garner's [1982] strategy and asked 13 independent judges to rate each textual unit according to its importance to a potential summary. The judges used a three-point scale and assigned a score of 2 to the units that they believed to be very important and should appear in a concise summary, 1 to those they considered moderately important, which should appear in a long summary, and 0 to those they considered unimportant, which should not appear in any summary. The judges were instructed that there were no right or wrong answers and no upper or lower bounds with respect to the number of textual units that they should select as being important or moderately important. The judges were all graduate students in computer science at the University of Toronto; I assumed that they had developed adequate comprehension and summarization skills on

Table 9.2
The scores assigned by the judges, analysts, and the discourse-based summarizer to the textual units of the text in Figure 9.4

Unit	Judges													Analysts		Program
	1	2	3	4	5	6	7	8	9	10	11	12	13	1	2	
1	0	2	2	2	0	0	0	0	0	0	0	0	0	3	3	3
2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2
3	0	2	0	2	0	0	0	0	0	0	0	0	1	3	2	3
4	2	1	2	2	2	2	2	2	2	2	2	2	2	6	5	6
5	1	1	0	1	1	1	0	1	2	1	0	2	2	4	3	4
6	0	1	0	1	1	1	0	1	1	1	0	2	2	4	3	4
7	0	2	1	0	0	0	1	1	1	0	0	0	0	4	3	3
8	0	1	0	0	0	0	0	0	0	0	0	0	0	4	3	3
9	0	0	2	0	0	0	0	0	0	0	1	0	1	1	0	1
10	0	2	2	2	0	0	2	0	0	0	0	0	0	3	4	3
11	0	0	0	2	0	0	0	1	0	0	0	0	1	3	4	3
12	2	2	2	2	2	2	2	2	2	0	1	2	2	5	4	5
13	1	1	0	0	0	1	0	1	0	0	0	2	0	3	3	3
14	1	0	0	0	0	1	1	0	0	0	0	2	0	3	3	3
15	0	0	0	0	0	1	0	0	0	0	0	1	0	2	3	3
16	0	1	1	0	1	0	0	0	2	0	0	1	0	4	3	4
17	0	1	0	0	0	0	0	0	1	0	0	1	0	2	1	2
18	2	1	1	0	1	0	1	0	2	0	1	1	2	4	3	4

their own, so no training session was carried out. Table 9.2 presents the scores that were assigned by each judge to the units in Figure 9.4.

As discussed in Section 6.3.7, the same texts were also given to two computational linguists, who were asked to build one RS-tree for each text. I took then the RS-trees built by the analysts and used the formalization in Chapter 3 to associate with each node in a tree its salient units. The salient units were computed recursively, associating with each leaf in an RS-tree the leaf itself, and to each internal node the salient units of the nucleus or nuclei of the rhetorical relation corresponding to that node. I then computed for each textual unit a score, by applying formula 9.1. Table 9.2 also presents the scores that were derived from the RS-trees that were built by each analyst for the text given in Figure 9.4 and the scores that were derived from the discourse tree that was built by the discourse-based summarizer.

Usually, the granularity of the trees that are built by the rhetorical parser is coarser than the granularity of those that are built manually. The last column in Table 9.2 reflects this phenomenon: all the units that were determined manually and that overlapped an elementary unit determined by the rhetorical parser were assigned the same score. For

Table 9.3
Percent agreement with the majority opinion

Units	Percent Agreement
All units	70.67
Very important units	65.66
Less important units	58.04
Unimportant units	73.86

example, units 1 and 3 in Figure 9.4 correspond to unit 1 in Figure 9.1. Because the score of unit 1 in the discourse structure that is built by the rhetorical parser is 3, both units 1 and 3 in Figure 9.4 are assigned the score 3.

Agreement among Judges

OVERALL AGREEMENT AMONG JUDGES. I measured the agreement of the judges with one another, using the notion of *percent agreement* that was defined by Gale [1992] and used extensively in discourse segmentation studies [Passonneau and Litman, 1993, Hearst, 1994]. Percent agreement reflects the ratio of observed agreements to possible agreements with the majority opinion. The percent agreements computed for each level of importance are given in Table 9.3. The agreements among judges for my experiment seem to follow the same pattern as those described by other researchers in summarization [Johnson, 1970]. That is, the judges are quite consistent with respect to what they perceive to be very important and unimportant, but less consistent with respect to what they perceive to be less important. In contrast with the agreement observed among judges, the percentage agreements computed for 1000 importance assignments that were randomly generated for the same texts followed a normal distribution with $\mu = 47.31, \sigma = 0.04$. These results suggest that the agreement among judges is significant.

AGREEMENT AMONG JUDGES WITH RESPECT TO THE IMPORTANCE OF EACH TEXTUAL UNIT. I considered a textual unit to be labeled consistently if a simple majority of the judges (≥ 7) assigned the same score to that unit. Overall, the judges labeled consistently 140 of the 160 textual units (87%). In contrast, a set of 1000 randomly generated importance scores showed agreement, on average, for only 50 of the 160 textual units (31%), $\sigma = 0.05$.

The judges consistently labeled 36 of the units as very important, 8 as less important, and 96 as unimportant. They were inconsistent with respect to 20 textual units. For example, for the text in Figure 9.4, the judges consistently labeled units 4 and 12 as very important, units 5 and 6 as less important, units 1, 2, 3, 7, 8, 9, 10, 11, 13, 14, 15, 17 as unimportant,

and were inconsistent in labeling unit 18. If we compute percent agreement figures only for the textual units for which at least seven judges agreed, we get 69% for the units considered very important, 63% for those considered less important, and 77% for those considered unimportant. The overall percent agreement in this case is 75%.

STATISTICAL SIGNIFICANCE. It has often been emphasized that agreement figures of the kind computed above could be misleading [Krippendorff, 1980, Passonneau and Litman, 1993]. Since the "true" set of important textual units cannot be independently known, we cannot compute how valid the importance assignments of the judges were. Moreover, although the agreement figures that would occur by chance offer a strong indication that our data are reliable, they do not provide a precise measurement of reliability.

To compute a reliability figure, I followed the same methodology as Passonneau and Litman [1993] and Hearst [1994] and applied Cochran's Q summary statistics to the data [Cochran, 1950]. Cochran's test assumes that a set of judges makes binary decisions with respect to a dataset. The null hypothesis is that the number of judges that take the same decision is randomly distributed. Since Cochran's test is appropriate only for binary judgments and since my main goal was to determine a reliability figure for the agreement among judges with respect to what they believe to be important, I evaluated two versions of the data that reflected only one importance level. In the first version I considered as being important the judgments with a score of 2 and unimportant the judgments with a score of 0 and 1. In the second version, I considered as being important the judgments with a score of 2 and 1 and unimportant the judgments with a score of 0. Essentially, I mapped the judgment matrices of each of the five texts into matrices whose elements ranged over only two values: 0 and 1. After these modifications were made, I computed for each version and each text the Cochran Q statistics, which approximates the χ^2 distribution with $N - 1$ degrees of freedom, where N is the number of elements in the dataset. In all cases I obtained probabilities that were very low: $p < 10^{-6}$. This means that the agreement among judges was extremely significant.

Although the probability was very low for both versions, it was lower for the first version of the modified data than for the second. Because of this, I considered as important only the units that were assigned a score of 2 by a majority of the judges.

As I have already mentioned, my ultimate goal was to determine whether there exists a correlation between the units that judges find important and the units that have nuclear status in the rhetorical structure trees of the same texts. Since the percentage agreement for the units that were considered very important was higher than the percentage agreement for the units that were considered less important, and since the Cochran's significance computed for the first version of the modified data was higher than the one computed for the second, I decided to consider the set of 36 textual units labeled by a majority of judges

with 2 as a reliable reference set of importance units for the five texts. For example, units 4 and 12 from the text in Figure 9.4 belong to this reference set.

Agreement between Analysts Once I determined the set of textual units that the judges believed to be important, I needed to determine the agreement between the analysts who built the discourse trees for the five texts. Because I did not know the distribution of the importance scores derived from the discourse trees, I computed the correlation between the analysts by applying Spearman's correlation coefficient on the scores associated to each textual unit. I interpreted these scores as ranks on a scale that measures the importance of the units in a text.[2]

The Spearman correlation coefficient between the ranks assigned for each textual unit on the bases of the RS-trees built by the two analysts was high for each of the five texts. It ranged from 0.645 to 0.960 at the $p < 0.0001$ level of significance. The Spearman correlation coefficient between the ranks assigned to the textual units of all five texts was 0.798, at the $p < 0.0001$ level of significance.

Agreement between the Analysts and the Judges with Respect to the Most Important Textual Units In order to determine whether there exists any correspondence between what readers believe to be important and the nuclei of the RS-trees, I selected, from each of the five texts, the set of textual units that were labeled as "very important" by a majority of the judges. For example, for the text in Figure 9.4, I selected units 4 and 12, i.e., 11% of the units. Overall, the judges selected 36 units as being very important, which is approximately 22% of the units in all the texts. The percentages of important units for the five texts were 11, 36, 35, 17, and 22 respectively.

I took the maximal scores computed for each textual unit from the RS-trees built by each analyst and selected a percentage of units that matched the percentage of important units selected by the judges. In the cases in which there were ties, I selected a percentage of units that was closest to the one computed for the judges. For example, I selected units 4 and 12, which represented the most important 11% of the units that were induced by formula 9.1 on the RS-tree built by the first analyst. However, I selected only unit 4, which represented 6% of the most important units that were induced on the RS-tree built by the second analyst, because units 10, 11, and 12 have the same score (see Table 9.2). If I had selected units

2. The Spearman correlation coefficient is based on the ranks of the data, and not on the data itself, and so is resistant to outliers. The null hypothesis tested by Spearman is that two variables are independent of each other, against the alternative hypothesis that the rank of a variable is correlated with the rank of another variable. The value of the statistic ranges from -1, indicating that high ranks of one variable occur with low ranks of the other variable, through 0, indicating no correlation between the variables, to $+1$, indicating that high ranks of one variable occur with high ranks of the other variable.

Table 9.4
The performance of a discourse-based summarizer that uses manually built trees

Granularity	Method	Recall	Precision
Clause-level	Judges	72.66	69.63
	Analysts	55.55	66.66
	Random	25.70	25.70
Sentence-level	Judges	78.11	79.37
	Analysts	67.23	78.06
	Random	38.40	38.40

10, 11, and 12 as well, I would have ended up selecting 22% of the units of the text in Figure 9.4, which is farther from 11 than 6. Hence, I determined for each text the set of important units as labeled by judges and as derived from the RS-trees of those texts.

I calculated for each text the recall and precision of the important units derived from the RS-trees, with respect to the units labeled important by the judges. The recall figure is given by the number of units identified correctly over the number of units considered important by the judges. The precision figure is given by the number of units identified correctly over the total number of units that were selected. The overall recall and precision was the same for both analysts: 55.55% recall and 66.66% precision. In contrast, the average recall and precision for the same percentages of units selected randomly 1000 times from the same five texts were both 25.7%, $\sigma = 0.059$. For a better understanding of the strengths and weaknesses of the approach, I also computed recall and precision figures for judges. These figures were computed by taking the average recall and precision of the summaries built by each judge individually when compared with the majority opinion of the other judges. The average recall and precision for the judges were 72.55% and 69.63% respectively.

In summarizing text, it is often useful to consider not only clause-like units, but full sentences. To account for this, I considered as important all the textual units that pertained to a sentence that was characterized by at least one important textual unit. For example, I labeled as important textual units 1 to 4 in the text in Figure 9.4, because they make up a full sentence and because unit 4 was labeled as important. For the adjusted data, I determined again the percentages of important units for the five texts and I recalculated the recall and precision for both analysts: the recall was 68.96% and 65.51% and the precision 81.63% and 74.50% respectively. (The average recall and precision figures are shown in Table 9.4.)

In contrast with these figures, the average recall and precision for the same percentages of units selected randomly 1000 times from the same five texts were 38.4%, $\sigma = 0.048$. And the average recall and precision for the judges were 78.11% and 79.37% respectively.

Table 9.4 summarizes these findings. These results confirm that there exists a strong correlation between the nuclei of the RS-trees that pertain to a text and what readers perceive to be important in that text. Given the values of recall and precision that I obtained, it is plausible that an adequate computational treatment of discourse theories would provide a significant part of what is needed for selecting accurately the important units in a text. However, the results also suggest that the summarization strategy developed so far is not enough if one wants to select important units at levels of performance that are indistinguishable from those of human judges.

9.2.3 An Evaluation of the Discourse-Based Summarization Program

To evaluate the summarization program, I followed the same method as in Section 9.2.2. That is, I used the importance scores assigned by formula 9.1 to the units of the discourse trees built by the rhetorical parser in order to compute statistics similar to those discussed in conjunction with the manual analyses. When the program selected only the textual units with the highest scores, in percentages that were equal to those of the judges, the recall was 51.35% and the precision was 63.33%. When the program selected the full sentences that were associated with the most important units, in percentages that were equal to those of the judges, the recall was 57.69% and the precision 51.72%.

I also compared the performance of the discourse-based summarizer with that of a commercial summarizer, the one included in the Microsoft Office97 package. I ran the Microsoft summarization program on the five texts from *Scientific American* and selected the same percentages of textual units as those considered important by the judges. When I selected percentages of text that corresponded only to the clause-like units considered important by the judges, the Microsoft program recalled 27.77% of the units, with a precision of 25.64%. When I selected percentages of text that corresponded to sentences considered important by the judges, the Microsoft program recalled 41.37% of the units, with a precision of 38.70%.

In order to provide a better understanding of the results in this section, I also considered two baseline algorithms: one baseline algorithm randomly selects from a text a number of units that matches the number of units that were considered important in that text by the human judges; the other baseline algorithm selects the first k units in a text, where k matches the number of units considered important by human judges.

Table 9.5 shows recall and precision results for the two baselines, Microsoft Office97, and discourse-based summarizers, as well as the results that would have been obtained if we had applied the score function 9.1 on the discourse trees that were built manually. In addition to the recall and precision figures, Table 9.5 also displays the corresponding F-values, computed using formula 9.3, which is given below. The F-value is always a number

Table 9.5
The performance of a discourse-based summarizer that uses the cue-phrase-based rhetorical parser

Granularity	Method	Recall	Precision	F-value
Clause-level	Judges	72.66	69.63	71.11
	Analyst-based summarizer	55.55	66.66	60.55
	Discourse-based summarizer	51.35	63.33	56.71
	Microsoft summarizer	27.77	25.44	26.55
	Lead baseline	39.68	39.68	39.68
	Random baseline	25.70	25.70	25.70
Sentence-level	Judges	78.11	79.37	78.73
	Analyst-based summarizer	67.23	78.06	72.24
	Discourse-based summarizer	57.69	51.72	54.54
	Microsoft summarizer	41.37	38.70	39.99
	Lead baseline	54.22	54.22	54.22
	Random baseline	38.40	38.40	38.40

between the values of recall and precision, and is higher when recall and precision are closer.

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9.3}$$

In Table 9.5, I use the term "Analyst-based Summarizer" as a name for a summarizer that identifies important units on the basis of discourse trees that are manually built. The recall and precision figures associated with the baseline algorithm that selects textual units randomly represent averages of 1000 runs.

The results in Table 9.5 show that at the clause level, the discourse-based summarizer performs quite well when compared with the analyst-based summarizer. Hence, although the trees built by the underlying rhetorical parser are not perfect (see Section 6.3.7), the nuclearity assignments seem to be good enough for the purpose of summarization. In contrast, at the sentence level, the performance of the discourse-based summarizer is much lower than that of the analyst-based summarizer. This suggests that assigning the same importance score to all clauses in a sentence is not a very good strategy.

Yet, overall, the discourse-based summarizer outperforms both baselines and the Microsoft summarizer. Surprisingly, the Microsoft summarizer performs worse than the lead baseline and slightly better than the random baseline.

10 Improving Summarization Performance through Rhetorical Parsing Tuning

10.1 Motivation

Using the discourse structure of text as an indicator of textual importance is only one of the many methods to obtain indicators that have been proposed in the literature. Some approaches assume that important textual units contain words that are used frequently [Luhn, 1958, Edmundson, 1968] or words that are used in the title and section headings [Edmundson, 1968]. Other approaches assume that important sentences are located at the beginning or the end of paragraphs [Baxendale, 1958] or at positions that can be determined through training for each particular text genre [Lin and Hovy, 1997]. (In fact, in the corpus of five *Scientific American* texts, we have already seen that the lead baseline algorithm has quite a good performance.) Other systems assume that important sentences in texts contain "bonus" words and phrases, such as *significant, important, in conclusion* and *In this paper we show,* while unimportant sentences contain "stigma" words such as *hardly* and *impossible* [Edmundson, 1968, Kupiec et al., 1995, Teufel and Moens, 1997]. And, yet, other systems assume that important sentences and concepts are the highest connected entities in more or less elaborate semantic structures [Skorochodko, 1971, Hoey, 1991, Salton and Allan, 1995, Mani and Bloedorn, 1998, Barzilay and Elhadad, 1997].

A variety of systems [Edmundson, 1968, Kupiec et al., 1995, Teufel and Moens, 1997, Lin, 1998, Mani and Bloedorn, 1998] were designed to integrate subsets of the heuristics mentioned above. In these approaches, each individual heuristic yields a probability distribution that reflects the importance of sentences. A combination of the probability distributions defined by each heuristic yields the sentences that are most likely to be included in a summary.

What all these multiple heuristic-based systems have in common is that they treat texts as *flat sequences of sentences*—no such system employs discourse-based heuristics. As a consequence, it is possible, for example, for a sentence to be assigned a high importance score on the basis of its position in the text and its semantic similarity with the title, although it is subsidiary to the main argument made in the text. In this chapter, I try to remedy this shortcoming, by taking advantage of the structure of discourse.

More precisely, I study the relationship between the structure of discourse and a set of summarization heuristics that are employed by current systems. A tight coupling of the two, which is achieved by applying a simple learning mechanism, gives two advantages over previous methods. First, two corpora of manually built summaries enable one to learn genre-specific combinations of heuristics that can be used for disambiguation during discourse parsing. Second, the discourse structures that the rhetorical parser derives enable one to select textual units that are not only important according to a variety of position-, title-, and lexically-based heuristics, but also central to the main claims of texts. By improving the disambiguation mechanisms of the cue-phrase-based discourse parser, I achieve a

significant increase in accuracy over the discourse-based summarization system described in the previous section.

10.2 An Enhanced Discourse-Based Framework for Text Summarization

10.2.1 Introduction

There are two ways in which one can integrate a discourse-based measure of textual saliency, such as that described above, with measures of saliency that are based on cohesion, position, similarity with the title, etc. The simplest way is to compute a probability distribution of the importance of textual units according to the discourse method and to combine it with all probability distributions produced by the other heuristics. In such an approach, the discourse heuristic is just one of the n heuristics that are employed by a system. Obtaining good summaries amounts then to determining a good way of combining the implemented heuristics.

Overall, a summarization system that works along the lines described above still treats texts as flat sequences of textual units, although the discourse method internally uses a more sophisticated representation. The shortcoming of such an approach is that it still permits the selection of textual units that do not play a central role in discourse. For example, if the text to be summarized consists only of units 7 and 8 in Figure 9.2, it may be possible that the combination of the position, title, and discourse heuristics will yield a higher score for unit 7 than for unit 8, although unit 8 is the nucleus of the text and expresses what is important. Unfortunately, if we interpret the text as a flat sequence of units, the rhetorical relation and the nuclearity assignments with respect to these units cannot be appropriately exploited.

A more complex way to integrate discourse, cohesion, position, and other summarization-based methods is to consider that the structure of discourse is the most important factor in determining saliency, an assumption supported by experiments done by Mani et al. [1998] and consistent with the experiments described in Section 9.2. In such an approach, we no longer interpret texts as flat sequences of textual units, but as tree structures that reflect the nuclearity and rhetorical relations that characterize each textual span. When discourse is taken to be central to the interpretation of text, obtaining good summaries amounts to finding the "best" discourse interpretations. In the rest of this section, I explore this approach.

10.2.2 Criteria for Measuring the 'Goodness"of Discourse Structures

In order to find the "best" discourse interpretations, i.e., the interpretations that yield summaries that are most similar to summaries generated manually, I considered seven metrics, which I discuss below.

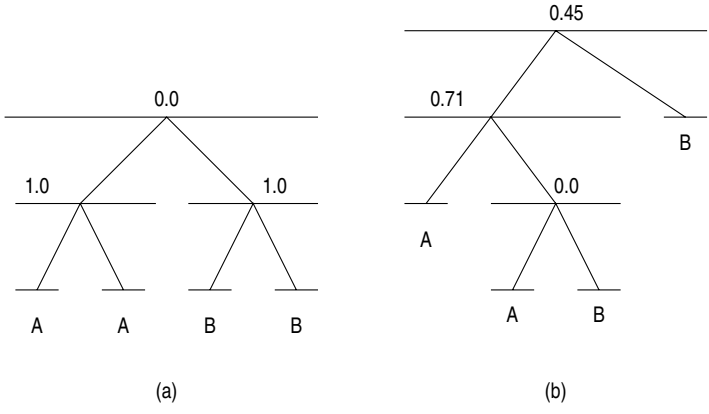

Figure 10.1
Two discourse structures of a hypothetical text "A. A. B. B.". The numbers associated with the internal nodes are clustering scores.

THE CLUSTERING-BASED METRIC. A common assumption of current text theories [Hoey, 1991, Hearst, 1997] is that well-written texts exhibit a well-defined topical structure. In the approach presented here, I assume that a discourse tree is "better" if it exhibits a high-level structure that matches as much as possible the topical boundaries of the text for which that structure is built.

In order to implement this intuition, I associate with each node of a tree a clustering score. For the leaves, this score is 0; for the internal nodes, the score is given by the similarity between the immediate children. The similarity is computed using the cosine metric discussed in Section 7.3.3 (Semantic-similarity-based features). I consider that a discourse tree A is "better" than another discourse tree B if the sum of the clustering scores associated with the nodes of A is higher than the sum of the clustering scores associated with the nodes of B.

Consider, for example, that a text has four sentences; the first two sentences contain only the word "A" and the last two sentences only the word "B". According to the clustering-based metric, the discourse structure in Figure 10.1a, which has a clustering score of $2.0(= 1.0 + 1.0 + 0.0)$, is better than the discourse structure in Figure 10.1b, which has a clustering score of $1.16(= 0.71 + 0.0 + 0.45)$. This is consistent with the intuitive notion that the discourse structure of such a text should be the tree in Figure 10.1a, because that text is made of two segments: the first segment, which contains the first two sentences, "talks about" "A"; while the second segment "talks about" "B".

THE MARKER-BASED METRIC. Naturally occurring texts use a wide range of discourse markers, which signal coherence relations between textual spans of various sizes. I assume that a discourse structure should reflect explicitly as many of the discourse relations that are signaled by discourse markers as possible. In other words, I assume that a discourse structure *A* is better than a discourse structure *B* if *A* uses more rhetorical relations that are explicitly signaled than *B*.

THE RHETORICAL-CLUSTERING-BASED METRIC. The clustering-based metric discussed above computes an overall similarity between two textual spans. However, in the discourse formalization proposed in Chapter 3, it is assumed that whenever a discourse relation holds between two textual spans, that relation also holds between the salient units (nuclei) associated with those spans. I extend this observation to similarity as well, by introducing the rhetorical-clustering-based metric, which measures the similarity between the salient units associated with two spans. For example, the clustering-based score associated with the root of the tree in Figure 9.2 measures the similarity between spans [1,6] and [7,10]. In contrast, the rhetorical-clustering-based score associated with the root of the same tree measures the similarity between units 2 and 8, which are the salient units of spans [1,6] and [7,10] respectively. In the light of the rhetorical-clustering-based metric, I consider that a discourse tree *A* is "better" than another discourse tree *B* if the sum of the rhetorical-clustering scores associated with the nodes of *A* is higher than the sum of the rhetorical-clustering scores associated with the nodes of *B*.

THE SHAPE-BASED METRIC. The only disambiguation metric that I used in Chapter 6 was the shape-based metric, according to which the "best" trees are those that are skewed to the right. According to the shape-based metric, I consider that a discourse tree *A* is "better" than another discourse tree *B* if *A* is more skewed to the right than *B* (see Section 6.3.5 for a mathematical formulation of the notion of skewedness).

THE TITLE-BASED METRIC. A variety of systems [Edmundson, 1968, Kupiec et al., 1995, Hovy and Lin, 1997, Teufel and Moens, 1997] assume that important sentences in a text use words that occur in the title. I measure the similarity between each textual unit and the title by applying the cosine metric discussed in Section 7.3.3 (Semantic-similarity-based features). I compute a title-based score for each discourse structure by computing the similarity between the title and the units that are promoted as salient in that structure. The intuition that I capture in this way is that a discourse structure should be constructed so that it promotes as close to the root as possible the units that are similar to the title. According to the title-based metric, I consider that a discourse structure *A* is "better" than a discourse structure *B* if the title-based score of *A* is higher than the title-based score of *B*.

THE POSITION-BASED METRIC. Research in summarization [Baxendale, 1958, Edmundson, 1968, Kupiec et al., 1995, Lin and Hovy, 1997] has shown that, in genres with stereotypical structure, important sentences are often located at the beginning or the end of paragraphs/documents. The position-based metric is consistent with this research; it assigns a positive score to each textual unit that belongs to the first two sentences or last two sentences of the first three or last two paragraphs. I compute a position-based score for each discourse structure by averaging the position-based scores of the units that are promoted as salient in that discourse structure. The intuitive notion this metric aims to formalize is that a discourse structure should be constructed so that it promotes as close to the root as possible the units that are located at the beginning or the end of a text. According to the position-based metric, I consider that a discourse structure A is "better" than a discourse structure B if the position-based score of A is higher than the position-based score of B.

THE CONNECTEDNESS-BASED METRIC. A heuristic that is often employed by current summarization systems is that of considering important the highest connected entities in more or less elaborate semantic structures [Skorochodko, 1971, Hoey, 1991, Salton and Allan, 1995, Mani and Bloedorn, 1998, Barzilay and Elhadad, 1997]. I implement this heuristic by computing the average cosine similarity of each textual unit in a text with respect to all the other units. I associate a connectedness-based score to each discourse structure by averaging the connectedness-based scores of the units that are promoted as salient in that discourse structure. As in the case of the other metrics, I consider that a discourse structure A is "better" than a discourse structure B if the connectedness-based score of A is higher than the connectedness-based score of B.

10.3 Combining Heuristics

10.3.1 The Approach

Discourse parsing is ambiguous in the same way sentence parsing is: the rhetorical parsing algorithm often derives more than one discourse structure for a given text. Each of the seven metrics listed above favors a different discourse interpretation: for example, the marker-based metric assumes that discourse trees that reflect explicitly marked relations are better than discourse trees that do not; in contrast, the title-based metric assumes that the best discourse trees are those that promote as close to the root as possible the textual units that are similar with the title.

In this section, I assume that the "best" discourse structures are given by a linear combination of the seven metrics. Hence, along the lines described in Section 10.2.2, I associate with each discourse structure a clustering-based score s_{clust}, a marker-based score s_{mark}, a

rhetorical-clustering-based score s_{rhet_clust}, a shape-based score s_{shape}, a title-based score s_{title}, a position-based score s_{pos}, and a connectedness-based score s_{con}; and I assume that the best tree of a text is that that corresponds to the discourse structure D that has the highest score $s(D)$. The score $s(D)$ is computed as shown in 10.1, where $w_{clust}, \ldots, w_{con}$ are weights associated with each metric.

$$\begin{aligned} s(D) = {} & w_{clust} \times s_{clust}(D) + w_{mark} \times s_{mark}(D) \\ & + w_{rhet\_clust} \times s_{rhet\_clust}(D) + w_{shape} \times s_{shape}(D) \\ & + w_{title} \times s_{title}(D) + w_{pos} \times s_{pos}(D) + w_{con} \times s_{con}(D). \end{aligned} \tag{10.1}$$

To avoid data skewedness, the scores that correspond to each metric are normalized to values between 0 and 1.

Given the above formulation, my goal is to determine combinations of weights that yield discourse structures that, in turn, yield summaries that are as close as possible to those generated by humans. In discourse terms, this amounts to using empirical summarization data for discourse parsing disambiguation.

10.3.2 Corpora Used in the Study

In order to evaluate the appropriateness for summarization of each of the heuristics, I have used two corpora: a corpus of forty newspaper articles from the TREC collection [Jing et al., 1998] and the corpus of five articles from *Scientific American* that I used to evaluate the performance of the cue-phrase-based rhetorical parser.

Five human judges selected sentences to be included in 10% and 20% summaries of each of the articles in the TREC corpus (see [Jing et al., 1998] for details). For each of the forty articles and for each cutoff figure (10% and 20%), I took the set of sentences selected by at least three human judges as the "gold standard" for summarization. For each of the five texts in the *Scientific American* corpus, I took the set of textual units for which at least seven judges agreed to be very important as the gold standard for summarization.

I built automatically discourse structures for the texts in the two corpora using various combinations of weights and I compared the summaries that were derived from these structures with the gold standards. The comparison employed traditional recall and precision figures. For both corpora, I attempted to mimic as closely as possible the summarization tasks carried out by human judges. For the TREC corpus, I automatically extracted summaries at 10% and 20% cutoffs; for the *Scientific American* corpus, I automatically extracted summaries that reflected the lengths of the summaries on which human judges agreed.

Table 10.1
The appropriateness of each of the seven metrics for text summarization in the TREC corpus—the 10% cutoff

Metric	Recall	Precision	F-value
Humans	83.20	75.95	79.41
Clustering	48.08	54.29	51.00
Marker	38.63	44.44	41.33
Rhetorical clustering	26.26	27.87	27.04
Shape	44.04	52.52	47.91
Title	58.93	67.67	63.00
Position	52.87	63.73	57.79
Connectedness	35.35	31.31	33.21
Lead	82.91	63.45	71.89
Random	9.44	9.44	9.44

10.3.3 Appropriateness for Summarization of the Individual Heuristics

THE TREC CORPUS. Initially, I evaluated the appropriateness for text summarization of each of the seven heuristics at both 10% and 20% cutoffs for the collection of texts in the TREC corpus. By assigning in turn value 1 to each of the seven weights, while the other six weights were assigned value 0, I estimated the appropriateness of using each individual metric for text summarization.

Tables 10.1 and 10.2 show the recall, precision, and F-value figures that pertain to discourse structures that were built for the TREC corpus, in order to evaluate the appropriateness for text summarization of each of the seven metrics at 10% and 20% cutoffs, respectively. For a better understanding of the impact of each heuristic, Tables 10.1 and 10.2 also show the recall, precision, and F-value figures associated with the human judges and with two baseline algorithms. The recall and precision figures for the human judges were computed by taking the average recall and precision of the summaries built by each human judge individually when compared with the gold standard. These recall and precision figures can be interpreted as summarization upper bounds for the collection of texts that they characterize. The recall and precision figures that pertain to the baseline algorithms are computed as follows: the lead-based algorithm assumes that important units are located at the beginning of texts; the random-based algorithm assumes that important units can be selected randomly.

The results in Table 10.1 show that, for newspaper articles, the title- and position-based metrics are the best individual metrics for distinguishing between discourse trees that are appropriate for generating 10% summaries and discourse trees that are not. The

Table 10.2
The appropriateness of each of the seven metrics for text summarization in the TREC corpus—the 20% cutoff

Metric	Recall	Precision	F-value
Humans	82.83	64.93	72.80
Clustering	40.99	43.61	42.26
Marker	37.91	38.78	38.34
Rhetorical clustering	23.10	24.68	23.86
Shape	46.54	49.73	48.08
Title	42.29	40.17	41.20
Position	37.48	40.97	39.14
Connectedness	29.87	32.78	31.26
Lead	70.91	46.96	56.50
Random	15.80	15.80	15.80

clustering- and shape-based metrics are less appropriate and the marker-, connectedness-, and rhetorical-clustering-based heuristics are the least appropriate. Interestingly, none of these heuristics taken in isolation is better than the lead-based algorithm. In fact, the results in Table 10.1 show that there is little difference between summaries generated by the lead-based algorithm and summaries generated by humans.

I was so puzzled by this finding that I investigated further this issue: by scanning the collection of forty articles, I came to believe that since most of them are very short and simple, they are inappropriate as a testbed for summarization research. To estimate the validity of this belief, I focused my attention on a subset of ten articles that seemed to use a more elaborate writing style, which did not follow straightforwardly the pyramid-based approach. When I evaluated the performance of the lead-based algorithm on this subset, I obtained figures of 66.00% recall and 43.66% precision at the 10% cutoff. This result suggests that as soon as more sophisticated texts are considered, the performance of the lead-based algorithm decreases significantly even within the newspaper genre.

The results in Table 10.2 show that, for newspaper articles, the shape-based metric is the best individual metric for distinguishing between discourse trees that are appropriate for 20% summaries and discourse trees that are not. Still, the shape-based heuristic is not better than the lead-based algorithm. According to the data in Table 10.2, the clustering-, marker-, title-, and position-based metrics are less successful in distinguishing between discourse structures that are appropriate for generating 20% summaries than the shape-based metric, and the rhetorical-clustering- and connectedness-based metrics are the least successful.

Table 10.3
The appropriateness of each of the seven metrics for text summarization in the *Scientific American* corpus—the clause-like unit case

Metric	Recall	Precision	F-value
Humans	72.66	69.63	71.11
Clustering	54.05	66.66	59.70
Marker	43.24	55.17	48.48
Rhetorical clustering	48.65	62.07	54.55
Shape	51.35	63.33	56.71
Title	40.54	55.56	46.88
Position	29.73	47.83	36.67
Connectedness	24.32	40.91	30.51
Lead	39.68	39.68	39.68
Random	25.70	25.70	25.70

Table 10.4
The appropriateness of each of the seven metrics for text summarization in the *Scientific American* corpus—the sentence case

Metric	Recall	Precision	F-value
Humans	78.11	79.37	78.73
Clustering	42.31	42.31	42.31
Marker	42.31	42.31	42.31
Rhetorical clustering	46.15	40.00	42.86
Shape	57.69	51.72	54.54
Title	30.77	33.33	32.00
Position	30.77	38.10	34.04
Connectedness	23.08	25.00	24.00
Lead	54.22	54.22	54.22
Random	38.40	38.40	38.40

THE SCIENTIFIC AMERICAN CORPUS. When I evaluated the appropriateness for text summarization of the heuristics at both clause and sentence levels for the collection of texts in the *Scientific American* corpus, I obtained a totally different distribution of the configuration of weights that yielded the highest recall and precision figures.

A close analysis of the results in Table 10.3 shows that, for *Scientific American* articles, the clustering-, rhetorical-clustering-, and shape-based metrics are the best individual metrics for distinguishing between discourse trees that are good for clause-based summarization and discourse trees that are not.

The results in Table 10.4 show that, for *Scientific American* articles, the shape-based metric is the best individual metric for distinguishing between discourse trees that are appropriate for sentence-based summarization. Surprisingly, the title-, position-, and connectedness-based metrics underperform even the random-based metric.

In contrast with the results for the TREC corpus, the lead-based algorithm performs significantly worse than human judges for the texts in the *Scientific American* corpus, despite the *Scientific American* texts being shorter than those in the TREC collection.

DISCUSSION. Overall, the recall and precision figures presented in this section suggest that no individual heuristic consistently guarantees success across different text genres. Moreover, the figures suggest that, even within the same genre, the granularity of the textual units that are selected for summarization and the overall length of the summary both affect the appropriateness of a given heuristic.

By focusing only on the human judgments, I noticed that the newspaper genre yields a higher consistency than the *Scientific American* genre with respect to what humans believe to be important. Also, the results in this section show that humans agree better on important sentences than on important clauses; and that within the newspaper genre, they agree better on what is very important (the 10% summaries) than on what is somewhat important (the 20% summaries).

10.4 Learning the Best Combinations of Heuristics

The individual applications of the metrics suggest what heuristics are appropriate for summarizing texts that belong to the text genres of the two corpora. In addition to this assessment, I was also interested in finding *combinations* of heuristics that yield good summaries. To this end, I employed a simple learning paradigm, which I describe below.

10.4.1 A GSAT-like Algorithm

In the framework that I proposed in this section, finding a combination of metrics that is best for summarization amounts to finding a combination of weights $w_{clust}, \ldots, w_{con}$ that maximizes the recall and precision figures associated with automatically built summaries. The *rhetorical parsing tuning* algorithm shown in Figure 10.2 performs a greedy search in the seven-dimensional space defined by the weights, using an approach that mirrors that proposed by Selman, Levesque, and Mitchell [1992] for solving propositional satisfiability problems.

The algorithm assigns initially to each member of the vector of weights $\vec{W}_{max}$ a random value in the interval [0, 1]. This assignment corresponds to a point in the n-dimensional space defined by the weights. The program then attempts *NoOfSteps* times to move incre-

Input: A corpus of texts C_T.
The manually built summaries S_T for the texts in C_T.
NoOfTries, NoOfSteps, Δw.

Output: The weights $\vec{W}_{max} = \{w_{clust}, w_{mark}, \ldots, w_{con}\}$ that yield the best summaries with respect to C_T and S_T.

1. $\vec{W}_{max} = \{w_{clust}, w_{mark}, \ldots, w_{con}\} = \{rand(0, 1), rand(0, 1), \ldots, rand(0, 1)\}$
2. **for** tries = 1 **to** *NoOfTries*
3. $\quad \vec{W}_t = \{w_{clust}, w_{mark}, \ldots, w_{con}\} = \{rand(0, 1), rand(0, 1), \ldots, rand(0, 1)\}$
4. $\quad F_t = \text{F\_RecallAndPrecision}(w_{clust}, w_{mark}, \ldots, w_{con})$;
5. $\quad$ **for** flips = 1 **to** *NoOfSteps*
6. $\quad\quad F_1 = \text{F\_RecallAndPrecision}(w_{clust} + \Delta w, w_{mark}, \ldots, w_{con})$;
7. $\quad\quad F_2 = \text{F\_RecallAndPrecision}(w_{clust} - \Delta w, w_{mark}, \ldots, w_{con})$;
8. $\quad\quad \ldots$
9. $\quad\quad F_{13} = \text{F\_RecallAndPrecision}(w_{clust}, w_{mark}, \ldots, w_{con} + \Delta w)$;
10. $\quad\quad F_{14} = \text{F\_RecallAndPrecision}(w_{clust}, w_{mark}, \ldots, w_{con} - \Delta w)$;
11. $\quad\quad F_{max} = max(F_t, F_1, F_2, \ldots, F_{14})$;
12. $\quad\quad F_t = \text{randomOf}(F_{max})$;
13. $\quad\quad \vec{W}_t = \text{weightsOf}(F_t)$;
14. $\quad$ **endfor**
15. $\quad \vec{W}_{max} = max(\vec{W}_{max}, \vec{W}_t)$;
16. **endfor**
17. **return** $\vec{W}_{max}$

Figure 10.2
A GSAT-like algorithm for improving summarization

mentally, in the n-dimensional space, along a direction that maximizes the F-measure of the recall and precision figures that pertain to the automatically built summaries.

For each point $\vec{W}_t$ the program computes the F-value of the recall and precision figures of the summaries that correspond to all the points in the neighborhood of $\vec{W}_t$ that are at distance Δw along each of the seven axes (lines 6–10 in Figure 10.2). From the set of fourteen points that characterize the neighborhood of the current configuration $\vec{W}_t$, the algorithm selects randomly (line 12) one of the weight configurations that yielded the maximum F-value (line 11). In line 13, the algorithm moves in the n-dimensional space to the position that characterizes the configuration of weights that was selected on line 12. After *NoOfSteps* iterations, the algorithm updates the configuration of weights $\vec{W}_{max}$ such that it reflects the combination of weights that yielded the maximal F-value of the recall

Table 10.5
The combination of heuristics that yielded the best summaries for the texts in the TREC corpus—10% compression

Method	w_{clust}	w_{mark}	w_{rhet_clust}	w_{shape}	w_{title}	w_{pos}	w_{con}	Training Corpus			Test Corpus		
								Rec	Prec	F-val	Rec	Prec	F-val
Humans								83.20	75.95	79.41	83.20	75.95	79.41
Program	0.65	0.62	−0.42	2.87	0.88	1.57	−0.58	71.33	81.94	76.27	63.33	75.00	68.67
	0.58	0.46	−0.36	1.19	0.54	1.11	−0.56	68.00	83.33	74.89	63.33	65.83	64.56
	0.57	0.90	0.52	2.79	0.80	2.13	−0.74	69.16	76.66	72.72	65.83	90.00	76.04
	0.25	−0.01	−0.32	1.62	0.46	1.79	0.14	69.72	83.88	76.15	62.50	59.16	60.78
Average								69.55	81.33	74.98	63.75	72.50	67.84
Lead								82.91	63.45	71.89	82.91	63.45	71.89

Table 10.6
The combination of heuristics that yielded the best summaries for the texts in the TREC corpus—20% compression

Method	w_{clust}	w_{mark}	w_{rhet_clust}	w_{shape}	w_{title}	w_{pos}	w_{con}	Training Corpus			Test Corpus		
								Rec	Prec	F-val	Rec	Prec	F-val
Humans								82.83	64.93	72.80	82.83	64.93	72.80
Program	−1.46	0.42	0.13	1.58	0.48	1.79	0.06	65.58	67.81	66.68	63.76	59.85	61.74
	−1.17	0.18	0.10	1.61	0.59	1.25	0.14	69.99	66.30	68.10	55.04	62.88	58.70
	0.07	0.40	−0.21	1.75	0.70	1.00	−1.17	65.84	62.80	64.28	69.00	70.66	69.82
	−0.94	0.62	0.37	1.27	0.88	1.57	−0.98	65.29	67.08	66.17	59.35	49.93	54.23
Average								66.67	66.00	66.33	61.79	60.83	61.31
Lead								70.91	46.96	56.50	70.91	46.96	56.50

and precision figures (line 15 in Figure 10.2). The algorithm repeats this process *noOfTries* times, in order to increase the chance of finding a maximum that is not local.

Since the length of the summaries that I automatically extracted was fixed in all cases, I chose to look for configurations of weights that maximized the F-value of the recall and precision figures. However, one can use the algorithm in Figure 10.2 to find configurations of weights that maximize only the recall or only the precision figure as well.

10.4.2 Results

THE TREC CORPUS. I have experimented with different values for *noOfTries, noOfSteps,* and Δw. To evaluate the performance of the approach, I performed a fourfold cross-

validation experiment. That is, I divided the collection of TREC documents into four groups, each group having ten documents. I used the algorithm in Figure 10.2 to determine the weights that would maximize the summarization performance on thirty documents and then tested the summarization program with the adjusted weights on the other ten documents. Tables 10.5 and 10.6 show the configuration of weights that yielded the maximal F-values on the training corpora, as well as the recall and precision figures obtained on the test corpora at both 10% and 20% cutoffs, for $noOfTries = 50$, $noOfSteps = 60$, and $\Delta w = 0.4$. The recall, precision, and F-values that pertain to the humans and the lead baselines are computed over all documents in the corpus.

The results in Table 10.5 show that although the combination of heuristics improves the performance of the discourse-based summarizer significantly, for short newspaper articles, it is difficult to beat the lead baseline. In most of the cases, selecting the first sentence of a short news story is the best choice. However, when longer summaries are involved, the discourse-based approach to summarization pays off. At 20% compression rate, the discourse-based summarizer outperforms the lead baseline and comes closer to human performance levels.

The Scientific American corpus. I also ran the algorithm shown in Figure 10.2 on the collection of five texts in the *Scientific American* corpus, with $noOfTries = 120$, $noOfSteps = 50$, and $\Delta w = 0.4$. This collection was too small to carry out the cross-validation evaluation that I carried out on the TREC corpus. Table 10.7 shows two configurations of weights that yielded maximal F-values of the recall and precision figures at the clause-like unit level and two configurations of weights that yielded maximal F-values at the sentence level. The best combination of weights for summarization at the clause-like unit level recalls 67.57% of the elementary units considered important by human judges, with a precision of 73.53% (on the training data). The F-value of the recall and precision figures for this configuration is 70.42%, which is less than 1% lower than the F-value that pertains to human judges and about 30% higher than the F-value that pertains to the lead-based algorithm. This result outperforms significantly the previous 51.35% recall and 63.33% precision figures that were obtained using only the shape-based heuristic (see Table 10.3).

The best combination of weights for summarization at the sentence level recalls 69.23% of the sentences considered important by human judges, with a precision of 64.29%. The F-value of the recall and precision figures for this configuration is 66.67%, which is about 12% lower than the F-value that pertains to human judges but about 12% higher than the F-value that pertains to the lead-based algorithm. These results suggest that although *Scientific American* articles cannot be summarized properly by applying a simple, lead-based heuristic, they can be by applying the discourse-based algorithm.

Table 10.7
The combination of heuristics that yielded the best summaries for the texts in the *Scientific American* corpus.

								Training Corpus		
Granularity	Method	w_{clust}	w_{mark}	w_{rhet_clust}	w_{shape}	w_{title}	w_{pos}	w_{con}	Rec	Prec	F-val
Clause-like	Humans								72.66	69.93	71.27
unit	Program	0.37	0.04	1.27	0.58	0.54	−1.16	−1.24	67.57	73.53	70.42
		1.34	0.27	0.69	0.58	−0.08	2.07	−0.53	62.16	71.87	66.66
	Lead								39.68	39.68	39.68
Sentence	Humans								78.11	79.37	78.73
	Program	0.41	0.36	0.14	1.75	0.65	−0.46	−0.73	69.23	64.29	66.67
		−1.51	0.13	0.65	0.84	0.52	−1.23	1.06	65.38	68.00	66.66
	Lead								54.22	54.22	54.22

DISCUSSION. Although the *Scientific American* corpus was too small to carry out a cross-validation evaluation procedure, the experiments with the two corpora suggest that choosing a discourse interpretation from a set of valid discourse interpretations affects the performance of a summarization system that applies discourse-based techniques. The experiments also suggest that the use of multiple heuristics is beneficial for increasing summarization performance.

The analysis of the patterns of weights in Tables 10.5, 10.6, and 10.7 shows that, for both corpora, no individual heuristic is a clear winner with respect to its contribution to obtaining good summaries. For the TREC corpus, with the exception of the rhetorical-clustering- and the connectedness-based heuristics at the 10% compression rate and the rhetorical-clustering-, clustering-, and connectedness-based heuristics at the 20% compression rate, all other heuristics seem to contribute consistently to the improvement in summarization quality. For the *Scientific American* corpus, when combined with other heuristics, the marker-, rhetorical-clustering-, shape-, and title-based heuristics seem to contribute consistently to the improvement in recall and precision figures in almost all cases. In contrast, the position-, and connectedness-based heuristics seem to be even detrimental with respect to the collection of texts that I considered.

Nevertheless, the conclusion that seems to be supported by the data in Tables 10.5, 10.6, and 10.7 is that the strength of a summarization system does not depend so much on the use of one heuristic, but rather on the ability of the system to use an optimal combination of heuristics. The data also shows that the optimal combinations need not necessarily follow a common pattern: for example, the combinations of heuristics that yield the highest F-values of the recall and precision figures for the 10% cutoff in the TREC corpus differ quite significantly.

In addition to the analysis of the patterns of weights that yielded optimal summaries in the two corpora, I also examined the appropriateness of using combinations of weights that were optimal for a given summary cutoff in order to summarize texts at a different cutoff. When the patterns of weights that yielded optimal summaries at 10% cutoff were used to summarize the texts in the TREC corpus at 20% cutoff, the overall F-value was about 7% lower than the average F-value computed for the combinations of weights that yielded the best 20% summaries. Similarly, when the patterns of weights that yielded optimal summaries at 20% cutoff were used to summarize the texts in the TREC corpus at 10% cutoff, the overall F-value was about 5% lower than the average F-value computed for the combinations of weights that yielded the best 10% summaries. These results suggest that there are at least two ways in which one can train a summarization system. If the system is going to be used frequently to summarize texts at a given cutoff, then it makes sense to train it to produce good summaries at that cutoff. However, if the system is going to be used to generate summaries of various lengths, then a different training methodology should be adopted, one that would ensure optimality across the whole cutoff spectrum, from 1 to 99%.

11 Discussion

11.1 Related Work

11.1.1 Natural Language Summarization ~~a~~ Psycholinguistic Perspective

The empirical experiment described in Chapter 9 confirms the hypothesis that the units that are promoted as important by our text theory correlate with the units considered important by human judges. The validation of this hypothesis is consistent with results obtained in other psycholinguistic studies, which have repeatedly confirmed that the structure of text is essential in summarizing text. For example, Cook and Mayer [1988] have shown that teaching students how to discriminate and use the structure of text helped them improve the recall of high-level information and answer application questions. Donlan [1980] has shown that the idea of subordination and text structure is important when teaching how to locate main ideas in history textbooks. An experiment described by Palmere et al. [1983] has demonstrated that a major idea that is supported by several subordinate propositions is better recalled than if it is supported by fewer propositions. And an experiment described by Lorch and Lorch [1985] has shown that readers use a representation of topic that help them recall the main ideas in a text. When the topic is explicitly represented and is found at the beginning of texts, the recall is better than when the topic is represented implicitly or when it is found at the end of a text.

Psychological experiments have confirmed not only the role of structure in summarization, but also the role of signaling. An experiment of Loman and Mayer [1983] has shown that signaling in text increases the recall of conceptual information and helps humans generate high-quality problem solutions. The signaling techniques studied by Loman and Mayer include (i) the specification of the structure of relations by means of cue phrases and discourse markers; (ii) the premature presentation of forthcoming material; (iii) the use of summary statements; and (iv) the use of pointer or bonus words, such as "more importantly," "unfortunately," etc. In fact, Glover et al. [1988] have shown that signaling even across chapters through "preview" and "recall" sentences has a strong effect on readers' recall of prose.

11.1.2 Natural Language Summarization ~~a~~ Computational Perspective

Two discourse theories have been used so far as basis for research in summarization: those of Sidner [1983] and Mann and Thompson [1988]. In an exploratory study, Gladwin, Pulman and Sparck Jones [1991] have applied manually Sidner's focusing algorithm [1983] in order to determine the entities that are salient in discourse. Their hypothesis was that the salient entities in a text are those that are in focus the largest number of times. Their

initial, informal evaluation suggested that there may exist a correlation between the entities in focus and the entities that are salient in a text, but this line of research has not been investigated further.

In contrast, the adequacy of using Mann and Thompson's theory in text summarization has been investigated more thoroughly. The idea that the nuclei of a discourse tree correlate with what readers label as important has been long hypothesized [Mann and Thompson, 1988, Matthiessen and Thompson, 1988, Sparck Jones, 1993b]. And more recently, Rino and Scott [1996] have discussed the role that not only nuclearity but also intentions and coherence can have in going from discourse structures to text summaries. The first discourse-based summarizer was built for Japanese by Ono et al. [1994], using the discourse parser of Sumita et al. [1992]. Since the discourse trees built by Sumita et al. [1992] do not have salient units associated with the nodes, an importance score is assigned to each sentence in a tree on the basis of the depth where it occurs. An evaluation performed on editorial and technical articles showed coverage figures of key sentences and most important key sentences in the range of 41% and 60% for the editorial articles and 51% and 74% for the technical papers, respectively. In a follow-up experiment, Miike et al. [1994] showed that when the abstracts generated by Ono et al. were presented to users in a standard, information retrieval selection task, the time required was about 80% of the time required to perform the same task using the original documents, with recall and precision remaining approximately the same.

11.2 Open Problems

11.2.1 Selecting the Most Important Units in a Text

The results presented in Chapters 9 and 10 confirm the suitability of using discourse structures for text summarization. The results also indicate that our discourse-based summarizer significantly outperforms the Microsoft Office97 summarizer, which, like the vast majority of summarizers on the market, relies primarily on the assumption that important sentences contain the words that are used most frequently in a given text.

In spite of the good results, in some cases, the recall and precision figures obtained with the discourse-based summarizer are still below the performance level of humans. I believe that there are two possible explanations for this: either the rhetorical parser does not construct adequate discourse trees; or the mapping from discourse structures to importance scores is too simplistic. I examine now, in turn, each of these explanations.

The results in Chapter 10 strongly support the first explanation. As we have seen, different discourse interpretations yield different summaries. By carefully choosing adequate discourse interpretations, one can increase the performance of a discourse-based summarizer.

[**Smart cards are becoming more attractive**[1]] [as the price of microcomputing power and storage continues to drop.[2]] [**They have two main advantages over magnetic-stripe cards.**[3]] [**First, they can carry 10 or even 100 times as much information**[4]] [—and hold it much more robustly.[5]] [**Second, they can execute complex tasks in conjunction with a terminal.**[6]] [For example, a smart card can engage in a sequence of questions and answers that verifies the validity of information stored on the card and the identity of the card-reading terminal.[7]] [A card using such an algorithm might be able to convince a local terminal that its owner had enough money to pay for a transaction[8]] [without revealing the actual balance or the account number.[9]] [Depending on the importance of the information involved,[10]] [security might rely on a personal identification number[11]] [such as those used with automated teller machines,[12]] [a midrange encipherment system,[13]] [such as the Data Encryption Standard (DES),[14]] [or a highly secure public-key scheme.[15]]

[Smart cards are not a new phenomenon.[16]] [**They have been in development since the late 1970s**[17]] [and have found major applications in Europe,[18]] [with more than a quarter of a billion cards made so far.[19]] [The vast majority of chips have gone into prepaid, disposable telephone cards,[20]] [but even so the experience gained has reduced manufacturing costs,[21]] [improved reliability[22]] [and proved the viability of smart cards.[23]] [**International and national standards for smart cards are well under development**[24]] [to ensure that cards, readers and the software for the many different applications that may reside on them can work together seamlessly and securely.[25]] [Standards set by the International Organization for Standardization (ISO), for example, govern the placement of contacts on the face of a smart card[26]] [so that any card and reader will be able to connect.[27]]

Figure 11.1
The *Smart Cards* text *(Scientific American, August 1996)*

I turn now to the other possible explanation, the one that concerns the mapping from discourse structures to importance scores. If one examines the results in Table 9.5, one can see that the difference in recall and precision between the discourse-based and analyst-based summarizers is lower in the clause case than the difference between the analyst-based summarizer and the upper bound given by the judges. This suggests that a better mapping between discourse structures and importance scores may have a more significant impact on the quality of a discourse-based summarization program than a better rhetorical parser. In order to understand this claim, I should examine the cases in which recall and precision figures were low even for the discourse trees that were built by the analysts, which were supposed to be "perfect."

Let us examine closely the correlation between the discourse structure built by the first analyst for the text in Figure 11.1, which is given in Figure 11.2, and the units that the judges considered important in the same text. In Figure 11.1, the elementary units that a majority of the judges agreed to be important are shown in bold.

Figure 11.2 shows the discourse structure built by the first analyst. Each elementary unit in the structure is labeled with a number from 1 to 27. The numbers shown in bold that are

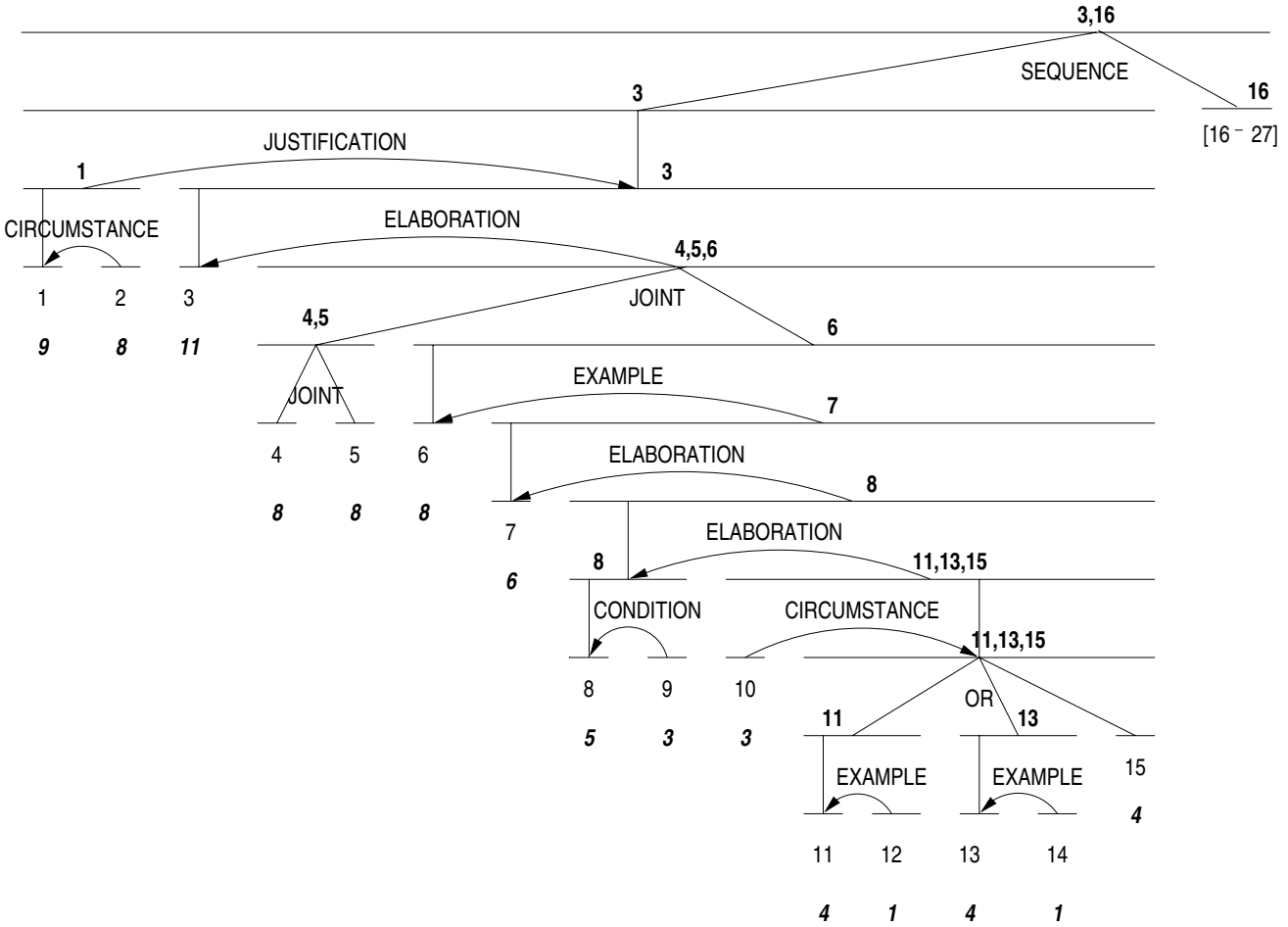

Figure 11.2
The discourse tree that was built for the text in Figure 11.1 by the first analyst

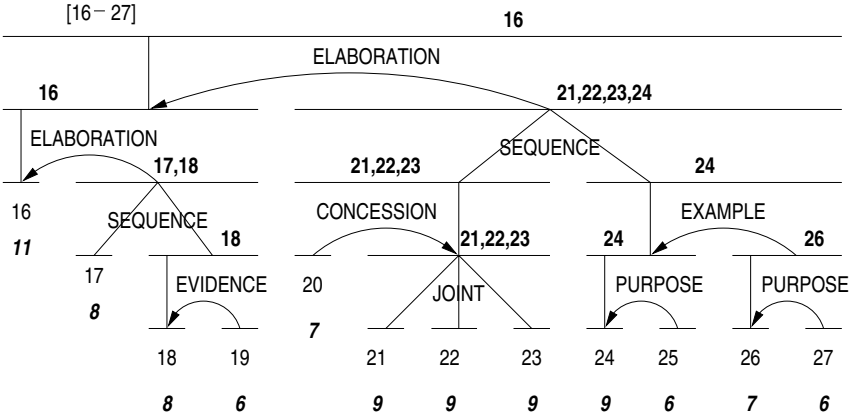

Figure 11.2 *(continued)*

associated with the non-elementary spans represent promotion units. The numbers shown in italics bold that are associated with the leaves represent the importance scores that are assigned by formula (9.1) to each elementary unit in the text. For example, the promotion units of span [1,27] are units 3 and 16, while the promotion units of span [10,15] are units 11, 13, and 15.

As I discussed before, when I evaluated the analyst-based summarizer, I selected from a partial ordering a number of units that reflected the number of units considered important in a text by the judges. In the text in Figure 11.1, six units were considered important: those labeled 1, 3, 4, 6, 17, 24. The partial ordering induced by formula (9.1) on the discourse structure of Figure 11.2 is that shown in 11.1 below.

$$3, 16 > 1, 21, 22, 23, 24 > 2, 4, 5, 6, 17, 18 > 26 > 19, 25, 27 > 8 > 11, 13, 15 > 9, 10 > 12, 14 \quad (11.1)$$

Selecting the first seven units in the partial ordering comes closest to the number of units that were considered important by the judges. Unfortunately, only three of the seven units that are selected by the analyst-based summarizer were considered important by a majority of the judges; these were units 1, 3, and 24. This gives a recall of 50% and a precision of 42.86%.

If one examines the discourse structure of the text in Figure 11.1 and the units that judges perceived to be important, one can notice a couple of interesting facts. For example, a majority of the judges labeled units 3, 4, and 6 as important. The discourse structure built by the analyst shows that an ELABORATION relation holds between units 4 and 3 and between units 6 and 3. Because units 4 and 6 are the satellites of the ELABORATION relation, they are assigned a lower score than unit 3. However, if one examines the text closely, one will find it natural to include in the summary not only the information that smart cards have two main advantages over magnetic-stripe cards (unit 3), but also the advantages per se, which are given in units 4 and 6. Hence, for certain kinds of ELABORATION relations, it seems adequate to assign a larger score to their satellites than formula 9.1 currently does. By examining the same discourse structure and the importance scores assigned by judges, one can see that none of the units in the span [7–15] were considered important. This observation seems to correlate with the fact that the whole span [7–15] is an exemplification of the information given in unit 6. If the observation that satellites of EXAMPLE relations are not important generalizes, then it would be appropriate to account for this in the formula that computes the importance scores.

Also interesting is the fact that judges considered unit 24 important, which seems to correlate with a topic shift. Again, if this observation generalizes, it will have to be properly accounted for by the formula that computes importance scores. To make things even more

difficult, consider the following two cases, in which the judges considered important only the first nucleus of a multinuclear relation. For example, although a rhetorical relation of JOINT holds between units 4 and 5 and a rhetorical relation of SEQUENCE holds between units 17 and 18, judges considered only units 4 and 17 important. According to formula 9.1, both pairs of units are assigned the same score. Obviously, mechanisms that are not inherent to the rhetorical structure of text are needed in order to explain why only one nucleus of a multinuclear relation is considered important by humans.

The discussion above suggests that there is definitely much more to assigning importance scores to textual units on the basis of a discourse structure than first meets the eye. Although formula 9.1 enables a discourse-based summarizer to derive summaries of good quality, there is definitely room for improvement. The experiments described in this part of the book suggest that there exists a correlation also between the types of relations that are used to connect various textual units and the importance of those units in a text. However, more elaborate experiments are needed in order to provide clear-cut evidence on the nature of this correlation.

11.2.2 Other Issues

Throughout this part of the book, I concentrated my attention only on the problem of selecting the most important units in a text. However, this solves only part of the summarization problem, because a complete system will also have to use the selected units in order to produce coherent text. I found that the summaries that are produced by the discourse-based summarizer read well—after all, the summarizer selects nuclei, which represent what is most essential for the writer's purpose and which can be understood independent of their satellites. Yet, I have not carried out any readability evaluation. One of the problems that the discourse-based summarizer still has is that of dangling references: in some cases, the selected units use anaphoric expressions to referents that were not selected. Dealing with these issues is, however, beyond the scope of this book.

11.3 Other Applications

Summarizing texts is only one of the possible applications of the formalization of valid text structures and the rhetorical parsing algorithms. In what follows, I discuss briefly four additional applications, which may also benefit from the work in this book.

SYNTACTIC PARSING. Current parsers that employ statistical techniques derive syntactic trees at labeled recall and precision levels of performance as high as 88% [Magerman, 1995, Collins, 1997, Charniak, 1997]. However, the longer the sentences given as input, the lower their performance. The segmentation algorithms in Chapters 6 and 7 determine

boundaries between elementary discourse units by relying not only on the local context, as most of the statistically based syntactic parsers, but also on the global one. For example, the discourse segmenter in Section 7.3.2 correctly learned to insert an elementary unit boundary after the first comma that follows the occurrence of the cue phrase *While* (see rule 6 in Figure 7.4). It is conceivable that by enabling the language models that are used in statistical parsing to account for such long-distance, discourse-specific phenomena, one can improve their performance. Since parsers work reasonably well on small clauses, it may be fruitful, for example, to explore a line of research in which a discourse segmenter determines the clauses of a sentence, a syntactic parser derives syntactic tree for each of these clauses, and a back-end component merges the syntactic trees of the individual clauses into syntactic trees that span the whole sentence given as input. These components can work both in serial or in parallel.

NATURAL LANGUAGE GENERATION. The information derived from the corpus studies in Sections 6.2 and 7.2 was exploited here only in the context of rhetorical parsing. However, as I have shown in [Marcu, 1997], the same information can be exploited in a Natural Language Generation (NLG) setting. The approach to NLG that I advocate in [Marcu, 1997] is well suited for cases in which one does not generate texts in order to satisfy a high-level communicative goal, but rather to convey information that has been already selected as relevant to a specific topic. For example, in a medical domain, a content selection module may select all information that is relevant to a given patient [DiMarco et al., 1995]. In a summarization setting, an extraction module such as that described in Chapter 9 may select all clauses that are important in a text. The task of an NLG system in this case is to produce a coherent output that subsumes all preselected information.

In order to solve this problem, I relied on an observation of Mann and Thompson [1988]. During the development of RST, they noticed that rhetorical relations exhibit strong patterns of ordering of their nuclei and satellites, which they called *canonical orderings*. The key idea of the approach to text planning that I proposed [Marcu, 1997] is to formalize both the strong tendency of semantic units that could be associated with the nuclei and satellites of various rhetorical relations to obey a given ordering; and the inclination of semantically and rhetorically related information to cluster into larger textual spans [Mooney et al., 1990, McCoy and Cheng, 1991]. In other words, this bottom-up approach to text planning assumes that global coherence can be achieved by satisfying the local constraints on ordering and clustering and by ensuring that the discourse tree that is eventually built is well formed.

In [Marcu, 1997], I showed how one can use the corpus data in Section 6.2 in order to estimate the clustering and ordering tendencies of the nuclei and satellites involved

in each type of rhetorical relation. I showed how these preferences can be encoded as constraints, and how the text planning problem can be solved by applying traditional constraint-satisfaction-based techniques.

MACHINE TRANSLATION. Previous research [Delin et al., 1994] has shown that in multilingual instructional texts, the same instructions are often realized using different rhetorical relations and rhetorical structures. In a preliminary study of Japanese texts and their English translations, Marcu et al. [2000] have noticed significant differences with respect to the way the same information is rhetorically rendered in the two languages. For example, what is conveyed in two or three sentences in one language is conveyed in only one sentence in another. What is conveyed using certain rhetorical relations and structures in one language, is conveyed using different rhetorical relations and structures in the other language. Obviously, if we translate texts on a sentence by sentence basis, we can end up with anomalous translations. Such translations are especially frequent when translating between languages that have very different syntactic structures, such as Japanese and English.

A possible research direction that would address these problems is to build a parallel corpus of, let's say, Japanese and English rhetorical structure trees. By employing techniques similar to those presented in Chapter 7, one can construct rhetorical parsers for Japanese. By using the parallel corpus, one can automatically learn rules that would enable one to map Japanese rhetorical structures into English rhetorical structures. And then one can translate the modified structure into English, thus obtaining texts that are not only syntactically, but also rhetorically well formed. For more details of such an approach see [Marcu et al., 2000].

INFORMATION EXTRACTION AND RETRIEVAL. The rhetorical parsers presented in the second part of the book enable one to determine the rhetorical structure of unrestricted texts. It is conceivable that once these structures are built, they can be saved on disk. If one saves the structure of all the texts in a collection, one can then query that collection using not only keywords, as traditional Information Retrieval systems, but also rhetorical relations. For example, one may be able to ask questions such as "what arguments support *x*?"; "what's the purpose of doing *y*?"; "is there anything that contradicts *z*?"; "what are the main steps/parts/members of *u*?"; "what is the consequence of doing *t*?"; etc. By selecting for example the text spans that contain the phrase *y* and by following the PURPOSE satellite links that are related to these spans, one can return information that is *rhetorically* relevant to query *y*. Posing and answering such queries is beyond the expressive power of current Information Retrieval systems.

11.4 Summary

Overall, the empirical and computational experiments that I described in this part of the book support at least the following conclusions.

1. Discourse trees are good indicators of textual importance.

2. The concepts of rhetorical analysis and nuclearity can be exploited algorithmically; the resulting discourse-based summarization system yields performance results that are not far from those obtained by humans.

3. For extracting 10% summaries of short articles of the news story genre, a simple lead-based algorithm is the most efficient solution.

4. For extracting longer summaries of short newspaper articles and for extracting any size summaries of complex (not necessarily news stories) newspaper articles, a simple lead-based algorithm does not provide a satisfactory solution. This assertion holds for other text genres, such as that of *Scientific American*, as well.

5. There is no magic key (heuristic) for obtaining good summarization results; rather the strength of a summarization system seems to come from its ability to combine a multitude of heuristics.

6. Combinations of heuristics that yield "optimal" results for certain summary extract lengths might not yield optimal results for different lengths.

7. Incorporating various heuristics into a discourse-based summarization framework improves its performance.

Bibliography

[Anscombre and Ducrot, 1983] Jean-Claude Anscombre and Oswald Ducrot. *L'argumentation dans la langue*. Bruxelles: Pierre Mardaga, 1983.

[Asher and Lascarides, 1994] Nicholas Asher and Alex Lascarides. Intentions and information in discourse. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL–94)*, pages 34–41, New Mexico State University, Las Cruces, NM, June 27-30 1994.

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

[Ballard et al., 1971] D. Lee Ballard, Robert Conrad, and Robert E. Longacre. The deep and surface grammar of interclausal relations. *Foundations of language*, 4:70–118, 1971.

[Barker and Szpakowicz, 1995] Ken Barker and Stan Szpakowicz. Interactive semantic analysis of clause-level relationships. In *Proceedings of the Second Conference of the Pacific Association for Computational Linguistics (PACLING–95)*, pages 22–30, Brisbane, Australia, 1995.

[Barton et al., 1985] Edward G. Barton, Robert C. Berwick, and Eric Sven Ristad. *Computational Complexity and Natural Language*. Cambridge, MA: The MIT Press, 1985.

[Barzilay and Elhadad, 1997] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL/EACL–97 Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain, July 11 1997.

[Bateman and Rondhuis, 1997] John A. Bateman and Klaas Jan Rondhuis. Coherence relations: Towards a general specification. *Discourse Processes*, 24:3–49, 1997.

[Baxendale, 1958] P. B. Baxendale. Machine-made index for technical literature—an experiment. *IBM Journal of Research and Development*, 2:354–361, 1958.

[Beeferman et al., 1999] Doug Beeferman, Adam Berger, and John Lafferty. Statistical models in text segmentation. *Machine Learning*, 34(1/3):177-210, February 1999.

[Bestgen and Costermans, 1997] Yves Bestgen and Jean Costermans. 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*, pages 201–218. Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

[Birnbaum, 1982] Lawrence Birnbaum. Argument molecules: a functional representation of argument structures. In *Proceedings of the Third National Conference on Artificial Intelligence (AAAI–82)*, pages 63–65, Pittsburgh, PA, August 18–20 1982.

[Birnbaum et al., 1980] Lawrence Birnbaum, Margot Flowers, and Rod McGuire. Towards an AI model of argumentation. In *Proceedings of the First National Conference on Artificial Intelligence (AAAI–80)*, pages 313–315, Stanford, CA, 1980.

[Blackburn et al., 1995] Patrick Blackburn, Wilfried Meyer-Viol, and Maarten de Rijke. A proof system for finite trees. Technical Report CLAUS–Report Nr. 67, University of Saarbrücken, Germany, October 1995.

[Brill, 1995] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565, December 1995.

[Briscoe, 1996] Ted Briscoe. 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 1996.

[Bruder and Wiebe, 1990] Gail A. Bruder and Janice M. Wiebe. 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 25-28 1990.

[Burstein et al., 1998] Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. Automated scoring using a hybrid feature identification technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics (COLING/ACL–98)*, pages 206–210, Montreal, Canada, August 1998.

[Carberry et al., 1993] Sandra Carberry, Jennifer Chu, Nancy Green, and Lynn Lambert. Rhetorical relations: Necessary but not sufficient. In Owen Rambow, editor, *Proceedings of the ACL Workshop on Intentionality and Structure in Discourse Relations*, pages 1–4, Columbus, OH, June 1993.

[Carletta et al., 1997] Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–32, March 1997.

[Caron, 1997] Jean Caron. Toward a procedural approach of the meaning of connectives. In Jean Costermans and Michel Fayol, editors, *Processing Interclausal Relationships. Studies in the Production and Comprehension of Text*, pages 53–74. Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

[Cawsey, 1991] Alison Cawsey. Generating interactive explanations. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI–91)*, volume 1, pages 86–91, Anaheim, CA, July 14–19 1991.

[Charniak, 1997] Eugene Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI–97)*, pages 598–603, Providence, RI, July 27–31 1997.

[Chou Hare and Borchardt, 1984] Victoria Chou Hare and Kathleen M. Borchardt. Direct instruction of summarization skills. *Reading Research Quarterly*, 20(1):62–78, Fall 1984.

[Cochran, 1950] William Gemmell Cochran. The comparison of percentages in matched samples. *Biometrika*, 37:256–266, 1950.

[Cohen, 1983] Robin Cohen. *A Computational Model for the Analysis of Arguments*. Ph.D. thesis, Department of Computer Science, University of Toronto, 1983. Also published as Technical Report CSRI–151, Computer Systems Research Institute, University of Toronto.

[Cohen, 1987] Robin Cohen. Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13(1–2):11–24, January–June 1987.

[Collins, 1997] Michael Collins. Three generative, lexicalized models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL–97)*, pages 16–23, Madrid, Spain, July 7-12 1997.

[Cook and Mayer, 1988] Linda K. Cook and Richard E. Mayer. Teaching readers about the structure of scientific text. *Journal of Educational Psychology*, 80(4):448–456, 1988.

[Corston-Oliver, 1998] Simon H. Corston-Oliver. 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, CA, March 23–25 1998.

[Crawford and Auton, 1996] James M. Crawford and Larry D. Auton. Experimental results on the crossover point in random 3SAT. *Artificial Intelligence*, 81(1-2):31–57, 1996.

[Cristea and Webber, 1997] Dan Cristea and Bonnie L. Webber. Expectations in incremental discourse processing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL–97)*, pages 88–95, Madrid, Spain, July 7–12 1997.

[Cristea et al., 1998] Dan Cristea, Nancy Ide, and Laurent Romary. Veins theory: A model of global discourse cohesion and coherence. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics (COLING/ACL–98)*, pages 281–285, Montreal, Canada, August 1998.

[Cristea et al., 1999] Dan Cristea, Nancy Ide, Daniel Marcu, and Valentin Tablan. 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, College Park, MD, June 21 1999.

[Crystal, 1991] David Crystal. *A Dictionary of Linguistics and Phonetics*. Oxford: Basil Blackwell, 3rd edition, 1991.

[Cumming and McKercher, 1994] Carmen Cumming and Catherine McKercher. *The Canadian Reporter: News Writing and Reporting*. San Diego, CA: Hartcourt Brace, 1994.

[Davis and Putnam, 1960] Martin Davis and Hilary Putnam. A computing procedure for quantification theory. *Journal of the Association for Computing Machinery*, 7(3):201–215, 1960.

[de Villiers, 1974] P. A. de Villiers. Imagery and theme in recall of connected discourse. *Journal of Experimental Psychology*, 103:263–268, 1974.

[Decker, 1985] Nan Decker. The use of syntactic clues in discourse processing. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics (ACL–85)*, pages 315–323, Chicago, IL, July 8–12 1985.

[Delin and Oberlander, 1992] Judy L. Delin and Jon Oberlander. Aspect-switching and subordination: the role of *it*-clefts in discourse. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING–92)*, pages 281–287, Nantes, France, August 23–28 1992.

[Delin et al., 1994] Judy L. Delin, Anthony Hartley, Cécile L. Paris, Donia R. Scott, and Keith Vander Linden. Expressing procedural relationships in multilingual instructions. In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 61–70, Kennebunkport, ME, June 1994.

[Di Eugenio, 1992] Barbara Di Eugenio. Understanding natural language instructions: the case of purpose clauses. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL–92)*, Newark, DE, pages 120–127, June 28–July 2 1992.

[Di Eugenio, 1993] Barbara Di Eugenio. *Understanding Natural Language Instructions: A Computational Approach to Purpose Clauses*. Ph.D. thesis, University of Pennsylvania, December 1993. Also published as Technical Report IRCS 93-52, The Institute for Research in Cognitive Science.

[Di Eugenio et al., 1997] Barbara Di Eugenio, Johanna D. Moore, and Massimo Paolucci. Learning features that predict cue usage. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL–97)*, pages 80–87, Madrid, Spain, July 7-12 1997.

[DiMarco et al., 1995] Chrysanne DiMarco, Graeme Hirst, Leo Wanner, and John Wilkinson. HealthDoc: Customizing patient information and health education by medical condition and personal characteristics. Proceedings of the Workshop on Artificial Intelligence in Patient Education, Glasgow, Scotland, pages 59–71, August 7–9 1995.

[Donlan, 1980] Dan Donlan. Locating main ideas in history textbooks. *Journal of Reading*, 24:135–140, 1980.

[Edmundson, 1968] H. P. Edmundson. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285, April 1968.

[Elhadad and McKeown, 1990] Michael Elhadad and Kathleen R. McKeown. Generating connectives. In *Proceedings of the International Conference on Computational Linguistics (COLING–90)*, volume 3, pages 97–102, Helsinki, Finland, 1990.

[Fellbaum, 1998] Christiane Fellbaum, editor. *Wordnet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press, 1998.

[Fox, 1987] Barbara Fox. *Discourse Structure and Anaphora*. Cambridge Studies in Linguistics; 48. Cambridge, England: Cambridge University Press, 1987.

[Fraser, 1990] Bruce Fraser. An approach to discourse markers. *Journal of Pragmatics*, 14:383–395, 1990.

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

[Gale et al., 1992] William Gale, Kenneth W. Church, and David Yarowsky. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL–92)*, pages 249–256, Newark, DE, June 28–July 2 1992.

[Gardent, 1994] Claire Gardent. Discourse multiple dependencies. Technical Report CLAUS-Report Nr. 45, Universität des Saarlandes, Saarbrücken, Germany, October 1994.

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

[Garner, 1982] Ruth Garner. Efficient text summarization: costs and benefits. *Journal of Educational Research*, 75:275–279, 1982.

[Gernsbacher, 1997] Morton Ann Gernsbacher. Coherence cues mapping during comprehension. In Jean Costermans and Michel Fayol, editors, *Processing Interclausal Relationships. Studies in the Production and Comprehension of Text*, pages 3–22. Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

[Givón, 1983] Talmy Givón. Topic continuity in discourse: an introduction. In Talmy Givón, editor, *Topic continuity in discourse: a quantitative cross-language study*, pages 1–41. Philadelphia, PA: John Benjamins Publishing Company, 1983.

[Givón, 1995] Talmy Givón. Coherence in text vs. coherence in mind. In Morton Ann Gernsbacher and Talmy Givón, editors, *Coherence in spontaneous text*, volume 31 of *Typological Studies in Language*, pages 59–115. Philadelphia, PA: John Benjamins Publishing Company, 1995.

[Gladwin et al., 1991] Philip Gladwin, Stephen Pulman, and Karen Sparck Jones. Shallow processing and automatic summarizing: A first study. Technical Report 223, University of Cambridge Computer Laboratory, Cambridge, England, May 1991.

[Glover et al., 1988] John A. Glover, Dale L. Dinnel, Dale R. Halpain, Todd K. McKee, Alice J. Corkill, and Steven L. Wise. Effects of across-chapter signals on recall of text. *Journal of Educational Psychology*, 80(1):3–15, 1988.

[Grimes, 1975] Joseph Evans Grimes. *The Thread of Discourse*. The Hague, Paris: Mouton, 1975.

[Grosz and Hirschberg, 1992] Barbara J. Grosz and Julia Hirschberg. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, pages 429–432, Banff, Canada, 1992.

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

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

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

[Hahn and Strube, 1997] Udo Hahn and Michael Strube. Centering in-the-large: Computing referential discourse segments. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL–97)*, pages 104–111, Madrid, Spain, July 7–12 1997.

[Halliday and Hasan, 1976] Michael A.K. Halliday and Ruqaiya Hasan. *Cohesion in English*. London, England: Longman, 1976.

[Halliday, 1994] Michael A.K. Halliday. *An Introduction to Functional Grammar*. London, England: Edward Arnold, second edition, 1994.

[Harabagiu and Maiorano, 1999] Sanda Harabagiu and Steven Maiorano. Knowledge-lean coreference resolution and its relation to textual cohesion and coreference. In *Proceedings of the ACL Workshop on Discourse/Dialogue Structure and Reference*, pages 29–38, University of Maryland, College Park, MD, June 21 1999.

[Harabagiu and Moldovan, 1996] Sanda Harabagiu and Dan I. Moldovan. 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, 1996.

[Hearst, 1994] Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL–94)*, pages 9–16, New Mexico State University, Las Cruces, NM, June 27–30 1994.

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

[Hermjakob and Mooney, 1997] Ulf Hermjakob and Raymond J. Mooney. Learning parse and translation decisions from examples with rich context. In *Proceedings of the 35th Annual Meeting of the Association for*

Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL–97), pages 482–489, Madrid, Spain, July 7–12 1997.

[Heurley, 1997] Laurent Heurley. 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*, pages 179–200. Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

[Hirschberg and Litman, 1987] Julia Hirschberg and Diane J. Litman. Now let's talk about *now*: Identifying cue phrases intonationally. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL–87)*, pages 163–171, Stanford, CA, 1987.

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

[Hirschberg and Nakatani, 1996] Julia Hirschberg and Christine H. Nakatani. A prosodic analysis of discourse segments in direction-given monologues. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL–96)*, pages 286–293, Santa Cruz, CA, June 24–27 1996.

[Hirschman et al., 1999] Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL–99)*, pages 325–332, University of Maryland, College Park, MD, June 20–26 1999.

[Hirst et al., 1993] Graeme Hirst, Susan W. McRoy, Peter Heeman, Philip Edmonds, and Diane Horton. Repairing conversational misunderstandings and non-understandings. In *International Symposium on Spoken Dialogue—New Directions in Human and Man-Machine Communication*, pages 185–196, Tokyo, Japan, Nov 10-12 1993.

[Hitzeman, 1995] Janet Hitzeman. Text type and the position of a temporal adverbial within the sentence. Technical Report Deliverable R1.3.2b, ESPRIT Research Project 6665, University of Edinburgh, Scotland, November 28 1995.

[Hitzeman et al., 1995] Janet Hitzeman, Marc Moens, and Claire Grover. Algorithms for analyzing the temporal structure of discourse. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL–95)*, Dublin, Ireland, 1995.

[Hobbs, 1990] Jerry R. Hobbs. *Literature and Cognition*. CSLI Lecture Notes Number 21, Stanford, CA: Cambridge University Press, 1990.

[Hobbs, 1993] Jerry R. Hobbs. Summaries from structure. In *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication*, Dagstuhl, Germany, December 13–17 1993.

[Hobbs, 1996] Jerry R. Hobbs. On the relation between the informational and intentional perspectives on discourse. In Eduard H. Hovy and Donia R. Scott, editors, *Computational and Conversational Discourse. Burning Issues – An Interdisciplinary Account*, chapter 6, pages 139–157. Heidelberg, Germany: Springer Verlag, 1996.

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

[Hoey, 1991] Michael Hoey. *Patterns of Lexis in Text*. Oxford, England: Oxford University Press, 1991.

[Hoover, 1997] Michael L. Hoover. Effects of textual and cohesive structure on discourse processing. *Discourse Processes*, 23:193–220, 1997.

[Horacek, 1992] Helmut Horacek. An integrated view of text planning. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation; 6th International Workshop on Natural Language Generation*, Trento, Italy, April 1992. Number 587 in Lecture Notes in Artificial Intelligence, pages 29–44. Heidelberg, Germany: Springer-Verlag, 1992.

[Hovy, 1988] Eduard H. Hovy. Planning coherent multisentential text. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (ACL–88)*, pages 163–169, State University of New York at Buffalo, NY, June 27–30 1988.

[Hovy, 1990] Eduard H. Hovy. Unresolved issues in paragraph planning. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation*, pages 17–45. New York: Academic Press, 1990.

[Hovy, 1993] Eduard H. Hovy. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63(1–2):341–386, October 1993.

[Hovy and Lin, 1997] Eduard H. Hovy and Chin-Yew Lin. Automated text summarization in SUMMARIST. In *Proceedings of the ACL/EACL–97 Workshop on Intelligent Scalable Text Summarization*, pages 18–24, Madrid, Spain, July 11 1997.

[Hovy and Lin, 1999] Eduard H. Hovy and Chin-Yew Lin. Automated multilingual text summarization and its evaluation. *Computational Intelligence*, To appear, 1999.

[Hovy and Maier, 1993] Eduard H. Hovy and Elisabeth Maier. Parsimonious or profligate: How many and which discourse structure relations? Unpublished Manuscript, 1993.

[Jing et al., 1998] Hongyan Jing, Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Summarization evaluation methods: Experiments and analysis. In *Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization*, pages 60–68, Stanford, CA, March 23–25 1998.

[Johnson, 1970] Ronald E. Johnson. Recall of prose as a function of structural importance of linguistic units. *Journal of Verbal Learning and Verbal Behaviour*, 9:12–20, 1970.

[Joshi, 1987] Aravind K. Joshi. An introduction to tree adjoining grammar. In Alexis Manaster-Ramer, editor, *Mathematics of Language*. Philadelphia, PA: John Benjamins, 1987.

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

[Kamp, 1979] J.A.W. Kamp. Events, instants, and temporal reference. In R. Bäuerle, U. Egli, and A. Karmiloff-Smith, editors, *Semantics from Different Points of View*. Berlin: Springer Verlag, 1979.

[Kautz and Selman, 1996] Henry Kautz and Bart Selman. Pushing the envelope: Planning, propositional logic, and stochastic search. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI–96)*, volume 2, pages 1194–1201, Portland, OR, August 4–8 1996.

[Keller, 1993] Bill Keller. Feature Logics, Infinitary Descriptions and Grammar. CSLI Lecture Notes Number 44, Stanford, CA: Cambridge University Press, 1993.

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

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

[Knott and Dale, 1996] Alistair Knott and Robert Dale. Choosing a set of coherence relations for text generation: a data-driven approach. In M. Zock, editor, *Trends in Natural Language Generation: an Artificial Intelligence Perspective*, pages 47–67. Heidelberg, Germany: Springer Verlag, 1996.

[Knott and Mellish, 1996] Alistair Knott and Chris Mellish. A feature-based account of the relations signalled by sentence and clause connectives. *Journal of Language and Speech*, 39(2–3):143–183, 1996.

[Knott and Sanders, 1998] Alistair Knott and Ted J. M. Sanders. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30:135–175, 1998.

[Kohavi et al., 1996] Ronny Kohavi, Dan Sommerfield, and James Dougherty Data mining using MLC++: A machine learning library in C++. In *Proceedings of the Internaltional Conference on Tools with Artificial Intelligence (ICTAI–96)*, pages 234-245. 1996. http://www.sgi.com/Technology/mlc.

[Krippendorff, 1980] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Beverly Hills, CA: Sage Publications, 1980.

[Kupiec et al., 1995] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, WA, 1995.

[Kurohashi and Nagao, 1994] Sadao Kurohashi and Makoto Nagao. 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 1123–1127, Kyoto, Japan, August 5–9 1994.

[Lascarides and Asher, 1991] Alex Lascarides and Nicholas Asher. Discourse relations and defeasible knowledge. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL–91)*, pages 55–62, Berkeley, CA, June 17–21 1991.

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

[Lascarides and Oberlander, 1992] Alex Lascarides and Jon Oberlander. Abducing temporal discourse. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation; 6th International Workshop on Natural Language Generation*, Trento, Italy, April 1992. Number 587 in Lecture Notes in Artificial Intelligence, pages 167–182. Heidelberg, Germany: Springer-Verlag, 1992.

[Lascarides et al., 1992] Alex Lascarides, Nicholas Asher, and Jon Oberlander. Inferring discourse relations in context. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL–92)*, pages 1–8, Newark, DE, June 28–July 2 1992.

[Lin, 1998] Chin-Yew Lin. Assembly of topic extraction modules in SUMMARIST. In *Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization*, pages 53–59, Stanford, CA, March 23–25 1998.

[Lin and Hovy, 1997] Chin-Yew Lin and Eduard H. Hovy. Identifying topics by position. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP–97)*, pages 283–290, Washington, D.C., March 31–April 3 1997.

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

[Lloyd, 1987] John Wylie Lloyd. *Foundations of Logic Programming*. Heidelberg, Germany: Springer Verlag, second edition, 1987.

[Loman and Mayer, 1983] Nancy Lockitch Loman and Richard E. Mayer. Signaling techniques that increase the understandability of expository prose. *Journal of Educational Psychology*, 75(3):402–412, 1983.

[Longacre, 1983] Robert E. Longacre. *The Grammar of Discourse*. New York: Plenum Press, 1983.

[Lorch and Lorch, 1985] Robert F. Lorch and Elizabeth Pugzles Lorch. Topic structure representation and text recall. *Journal of Educational Psychology*, 77(2):137–148, 1985.

[Luhn, 1958] H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, April 1958.

[Magerman, 1995] David M. Magerman. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL–95)*, pages 276–283, Cambridge, MA, June 26-30 1995.

[Maier, 1993] Elisabeth Maier. The extension of a text planner for the treatment of multiple links between text units. In *Proceedings of the Fourth European Workshop on Natural Language Generation (ENLG–93)*, pages 103–114, Pisa, Italy, April 28–30 1993.

[Mani and Bloedorn, 1998] Inderjeet Mani and Eric Bloedorn. Machine learning of generic and user-focused summarization. In *Proceedings of Fifteenth National Conference on Artificial Intelligence (AAAI–98)*, pages 821–826, Madison, Wisconsin, July 26–30 1998.

[Mani et al., 1998] Inderjeet Mani, Eric Bloedorn, and Barbara Gates. Using cohesion and coherence models for text summarization. In *Proceedings of the AAAI'98 Spring Symposium on Intelligent Text Summarization*, pages 69–78, Stanford, CA, March 23–25 1998.

[Mann and Thompson, 1988] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[Marcu, 1997] Daniel Marcu. From local to global coherence: A bottom-up approach to text planning. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI–97)*, pages 629–635, Providence, Rhode Island, July 28-31 1997.

[Marcu, 1998a] Daniel Marcu. *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. thesis, Department of Computer Science, University of Toronto, January 1998.

[Marcu, 1998b] Daniel Marcu. *Instructions for Manually Annotating the Discourse Structures of Texts*. Information Sciences Institute, University of Southern California, 1998.

[Marcu, 1999a] Daniel Marcu. Discourse trees are good indicators of importance in text. In Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, chapter 11, pages 123–136. Cambridge, MA: The MIT Press, 1999.

[Marcu, 1999b] Daniel Marcu. A formal and computational synthesis of Grosz and Sidner's and Mann and Thompson's theories. In *Proceedings of the Workshop on Levels of Representation in Discourse*, pages 101–108, Edinburgh, Scotland, July 7–9 1999.

[Marcu, 1999c] Daniel Marcu. *Instructions for Manually Annotating the Discourse Structures of Texts*. Information Sciences Institute, University of Southern California, 1999.

[Marcu, 2000] Daniel Marcu. The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. *Computational Linguistics*, To appear, 2000.

[Marcu et al., 1999a] Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*, pages 48–57, University of Maryland, College Park, MD, June 22 1999.

[Marcu et al., 1999b] Daniel Marcu, Magdalena Romera, and Estibaliz Amorrortu. Experiments in constructing a corpus of discourse trees: Problems, annotation choices, issues. In *Proceedings of the Workshop on Levels of Representation in Discourse*, pages 71–78, Edinburgh, Scotland, July 7–9 1999.

[Marcu et al., 2000] Daniel Marcu, Lynn Carlson, and Maki Watanabe. The Automatic Translation of Discourse Structures. In *Proceedings of the First Annual Meeting of the North American Association for Computational Linguistics*, Seattle, WA, April 29–May 4 2000.

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

[Matthiessen and Thompson, 1988] Christian Matthiessen and Sandra A. Thompson. The structure of discourse and 'subordination'. In J. Haiman and Sandra A. Thompson, editors, *Clause combining in grammar and discourse*, volume 18 of *Typological Studies in Language*, pages 275–329. Philadelphia, PA: John Benjamins Publishing Company, 1988.

[Maxwell and Kaplan, 1993] John T. Maxwell and Ronald M. Kaplan. The interface between phrasal and functional constraints. *Computational Linguistics*, 19(4):571–590, December 1993.

[Maybury, 1992] Mark T. Maybury. Communicative acts for explanation generation. *International Journal of Man–Machine Studies*, 37:135–172, 1992.

[McCoy and Cheng, 1991] Kathleen F. McCoy and Jeannette Cheng. Focus of attention: Constraining what can be said next. In Cécile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 103–124. London, Boston, Dordrecht: Kluwer Academic Publishers, 1991.

[McRoy, 1993] Susan W. McRoy. *Abductive Interpretation and Reinterpretation of Natural Language Utterances*. Ph.D. thesis, Department of Computer Science, University of Toronto, 1993.

[Meteer, 1992] Marie W. Meteer. *Expressibility and the Problem of Efficient Text Planning*. London: Pinter Publishers, 1992.

[Miike et al., 1994] Seiji Miike, Etsuo Itoh, Kenji Ono, and Kazuo Sumita. A full-text retrieval system with a dynamic abstract generation function. In *Proceedings of the Seventeenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 152–161, Dublin, Ireland, July 3–6 1994.

[Moens and Steedman, 1988] Marc Moens and Mark Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28, 1988.

[Montague, 1973] Richard Montague. The proper treatment of quantification in ordinary english. In K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes, editors, *Approaches to Natural Language*. Dordrecht: Kluwer Academic Publishers, 1973.

[Mooney et al., 1990] David J. Mooney, Sandra Carberry, and Kathleen F. McCoy. The generation of high-level structure for extended explanations. In *Proceedings of the International Conference on Computational Linguistics (COLING–90)*, volume 2, pages 276–281, Helsinki, Finland, 1990.

[Moore and Paris, 1993] Johanna D. Moore and Cécile L. Paris. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651–694, 1993.

[Moore and Pollack, 1992] Johanna D. Moore and Martha E. Pollack. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544, 1992.

[Moore and Swartout, 1991] Johanna D. Moore and William R. Swartout. A reactive approach to explanation: Taking the user's feedback into account. In Cécile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 3–48. Dordrecht: Kluwer Academic Publishers, 1991.

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

[Morrow, 1986] Daniel G. Morrow. Grammatical morphemes and conceptual structure in discourse processing. *Cognitive Science*, 10:423–455, 1986.

[Moser and Moore, 1995] Megan Moser and Johanna D. Moore. Investigating cue selection and placement in tutorial discourse. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL–95)*, pages 130–135, Cambridge, MA, June 26-30 1995.

[Moser and Moore, 1996] Megan Moser and Johanna D. Moore. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419, September 1996.

[Moser and Moore, 2000] Megan Moser and Johanna D. Moore. On the correlation of cues with discourse structure: Results from a corpus study. *Forthcoming*, 2000.

[Nakatani et al., 1995] Christine H. Nakatani, Julia Hirschberg, and Barbara J. Grosz. Discourse structure in spoken language: Studies on speech corpora. In *Working Notes of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 106–112, Stanford, CA, March 1995.

[Nicholas, 1994] Nick Nicholas. Problems in the application of Rhetorical Structure Theory to text generation. Master's thesis, University of Melbourne, Australia, June 1994.

[Noordman and Vonk, 1997] Leo G. M. Noordman and Wietske Vonk. Toward a procedural approach of the meaning of connectives. In Jean Costermans and Michel Fayol, editors, *Processing Interclausal Relationships. Studies in the Production and Comprehension of Text*, pages 75–93. Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

[Norvig, 1992] Peter Norvig. *Paradigms of Artificial Intelligence Programming: Case Studies in Common Lisp*. San Mateo, CA: Morgan Kaufmann Publishers, 1992.

[Nunberg, 1990] Geoffrey Nunberg. *The linguistics of punctuation*. CSLI Lecture Notes 18, Stanford, CA. Chicago, IL: University of Chicago Press, 1990.

[Ono and Thompson, 1996] Tsuyoshi Ono and Sandra A. Thompson. Interaction and syntax in the structure of conversational discourse: Collaboration, overlap, and syntactic dissociation. In Eduard H. Hovy and Donia R. Scott, editors, *Computational and Conversational Discourse. Burning Issues—An Interdisciplinary Account*, chapter 3, pages 67–96. Heidelberg, Germany: Springer Verlag, 1996.

[Ono et al., 1994] Kenji Ono, Kazuo Sumita, and Seiji Miike. Abstract generation based on rhetorical structure extraction. In *Proceedings of the International Conference on Computational Linguistics (Coling–94)*, pages 344–348, Kyoto, Japan, August 5–9 1994.

[Oversteegen, 1997] Leonoor E. Oversteegen. On the pragmatic nature of causal and contrastive connectives. *Discourse Processes*, 24:51–85, 1997.

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

[Palmere et al., 1983] Mark Palmere, Stephen L. Benton, John A. Glover, and Royce R. Ronning. Elaboration and recall of main ideas in prose. *Journal of Educational Psychology*, 75(6):898–907, 1983.

[Pander Maat, 1998] Henk Pander Maat. Classifying negative coherence relations on the basis of linguistic evidence. *Journal of Pragmatics*, 30:177–204, 1998.

[Pascual and Virbel, 1996] Elsa Pascual and Jacques Virbel. 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 1996.

[Passonneau, 1997] Rebecca J. Passonneau. Using centering to relax informational constraints on discourse anaphoric noun phrases. *Language and Speech*, 39(1-2), 1997. Special Issue devoted to Discourse, Syntax, and Information.

[Passonneau, 1998] Rebecca J. Passonneau. Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In Ellen Prince, Aravind K. Joshi, and Marilyn Walker, editors, *Centering Theory in Discourse*. Oxford, England: Oxford University Press, 1998.

[Passonneau and Litman, 1993] Rebecca J. Passonneau and Diane J. Litman. Intention-based segmentation: human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL–93)*, pages 148–155, Columbus, OH, June 22-26 1993.

[Passonneau and Litman, 1996] Rebecca J. Passonneau and Diane J. Litman. Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence, and linguistic devices. In Eduard H. Hovy and Donia R. Scott, editors, *Computational and Conversational Discourse. Burning Issues – An Interdisciplinary Account*, Heidelberg, Germany: Springer Verlag, 1996.

[Passonneau and Litman, 1997] Rebecca J. Passonneau and Diane J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–140, March 1997.

[Pfau, 1995] Michael Pfau. Designing messages for behavioral inoculation. In E. Maibach and R. L. Parrott, editors, *Designing Health Messages. Approaches from Communication Theory and Public Health Practice*, pages 99–113. Beverly Hills, CA: Sage Publications, 1995.

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

[Polanyi, 1993] Livia Polanyi. Linguistic dimensions of text summarization. In *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication*, Dagstuhl, Germany, December 13–17 1993.

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

[Polanyi and van den Berg, 1996] Livia Polanyi and Martin H. van den Berg. 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, Amsterdam, The Netherlands, 1996.

[Prince, 1978] Ellen F. Prince. A comparison of *it*-clefts and *wh*-clefts in discourse. *Language*, 54:883–906, 1978.

[Quinlan, 1993] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

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

[Reynar, 1999] Jeff Reynar. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL–99)*, pages 357-364, University of Maryland, College Park, MD, June 20–26 1999.

[Rino and Scott, 1996] Lucia Helena Machado Rino and Donia R. Scott. A discourse model for gist preservation. In *Proceedings of the Thirteen Brazilian Symposium on Artificial Intelligence*, Curitiba, Brazil, October 1996.

[Rogers, 1994] James Rogers. *Studies in the Logic of Trees with Applications to Grammar Formalisms*. Ph.D. thesis, University of Delaware, Department of Computer Science, 1994.

[Rogers, 1996] James Rogers. A model-theoretic framework for theories of syntax. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL–96)*, pages 10–16, Santa Cruz, CA, June 24–27 1996.

[Rösner and Stede, 1992] Dietmar Rösner and Manfred Stede. Customizing RST for the automatic production of technical manuals. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation; 6th International Workshop on Natural Language Generation*, Trento, Italy, April 1992. Number 587 in Lecture Notes in Artificial Intelligence, pages 199–214. Heidelberg, Germany: Springer-Verlag, 1992.

[Russell and Norvig, 1995] Stuart Russell and Peter Norvig. *Artificial Intelligence. A Modern Approach*. Englewood Cliffs, New Jersey: Prentice Hall, 1995.

[Sacks et al., 1974] Harvey Sacks, Emmanuel Schegloff, and Gail Jefferson. A simple systematics for the organization of turntaking in conversation. *Language*, 50:696–735, 1974.

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

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

[Sanders, 1997] Ted J. M. Sanders. Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes*, 24:119–147, 1997.

[Sanders et al., 1992] Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. Toward a taxonomy of coherence relations. *Discourse Processes*, 15:1–35, 1992.

[Sanders et al., 1993] Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics*, 4(2):93–133, 1993.

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

[Scha and Polanyi, 1988] Remko Scha and Livia Polanyi. An augmented context free grammar for discourse. In *Proceedings of the International Conference on Computational Linguistics (COLING–88)*, pages 573–577, Budapest, Hungary, August 1988.

[Schiffrin, 1987] Deborah Schiffrin. *Discourse Markers*. Cambridge, England: Cambridge University Press, 1987.

[Schilder, 1997] Frank Schilder. 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 1997.

[Schneuwly, 1997] Bernard Schneuwly. 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*, pages 245–263. Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

[Scott and de Souza, 1990] Donia R. Scott and Clarisse Sieckenius de Souza. Getting the message across in RST-based text generation. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation*, pages 47–73. New York: Academic Press, 1990.

[Segal and Duchan, 1997] Erwin M. Segal and Judith F. Duchan. 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*, pages 95–119. Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

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

[Selman et al., 1992] Bart Selman, Hector Levesque, and David Mitchell. A new method for solving hard satisfiability problems. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI–92)*, pages 440–446, San Jose, CA, July 1992.

[Selman et al., 1994] Bart Selman, Henry Kautz, and Bram Cohen. Noise strategies for improving local search. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI–94)*, pages 337–343, Seattle, WA, July 31–August 4 1994.

[Sherrard, 1989] Carol Sherrard. Teaching students to summarize: Applying textlinguistics. *System*, 17(1), 1989.

[Shiuan and Ann, 1996] Peh Li Shiuan and Christopher Ting Hian Ann. 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 1996.

[Sidner, 1981] Candace L. Sidner. Focusing for interpretation of pronouns. *Computational Linguistics*, 7(4):217–231, October–December 1981.

[Sidner, 1983] Candace L. Sidner. Focusing in the comprehension of definite anaphora. In M. Bady and R. Berwick, editors, *Computational Models of Discourse*, pages 267–330. Cambridge, MA: The MIT Press, 1983.

[Siegel and McKeown, 1994] Eric V. Siegel and Kathleen R. McKeown. 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, July 31–August 4 1994.

[Simmons and Yu, 1992] Robert F. Simmons and Yeong-Ho Yu. The acquisition and use of context-dependent grammars for English. *Computational Linguistics*, 18(4):391–418, 1992.

[Siskind and McAllester, 1993a] Jeffrey M. Siskind and David A. McAllester. Nondeterministic Lisp as a substrate for Constraint Logic Programming. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI–93)*, pages 133–138, Washington, D.C., July 11-15 1993.

[Siskind and McAllester, 1993b] Jeffrey M. Siskind and David A. McAllester. Screamer: A portable efficient implementation of nondeterministic Common Lisp. Technical Report IRCS–93–03, University of Pennsylvania, Institute for Research in Cognitive Science, July 1 1993.

[Skorochodko, 1971] E.F. Skorochodko. Adaptive method of automatic abstracting and indexing. In *Information Processing*, volume 2, pages 1179–1182, 1971.

[Sparck Jones, 1993a] Karen Sparck Jones. Summarising: analytic framework, key component, experimental method. In *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication*, Dagstuhl, Germany, December 13–17 1993.

[Sparck Jones, 1993b] Karen Sparck Jones. What might be in a summary? In *Information Retrieval 93: Von der Modellierung zur Anwendung*, pages 9–26, Universitätsverlag Konstanz, 1993.

[Sparck Jones, 1999] Karen Sparck Jones. Introduction to text summarization. In Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, pages 1–12. Cambridge, MA: The MIT Press, 1999.

[Spooren, 1997] Wilbert P. M. Spooren. The processing of underspecified coherence relations. *Discourse Processes*, 24:149–168, 1997.

[Sumita et al., 1992] Kazuo Sumita, Kenji Ono, T. Chino, Teruhiko Ukita, and Shin'ya Amano. A discourse structure analyzer for Japanese text. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, volume 2, pages 1133–1140, Tokyo, Japan, June 1–5 1992.

[Sweetser, 1990] Eve Sweetser. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge, England: Cambridge University Press, 1990.

[Talmy, 1983] Leonard Talmy. How language structures space. In H. Pick and L. Acredolo, editors, *Spatial Orientation: Theory, Research, and Application*. New York: Plenum Press, 1983.

[Teufel and Moens, 1997] Simone Teufel and Marc Moens. Sentence extraction as a classification task. In *Proceedings of the ACL/EACL–97 Workshop on Intelligent Scalable Text Summarization*, pages 58–65, Madrid, Spain, July 11 1997.

[Tomita, 1985] Masaru Tomita. *Efficient Parsing for Natural Language, A Fast Algorithm for Practical Systems*. Dordrecht: Kluwer Academic Publishers, 1985.

[Toulmin et al., 1979] Stephen Toulmin, Richard Rieke, and Allan Janik. *An Introduction to Reasoning*. London, England: Macmillan Publishing, 1979.

[Traxler and Gernsbacher, 1995] Matthew J. Traxler and Morton Ann Gernsbacher. Improving coherence in written communication. In Morton Ann Gernsbacher and T. Givón, editors, *Coherence in spontaneous text*,

volume 31 of *Typological Studies in Language*, pages 215–237. Philadelphia, PA: John Benjamins Publishing Company, 1995.

[van den Berg, 1996] Martin H. van den Berg. 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, Amsterdam, The Netherlands, 1996.

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

[van Dijk, 1979] Teun A. van Dijk. Pragmatic connectives. *Journal of Pragmatics*, 3:447–456, 1979.

[Vander Linden and Martin, 1995] Keith Vander Linden and James R. Martin. Expressing rhetorical relations in instructional text: A case study of the purpose relation. *Computational Linguistics*, 21(1):29–58, March 1995.

[Vonk et al., 1992] Wietske Vonk, Lettica G. M. M. Hustinx, and Wim H. G. Simons. The use of referential expressions in structuring discourse. *Language and Cognitive Processes*, 7(3,4):301–333, 1992.

[Walker, 1996] Marilyn A. Walker. Limited attention and discourse structure. *Computational Linguistics*, 22(2):255–264, 1996.

[Walker, 1998] Marilyn A. Walker. Centering, anaphora resolution, and discourse structure. In Ellen Prince, Aravind K. Joshi, and Marilyn Walker, editors, *Centering Theory in Discourse*. Oxford, England: Oxford University Press, 1998.

[Webber, 1988a] Bonnie L. Webber. Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistic (ACL–88)*, pages 113–122, State University of New York at Buffalo, NY, June 27-30 1988.

[Webber, 1988b] Bonnie L. Webber. Tense as discourse anaphor. *Computational Linguistics*, 14(2):61–72, June 1988.

[Webber, 1991] Bonnie L. Webber. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135, 1991.

[Webber, 1998] Bonnie L. Webber. Anchoring a lexicalized tree-adjoining grammar for discourse. In Manfred Stede, Leo Wanner, and Eduard H. Hovy, editors, *Proceedings of the COLING/ACL–98 Workshop on Discourse Relations and Discourse Markers*, pages 86–92, Montreal, Quebec, Canada, August 15th 1998.

[Webber et al., 1999] Bonnie L. Webber, Alistair Knott, Matthew Stone, and Aravind K. Joshi. Discourse relations: A structural and presuppositional account using lexicalized TAG. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL–99)*, pages 41–48, University of Maryland, College Park, MD, June 20–26 1999.

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

[Winograd, 1984] Peter N. Winograd. Strategic difficulties in summarizing texts. *Reading Research Quaterly*, 19(4):404–425, Summer 1984.

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

[Zadrozny and Jensen, 1991] Wlodek Zadrozny and Karen Jensen. Semantics of paragraphs. *Computational Linguistics*, 17(2):171–210, June 1991.

[Zukerman and McConachy, 1995] Ingrid Zukerman and Richard McConachy. Generating explanations across several user models: Maximizing belief while avoiding boredom and overload. In *Proceedings of the Fifth European Workshop on Natural Language Generation (ENLG–95)*, pages 143–161, Leiden, The Netherlands, May 20-22 1995.

Author Index

Subject and Notation Index

summarization, question answering, and information retrieval systems.

Daniel Marcu is a Research Scientist at the Information Sciences Institute at the University of Southern California and Research Assistant Professor in the university's Department of Computer Science.

A Bradford Book

OF RELATED INTEREST

FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING

CHRISTOPHER D. MANNING AND HINRICH SCHÜTZE

Statistical approaches to processing natural language text have become dominant in recent years. This foundational text is the first comprehensive introduction to statistical natural language processing (NLP) to appear. The book contains all the theory and algorithms needed for building NLP tools. It provides broad but rigorous coverage of mathematical and linguistic foundations, as well as detailed discussion of statistical methods, allowing students and researchers to construct their own implementations. The book covers collocation finding, word sense disambiguation, probabilistic parsing, information retrieval, and other applications.

ADVANCES IN AUTOMATIC TEXT SUMMARIZATION

EDITED BY INDERJEET MANI AND MARK T. MAYBURY

With the rapid growth of the World Wide Web and electronic information services, information is becoming available on-line at an incredible rate. One result is the oft-decried information overload. No one has time to read everything, yet we often have to make critical decisions based on what we are able to assimilate. The technology of automatic text summarization is becoming indispensable for dealing with this problem. Text summarization is the process of distilling the most important information from a source to produce an abridged version for a particular user or task. This book presents the key developments in the field in an integrated framework and suggests future research areas.

The MIT Press
Massachusetts Institute of Technology
Cambridge, Massachusetts 02142
http://mitpress.mit.edu

ISBN 0-262-13372-5
90000
9 780262 133722