The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts

by

Daniel Marcu

Department of Computer Science
University of Toronto
Toronto, Canada

December 1997

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

This thesis is an inquiry into the nature of the high-level, rhetorical structure of unrestricted natural language texts, computational means to enable its derivation, and two applications (in automatic summarization and natural language generation) that follow from the ability to build such structures automatically.

The thesis proposes a first-order formalization of the high-level, rhetorical structure of text. The formalization assumes that text can be sequenced into elementary units; that discourse relations hold between textual units of various sizes; that some textual units are more important to the writer's purpose than others; and that trees are a good approximation of the abstract structure of text. The formalization also introduces a linguistically motivated compositionality criterion, which is shown to hold for the text structures that are valid.

The thesis proposes, analyzes theoretically, and compares empirically four algorithms for determining the valid text structures of a sequence of units among which some rhetorical relations hold. Two algorithms apply model-theoretic techniques; the other two apply proof-theoretic techniques.

The formalization and the algorithms mentioned so far correspond to the theoretical facet of the thesis. An exploratory corpus analysis of cue phrases provides the means for applying the formalization to unrestricted natural language texts. A set of empirically motivated algorithms were designed in order to determine the elementary textual units of a text, to hypothesize rhetorical relations that hold among these units, and eventually, to derive the discourse structure of that text. The process that finds the discourse structure of unrestricted natural language texts is called rhetorical parsing.

The thesis explores two possible applications of the text theory that it proposes. The first application concerns a discourse-based summarization system, which is shown to significantly outperform both a baseline algorithm and a commercial system. An empirical psycholinguistic experiment not only provides an objective evaluation of the summarization system, but also confirms the adequacy of using the text theory proposed here in order to determine the most important units in a text. The second application concerns a set of text planning algorithms that can be used by natural language generation systems in order to construct text plans in the cases in which the high-level communicative goal is to map an entire knowledge pool into text.

Acknowledgements

As an undergraduate student, I one day came across Bill Woods's ACM paper on augmented transition networks. After I read it, I thought that finding computational means for understanding natural language is so cool that I had to try it by myself. This thesis is the written proof that I did try it...

The more I think about everything that happened between my reading of Bill Woods's paper and now, the more I believe that my ability to write this thesis is fundamentally rooted in a fortunate sequence of events, moral support from friends and family, and advice from scientists of extremely high caliber.

Most of all I have learnt from my advisor, Graeme Hirst. He taught me how to read and how to write; how to trust people and how to let them express and follow up on their own ideas; how to demolish and how to build arguments; how to talk and how to keep silent. But most of all, he taught me how to love language and to see in it more than a string of characters that is subject to immediate formalization and processing.

Hector Levesque and Ray Reiter shaped not only my formal training, but also my way of approaching science. They taught me logics and mathematics and they taught me to lean back on the chair and ask: "what's the scientific problem that you solve?". They also taught me that sometimes it is more important to focus on the shortcomings and weak parts of a theory than on the aspects where that theory proves to be successful.

Chrysanne DiMarco convinced me that The University of Toronto would be an excellent choice for pursuing my graduate studies and introduced me to Graeme. She has been helpful and supportive ever since.

Eduard Hovy motivated me with his limitless enthusiasm and readiness to give anyone around a good idea. He became a good friend who I could talk to about much more than science.

Derek Corneil taught me algorithms and graph theory with an ardor that is very difficult to match. His feedback was instrumental during every stage of the thesis.

Michael Cummings, Ron Smith, and Kathy McKeown, my external examiner, were all invaluable in providing comments and suggestions on this thesis, which ranged from linguistic and psycholinguistic to computational.

On the non-scientific side of my life, my parents never ceased to be my most faithful and devoted friends in spite of being thousands of miles away. They knew nothing about computer science, but what they taught me proved to me more essential and fundamental than all my University courses put together. There are no words to thank them for that.

Aside from my parents, teachers, and professors, I believe that those who I owe the most are my friends. Because of them, I consider myself to be one of the most fortunate persons in the world. And because they taught, influenced, and helped me so much (even when we disagreed completely), I dedicate this thesis to them. The order in which they are listed on the dedication page approximates to the best of my knowledge the order in which we became friends.

Contents

List of Tables

List of Figures

To the friends who have helped and influenced me the most,

Marin, Cornel, Vasile, Cuţu, More, Adi, Rareş, Vivi, Călin, Doina, Bilă, Ion, Juvete, Horace, Pelicanii, Ţicrea, Grir, Caşu, Almi, Băsă, E6, Reli, Ciupe, Brîndu, Oana, Monica, Cipi, Ed, Gelu, Melanie, Jin, Bil, Laura, Alex, Attila.

Chapter 1

Introduction

1.1 Motivation

Research in linguistics and computational linguistics has long pointed out that text is not just a simple sequence of clauses and sentences, but rather, a highly elaborate structure. Still, a formal theory of 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 (1.1) and (1.2) below, 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 without noticing any anomalies. Yet, only the sequence shown in (1.1) is coherent, i.e., is understandable text. The sequence shown in (1.2) does not make too much sense; consider just its first sentence: it is clear that we cannot start a text with an explicitly marked example.

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

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

(1.2) 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.

The fact that sequence (1.1) is coherent text, while sequence (1.2) is merely a collection of sentences, although each is exemplary when taken in isolation, suggests that extra-sentential factors play a major 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. With respect to text (1.1), such inferences will explain that "50 percent farther from the sun than Earth" is just some parenthetical information that is not central to the understanding of the whole text; that "Surface temperatures typically average about -60 degrees Celsius (-76 degrees Fahrenheit) at the equator and can dip to -123 degrees C near the poles" is just an elaboration of the fact that "Mars experiences frigid weather conditions"; and that it is "the low atmospheric pressure" that causes the liquid water to evaporate.

One possible way to represent these inferences explicitly is by means of a tree structure such as that shown in figure 1.1, where each leaf of the tree is associated with a contiguous textual span; the parenthetical units are enclosed within curly brackets; the internal nodes are labelled with the names of the rhetorical relations that hold between the textual spans

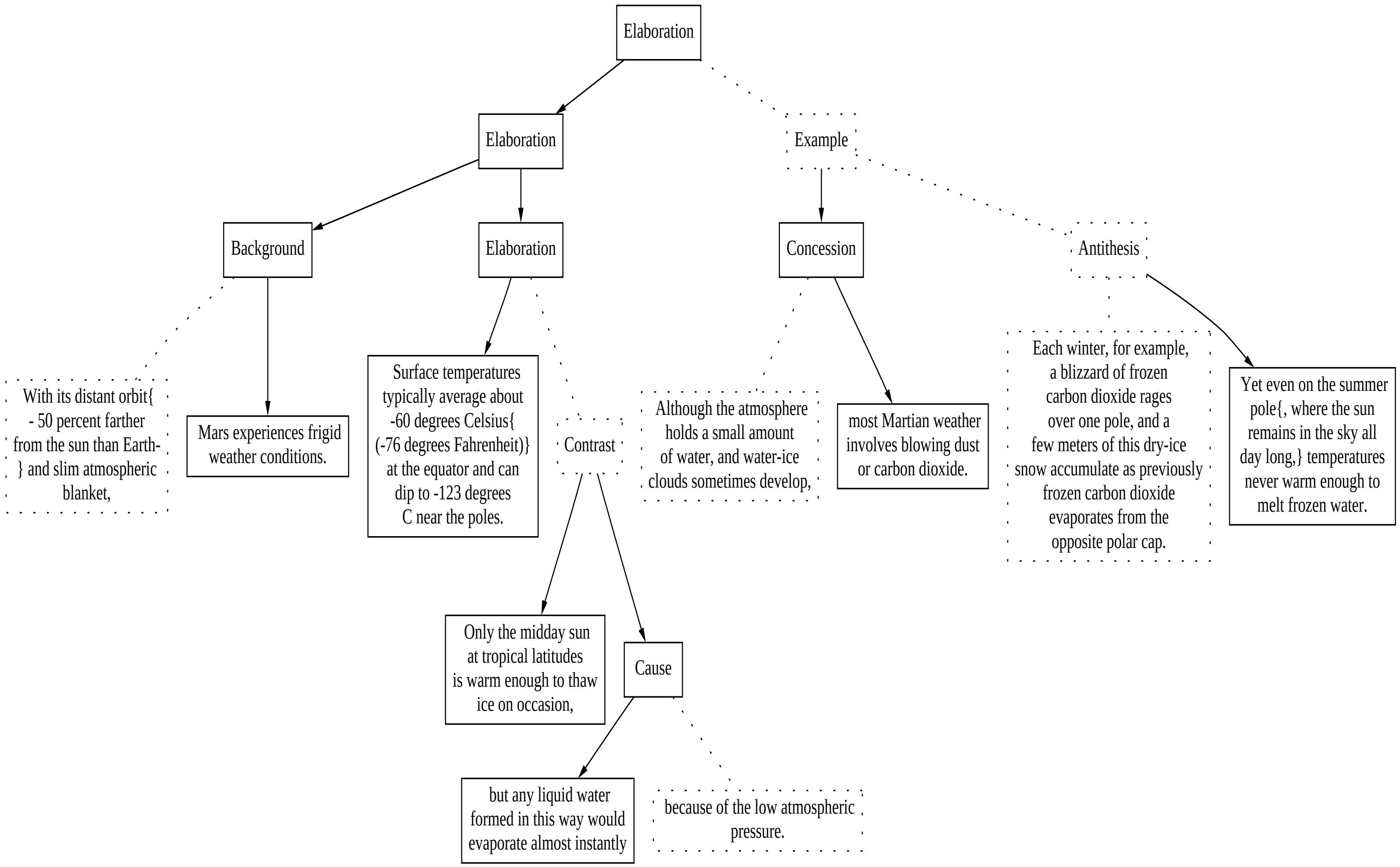

Figure 1.1: A tree-like structure that shows the rhetorical relations between the textual units of (1.1).

that are subsumed by their child nodes; and solid boxes and lines denote textual spans that are important to the writer's purpose. For example, the textual unit "most Martian weather involves blowing dust or carbon dioxide" is surrounded by a solid box and the unit "Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop," is surrounded by a dotted box, because the former represents something that is more essential to the writer's purpose than the latter.

During the continuous refinement of the text and discourse theories that have been proposed so far, it has become clear that an adequate 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 1.1? 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? Are these relations grounded in the events and the world that the text describes? Or are they grounded in 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 thesis is an attempt to answer some of these questions. More precisely, it is an inquiry into the formal properties of the high-level structure of unrestricted natural language text, the computational means that would enable its derivation, and two applications in automatic summarization and natural language generation that follow from the ability to automatically derive such structures.

1.2 Overview of the thesis

Previous discourse and text theories can be partitioned into two classes.

- In the first class, we find the theories developed in the traditional, truth-based semantic perspective on language [Kamp, 1981, Lascarides and Asher, 1991, Lascarides *et al.*, 1992, Lascarides and Oberlander, 1992, Lascarides and Asher, 1993, Asher, 1993, Kamp and Reyle, 1993, Asher and Lascarides, 1994, Kameyama, 1994, Gardent, 1994, Polanyi and van den Berg, 1996, van den Berg, 1996, Gardent, 1997, Schilder, 1997, Cristea and Webber, 1997]. These theories have a grammar as their backbone and rely on sophisticated logics of belief and default logics in order to intertwine and characterize the sentence- and discourse-based linguistic phenomena. Although these theories can be used to explain why "he" is a co-referent of "John" and "it" a co-referent of "donkey" in example (1.3) below, and to infer that "Max fell" because "John pushed him" in example (1.4), they are not tractable and cannot handle naturally occurring texts, such as that shown in (1.1).

(1.3) John has a donkey. He beats it.

(1.4) Max fell. John pushed him.

- In the second class, we find the theories that aim at characterizing the constraints that pertain to the structure of unrestricted texts and the computational mechanisms that would enable the derivation of these structures [van Dijk, 1972, Zock, 1985, Grosz and Sidner, 1986, Mann and Thompson, 1988, Polanyi, 1988, Hobbs, 1990, Polanyi, 1996]. Because these theories are either informal or incompletely specified, so far, they have been only manually applied to text analysis.

In this thesis, I explore the ground found at the intersection of these two lines of research. More specifically, I provide a theory and a fully specified formalization of text structures that is general enough to enable its applicability to unrestricted natural language texts, and yet simple enough to yield tractable, text-structure derivation algorithms.

The mathematics of text structures

In formalizing the structure of unrestricted texts (in chapter 2), I first distill the features that are common to previous approaches and show that most discourse theories acknowledge 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. However, as I will show, none of the present theories propose a clearly defined compositionality criterion, one that would spell out the conditions that have to be satisfied when two textual units are put together in a tree structure in order to create a larger unit, and would explain how the rhetorical relations that hold between large textual units relate to rhetorical relations that hold between elementary units. The lack of such a criterion not only prevents us from correctly classifying a given text structure as being valid or invalid, but also from deriving all the valid structures of a text. In sections 2.3.2 and 2.4, I use the theories proposed by Mann and Thompson [1988], Grosz and Sidner [1986], Hobbs [1990], and Polanyi [1988, 1996] in order to show that such a compositionality criterion is inherent primarily to the structure of discourse, rather than to the taxonomies of rhetorical relations that have been proposed by various researchers.

In section 2.5, I show that the difference between linguistic and nonlinguistic constructs that are more important to the writer's purpose (usually called *nuclei*) and constructs that are less important (usually called *satellites*) can constitute the foundation of a compositionality criterion of valid text structures. This criterion (proposition 2.1) specifies that if a relation holds between two nodes of the tree structure of a text, that relation also holds between some linguistic and nonlinguistic constructs that pertain to the most important constituents of those nodes. In spite of its large range applicability, the formalization of this criterion proves to be beyond the current state of the art in computational linguistics and artificial intelligence. Hence, I propose instead a stronger criterion, one that is easily formalized. The strong compositionality criterion (proposition 2.2) stipulates that if a relation holds between two textual spans of the tree structure of a text, that relation also holds between the most important units of the constituent spans. Hence, the strong compositionality criterion leaves implicit the nonlinguistic constructs that characterize the weak criterion and focuses only on textual units as the linguistic entities of interest.

In section 2.6, I formalize the strong compositionality criterion and the features listed at the beginning of this section in the language of first-order logic. The resulting formalization is general with respect to the taxonomy of rhetorical relations that it can rely upon; as an example, I show how one can obtain, as a by-product, a formalization of Rhetorical Structure Theory (RST) [Mann and Thompson, 1988].

Using the formalization of RST that I propose in section 2.6 and Moser and Moore's [1996] discussion of the relationship between RST and Grosz and Sidner's intention-based discourse theory [1986], I propose a formal account of both theories (see section 2.7). The melding of structure- and intention-based constraints enables the derivation of intentional inferences on the basis of the structure of text and provides a means for using intentional judgments for reducing the ambiguity of text structures.

The automatic derivation of text structures: an algorithmic perspective

The formalization proposed in chapter 2 focuses only on the mathematical properties of text structures, but says nothing about any algorithms that can be used to derive them. In chapter 3, I explore the problem of text structure derivation (see definition 2.2) from an algorithmic perspective. More precisely, I investigate how, given a sequence of elementary units and a set of rhetorical relations that hold among these units, one can derive all the valid text structures of the sequence.

I study theoretically and compare empirically four paradigms that solve the problem of text structure derivation. I show how the problem of text structure derivation can be encoded as

- a classical constraint-satisfaction problem (section 3.2);
- a propositional satisfiability problem (section 3.3);
- a theorem-proving problem (section 3.4);
- a parsing problem using a grammar in Chomsky normal form (section 3.5).

The four paradigms yield sound and complete algorithms for deriving the structure of text.

In contrast with previous approaches to discourse analysis, the algorithms that I propose in chapter 3 no longer assimilate the task of discourse processing with an incremental process in which discourse units are sequentially examined and added to a continuously updated discourse tree. Rather, the algorithms assume that the elementary textual units and the relations between them can be determined beforehand. As a consequence, the algorithms that I propose no longer need the notion of "right frontier", which is pervasive in incremental approaches to discourse analysis, and no longer have to deal with nonmonotonicity, which occurs when some decisions made during the incremental processing of discourse need to be "undone" at a later stage.

A corpus analysis of cue phrases

The algorithms presented in chapter 3 provide a computational solution to the problem of text structure derivation. However, this problem takes as its input the sequence of elementary units that make up a text and the rhetorical relations that hold among them. If any of the algorithms discussed in chapter 3 is to be applicable on real texts, we need to also automate the process of determining the elementary units of a text and the rhetorical relations that hold among them.

In chapter 4, I discuss a set of linguistic devices that can be exploited to provide solutions to both problems. For the rest of the thesis, I choose to explore how well we can solve the problem of text structure derivation by relying mostly on the discourse function of cue phrases, i.e., words such as *however, although,* and *but*, and by applying only shallow techniques that do not require syntactic and semantic analysis of the text.

The main assumption behind the use of cue phrases is that they are an accurate-enough indicator of the boundaries between elementary textual units and of the rhetorical relations that hold between them. In section 4.3, I discuss in detail how the ambiguity of cue phrases is managed by the formalization presented in chapter 2.

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 in order to determine the elementary textual units that are found in their vicinity and to hypothesize rhetorical relations between these units. In order to overcome this lack of data, I designed an exploratory, empirical study of my own (section 4.4). I used previously published lists of cue phrases [Halliday and Hasan, 1976, Grosz and Sidner, 1986, Martin, 1992, Hirschberg and Litman, 1993, Knott, 1995, Fraser, 1996] and created a set of 460. For each cue phrase in the list, I extracted from the Brown Corpus a number of text fragments that contained that cue phrase. Overall, I selected more than 7600 text fragments. I manually analyzed 2100 of these texts and, on the basis of the data in the corpus and the intuitions that I developed during the analysis, I associated with each cue phrase information that enables

- its automatic 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.

Chapter 4 discusses in detail the materials and methods of the corpus analysis and provides some general results. In chapters 5 and 7, I subsequently establish the connection between the corpus analysis and the algorithms that derive text structures for unrestricted texts in the context of discourse analysis, and build valid text plans in the context of natural language generation.

The rhetorical parsing of unrestricted natural language texts

The text theory developed in chapter 2, the algorithms developed in chapter 3, and the corpus analysis presented in chapter 4 provide the foundations for a rhetorical parsing algorithm, which is presented in chapter 5. The rhetorical parsing algorithm takes as input natural language text and returns the discourse structure of that text.

In chapter 5, I first discuss the advantages and disadvantages that would result from adopting the position that there exists some correlation between the structure of text and the sentence, paragraph, and section boundaries that are used by writers. The rhetorical parsing algorithm assumes that such a correlation exists, i.e., it assumes that clauses, sentences, paragraphs, and sections provide an underspecified representation of the structure of text. Exploiting this structure improves the computational properties of the rhetorical parsing algorithm.

The rhetorical parsing algorithm first determines the set of all cue phrases that occur in the text that is given as input. In the second step, the rhetorical parser uses information derived from the corpus analysis in order to determine the elementary units of the text and the cue phrases that have a discourse function. Section 5.3 discusses in detail an algorithm that identifies discourse markers and clause-like unit boundaries using only surface-based methods and evaluates the algorithm against three texts. The texts total more than 7000 words and belong to three different genres.

Once the elementary units have been identified, the rhetorical parser uses again information derived from the corpus in order to make disjunctive hypotheses with respect to the rhetorical relations that hold between different units. Section 5.4 presents two algorithms that are used to hypothesize discourse relations: one of them is based on coherence, while the other is based on cohesion. The coherence-based algorithm is rooted in the corpus analysis of cue phrases. The cohesion-based hypothesizes rhetorical relations by measuring the degree of overlap between the words that are used by two textual units.

The algorithms developed in chapter 3 assumed that the rhetorical relations that hold between elementary units were precisely known. However, as we have seen, the rhetorical parser makes merely disjunctive hypotheses. In order to deal with this issue, I consider, in section 5.5, a disjunctive formulation of the problem of text structure derivation. That is, I consider the problem of text structure derivation to be the following: given a sequence of textual units and a set of disjunctive rhetorical relations that hold among these units, find all valid text structures of the sequence. In section 5.5, I discuss how the most efficient algorithms that were developed in chapter 3 can be modified such that they can handle disjunctive hypotheses as well. More precisely, I develop a proof-theoretic approach for the disjunctive case and I show how disjunctive hypotheses can be compiled into a parsing problem with a grammar in Chomsky normal form.

In section 5.5, I discuss how these approaches can be implemented and integrated with the rhetorical parser. I end the chapter with a discussion of ambiguity in discourse processing and a proposal on how one can deal with it.

All the algorithms that pertain to the rhetorical parser have been fully implemented. When the rhetorical parser takes text (1.1) as input, it produces a text structure similar to that shown in figure 1.1.

The summarization of natural language texts

Researchers in computational linguistics [Mann and Thompson, 1988, Matthiessen and Thompson, 1988, Sparck Jones, 1993b] have long hypothesized that discourse structures can be used in natural language summarization. That is, they have suggested that there is a correlation between the textual units that are assigned a nuclear status in a text structure and what readers perceive as being important in the corresponding text. However, to date, no empirical experiment has tested the validity of this hypothesis.

In chapter 6, I describe such an experiment, which shows that, indeed, text structures *can* be used effectively in order to select the most important units in a text. In addition, the experiment provides a clear insight into the nature of the discourse-based summarization problem, because it uncovers both its strengths and limitations, independent of any particular implementation.

This result leads me to propose a discourse-based summarization algorithm: the algorithm takes as input a natural language text and a number p between 1 and 100, which corresponds to the percentage of important units that the algorithm is to select from the given text. The discourse-based summarizer uses the rhetorical parsing algorithm in order to derive 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 (see section 6.2). The $p\%$ units with highest score provide a summary of the text. An evaluation of the discourse-based summarization program has shown that it significantly outperforms both a baseline algorithm and Microsoft's Office97 summarizer.

From local to global coherence: A bottom-up approach to text planning

In chapter 7, I explore an application of the formalization of text structures in the area of text planning. Traditionally, flexible approaches to text planning assimilated the problem of text-plan derivation with a top-down, hierarchical expansion process. In section 7.1, I show that in spite of their adequacy in goal-driven settings, top-down planning techniques are not appropriate when the high-level communicative goal boils down to "tell everything that is in this knowledge base" or "tell everything that is in this chosen subset". The solution that I propose to this problem is bottom-up.

The intuition behind the bottom-up, text-planning algorithms, which I present in section 7.4, is that global coherence can be achieved by satisfying as many as possible of the local coherence constraints on ordering and adjacency. The corpus analysis discussed in chapter 4 provides evidence that different rhetorical relations are characterized by different preferences with respect to the order in which they realize their satellites and nuclei and with respect to their tendency of clustering their satellites and nuclei into larger textual spans. Besides providing a solution to the text planning problem in the cases in which the high-level communicative goal is "tell everything that is in this knowledge base", the bottom-up approach also enables a simple solution to the problem of generating text plans that satisfy multiple communicative goals.

The bottom-up text planning algorithms were incorporated into HealthDoc [DiMarco *et al.*, 1997, Hirst *et al.*, 1997], a natural language system that generates texts that are tailored to particular audiences.

Conclusions

In the last chapter, I critically review the main contributions of the thesis and point to future research directions.

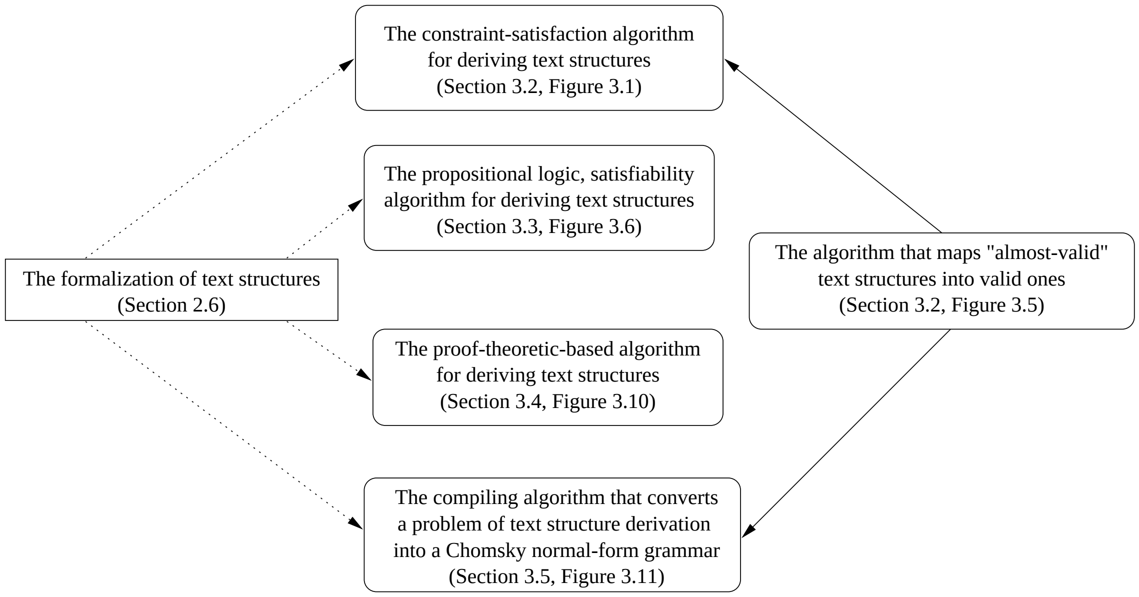

Figure 1.2: The algorithms that find a solution to the problem of text structure derivation that is given in definition 2.2.

1.3 Maps of the thesis

General remarks on the layout of the thesis

In the previous section, I presented a chapter by chapter overview of the main topics that I address in this thesis. As we saw, the thesis dwells on topics that range from formal, knowledge representation issues in text theory to issues in algorithms, linguistics, psycholinguistics, and language engineering. Because of its diversity, I found it inappropriate to cluster the discussion of the literature in a single chapter. Instead, I preferred to discuss the relevant research in connection with each particular topic. I hope that this will enable the reader who is interested in only a particular aspect of the thesis to find her way around easier. For the same reason, I have included a short summary at the end of each chapter.

A map of the algorithms in the thesis

Throughout the thesis, I present a number of algorithms: between some of them exist some obvious connections. Figures 1.2 and 1.3 make explicit the connections between the most important ones. The first class of algorithms, that presented in figure 1.2, concerns the theoretical facet of the problem of text structure derivation. The dotted arrows denote that the algorithms referred to by nodes surrounded by rounded boxes rely upon the formalization of text structures presented in section 2.6. The solid arrows denote “uses” relations: the destination of an arrow corresponds to an algorithm that uses the algorithm from which the arrow originates.

The second class of algorithms concerns natural language applications. As figure 1.3 shows, the rhetorical parser relies upon six algorithms and constitutes the basis of the discourse summarizer. Some of the algorithms that are used by the rhetorical parser and the text planning algorithms rely heavily on the exploratory analysis of cue phrases that is discussed in chapter 4.

A rhetorical map of the thesis

In order to facilitate better navigation through the thesis, I also provide a rhetorical map of it (see figure 1.4) in the style of the text structure diagram shown in figure 1.1. A reader without background in discourse theories will probably have a much better understanding of the meaning of the rhetorical map shown in figure 1.4 after reading chapter 2.

In figure 1.4, the leaves of the tree-like map correspond to the chapters of the thesis. Internal nodes correspond to the relations between the spans of the thesis that are subsumed by the immediate children. Solid lines and boxes correspond to the most important parts, the nuclei of the representation. Dotted lines and boxes correspond to the satellites. Hence, in chapter 1 I "motivate" the work presented in chapters 2 to 7. The formalization of text structures discussed in chapter 2 is provided an "algorithmic solution" in chapter 3. The corpus analysis in chapter 4 "enables" the development of the rhetorical parser in chapter 5. An immediate "application" of the rhetorical parser is the discourse-based summarization program that is presented in chapter 6. In fact, both the rhetorical parser and the text planning algorithms presented in chapter 7 can be "jointly" seen as "applications" of the formalization of text structures presented in chapter 2. Chapter 8 "summarizes" the results presented in the whole thesis.

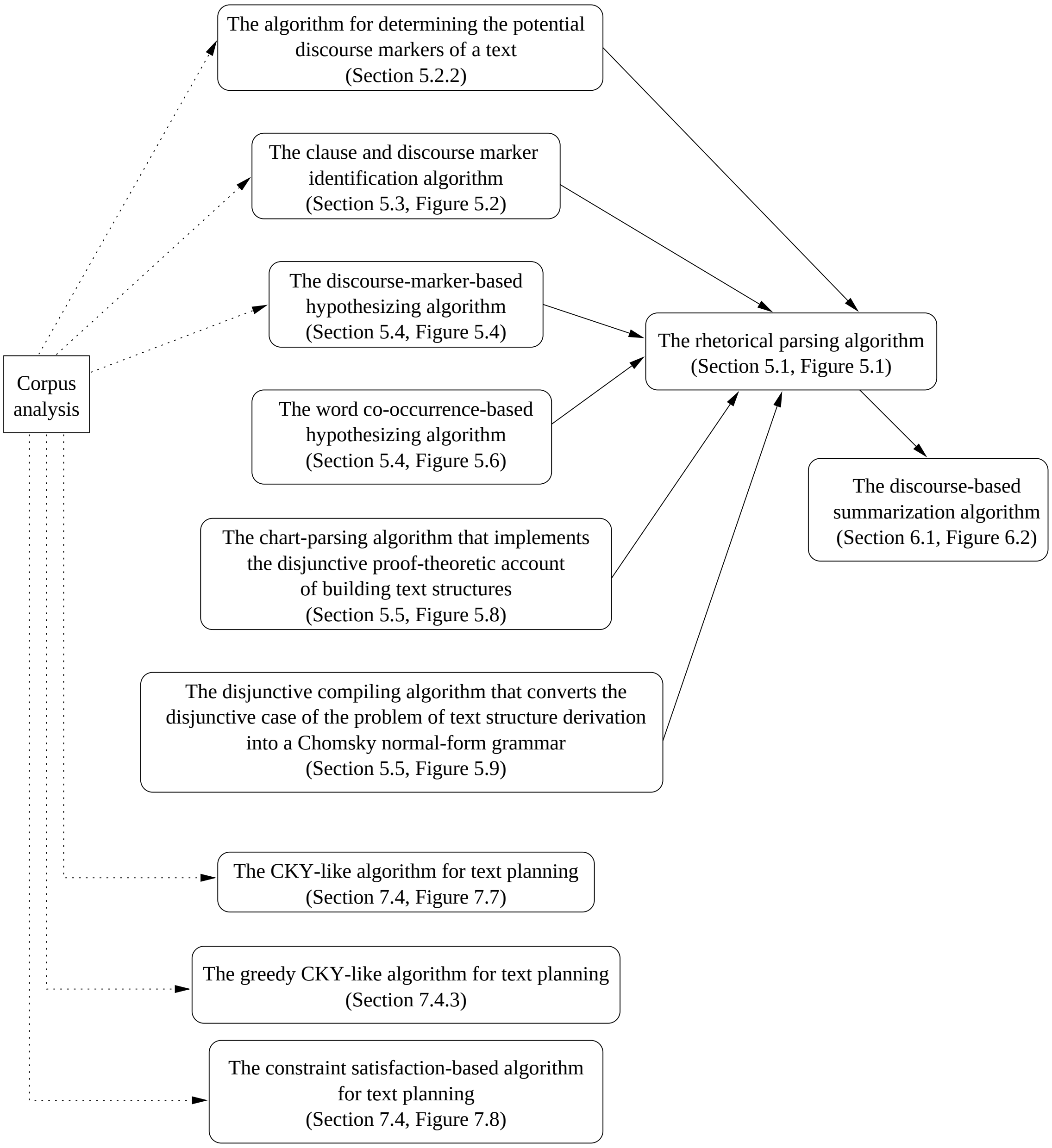

Figure 1.3: Algorithms that concern applications of the formalization of text structures in rhetorical parsing, summarization, and text planning.

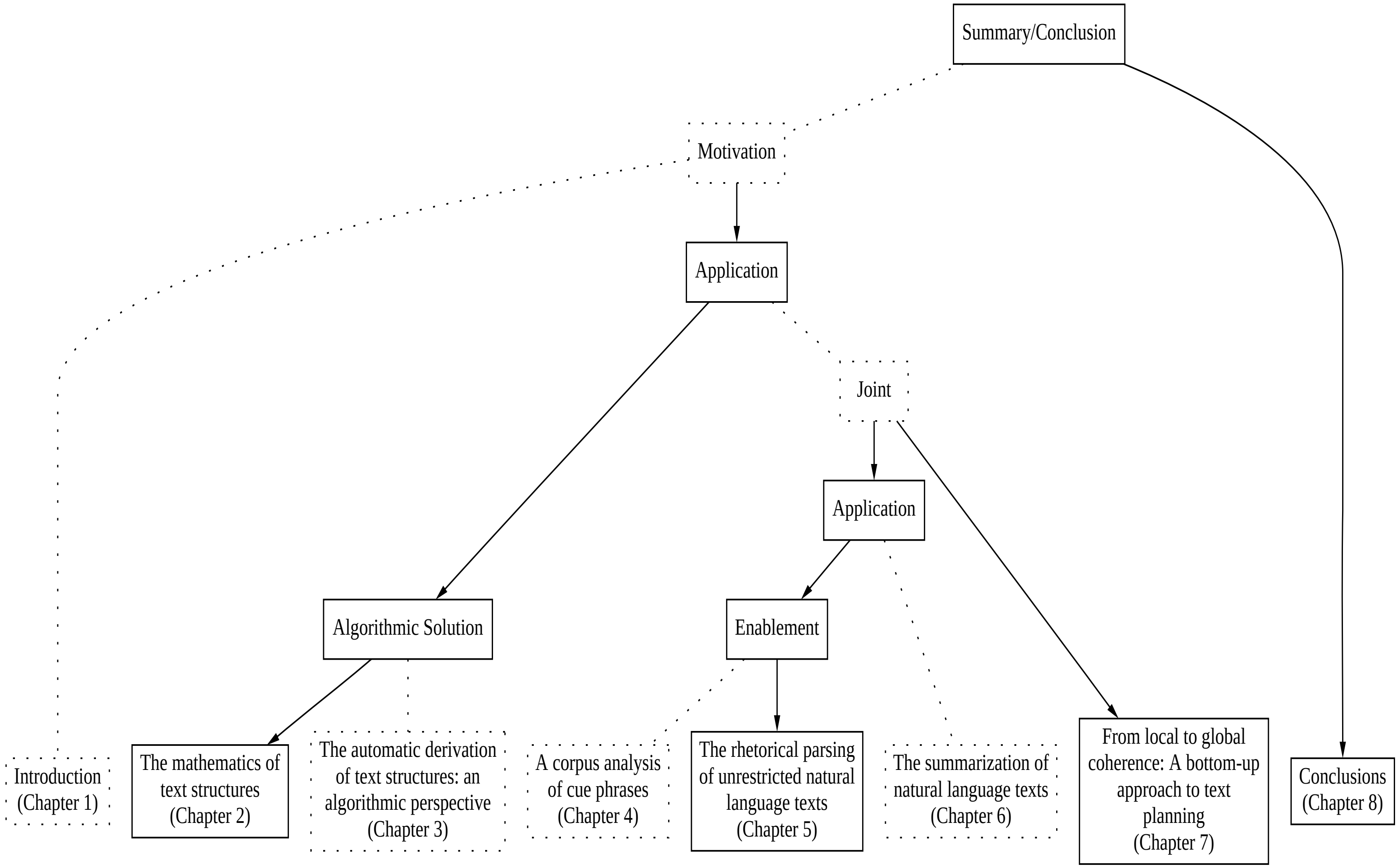

Figure 1.4: A rhetorical map of the thesis

Chapter 2

The mathematics of text structures

2.1 Preamble

As I have mentioned in the introduction, one of the goals of this thesis is to provide a theory of text structures that is general enough to enable its applicability to unrestricted natural language texts, and simple enough to yield tractable, text-derivation algorithms. In this chapter, I first discuss the essential features of discourse structures that have been proposed by previous researchers. I show that none of the current discourse theories provides a compositionality criterion that would explain how rhetorical relations that hold between large spans relate to rhetorical relations that hold between small spans. I provide such a criterion and a first-order formalization of the constraints that characterize the valid structures of text. I end the chapter by showing how the formalization can be extended to handle both structural and intentional constraints.

2.2 A formalization of text structures from first principles

2.2.1 The essential features of text structures

If we examine carefully the claims that current theories make with respect to the *structure* of text and discourse, we will find significant commonalities. Essentially, all these theories acknowledge that the elementary textual units are non-overlapping spans of text; that there exist rhetorical, coherence, and cohesive relations 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-like structure. 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], 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 *clause-like* spans to be the elementary units of text, the text fragment in (2.1) can be broken into 6 units, as shown below. The elementary units are delimited by square brackets.[1]

(2.1) [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]]

Rhetorical, coherence, and cohesive relations hold between textual units of various sizes. The nature, number, and taxonomy of the relations that hold between textual units continue to be controversial issues. At one end of a spectrum of influential proposals, we have the ground-breaking 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, 1997] for an overview). Although unprincipled, these approaches provided the first "complete" taxonomy of the 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 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]; from the linguistic evidence (such as cue phrases) of some linguistic psychological constructs that are used during text processing [Knott, 1995]; 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 presupposition that rhetorical, coherence, and cohesive relations *need* to be considered if one is to account for the meaning of text.

For example, we can say that a rhetorical relation of ELABORATION holds between units

[1]See pages 125 and 133 for a discussion of the difference between clauses and clause-like units.

1 and 2 in text (2.1), because unit 2 provides some extra information with respect to unit 1. And we can say that a rhetorical relation of BACKGROUND or JUSTIFICATION holds between the span that ranges over units 1 to 3, [1,3], and unit 4, because the information given in span [1,3] merely sets the stage for presenting the information in 4.

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 distiction 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 same distinction permeates the dominance relations that hold between the intentions associated with discourse segments in Grosz and Sidner's theory [1986] and is central to Mann and Thompson's theory [1988], in which the units between which a rhetorical relation holds are explicitly labelled as *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, 1997] reflect the same difference in the relative importance of the units that are members of these structures.

For example, units 5 and 6 in text (2.1) 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 JOINT holds between units 5 and 6. In contrast, if we reconsider span [1,3] and unit 4, we easily notice that unit 4 expresses what is most essential for the writer's purpose: the role that units 1–3 play is subsidiary to the role played by unit 4. Hence, we can say that a hypotactic relation of JUSTIFICATION or BACKGROUND holds between span [1,3] and unit 4.

The abstract structure of most texts is a tree-like structure. 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, Asher, 1993, Lascarides and Asher, 1993, Polanyi, 1996, Moser and Moore, 1996, Walker, 1997]. For example, a possible tree-like representation of the discourse structure that pertains to units 1–6 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.

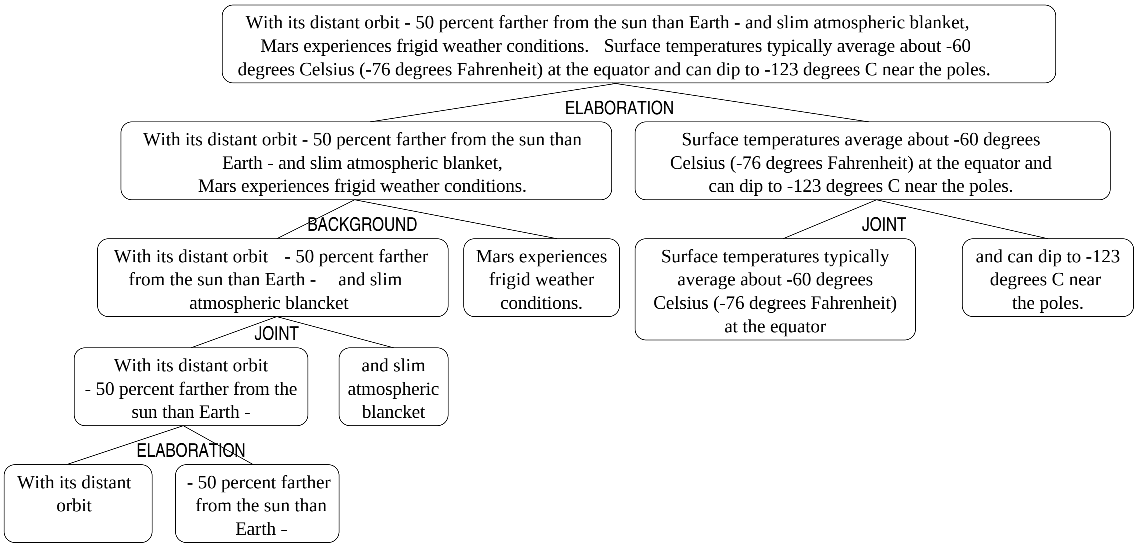

Figure 2.1: An example of a tree-like discourse structure that corresponds to text (2.1).

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 some criticism: it seems that certain classes of texts, such as argumentative texts [Toulmin *et al.*, 1979, Birnbaum *et al.*, 1980, Birnbaum, 1982] 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, the empirical experiments that I will describe in chapter 4 show that trees are an adequate representation in the majority of the cases. (In fact, Cohen [1983, 1987] shows that even arguments can be modelled 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 thesis.

2.2.2 The problem of formalizing text structures

The four features that I discuss in section 2.2.1 constitute the foundations of my formalization. In other words, I take as axiomatic that any text can be partitioned into a sequence of non-overlapping, elementary textual units and that a text structure, i.e., a tree, can be associated with the text such that:

- There exists a bijection between the leaves of the tree and the elementary textual units;
- 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. Had such constraints not been 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 taxonomy of relations that it relies upon. The only assumption behind the formalization is that such a taxonomy exists and that some relations in this taxonomy are paratactic, while others are hypotactic.

Presenting the formalization only in abstract terms will make the reading difficult. To avoid this, I will mainly use in my examples the taxonomy 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 relations*. However, the reader should understand that I take *rhetorical relation* to be just a general term that subsumes all the other kinds of relations that a text theory might need, such as coherence, argumentative, and cohesion relations. For the uninitiated reader, I first provide a short introduction to Mann and Thompson's theory and taxonomy of relations.

2.3 Rhetorical Structure Theory

2.3.1 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, 1988b, Scott and de Souza, 1990, Moore and Swartout, 1991, 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, 1990b, Rösner and Stede, 1992, Maier, 1993, Hovy and Maier, 1997], the problematic mapping between rhetorical relations and speech acts [Hovy, 1990b] 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 multi-nuclear. 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;

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:[B1]] [we know that 3,000 teens start smoking each day.[C1]]

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

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

According to Mann and Thompson [1988], a canonical analysis of a text is a set of

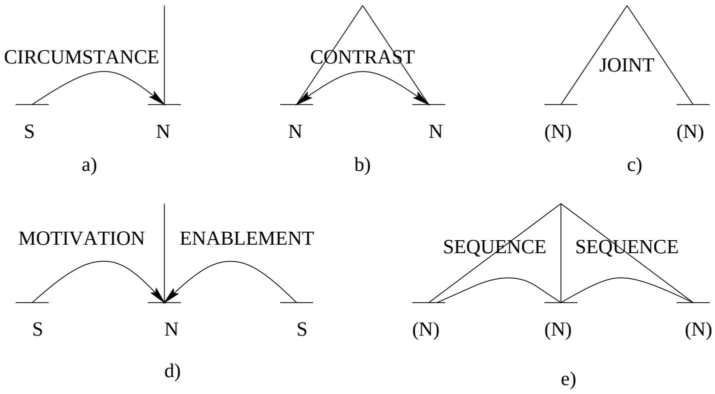

Figure 2.3: Examples of the five types of schema that are used in RST [Mann and Thompson, 1988, p. 247]. The arrows link the satellite to the nucleus of a rhetorical relation. Arrows are labeled with the name of the rhetorical relation that holds between the units over which the relation spans. The horizontal 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.

---

schema applications for which the following constraints hold:

(2.2)

- **Completeness:** One schema application (the root) spans the entire text.
- **Connectedness:** Except for the root, each text span in the analysis is either a minimal unit or a constituent of another schema application of the analysis.
- **Uniqueness:** Each schema application involves a different set of text spans.
- **Adjacency:** The text spans of each schema application constitute one contiguous text span.

Obviously, the formulation of the constraints that Mann and Thompson put on the discourse structure (2.2) 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 acknowledgement 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 pieces of text, RST is fully compatible with the essential features of text structures that I discussed in section 2.2.1.

2.3.2 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 the following text (in which each textual unit is labelled for reference):

(2.3) [No matter how much one wants to stay a non-smoker,$^{\mathrm{A}_1}$] [the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life.$^{\mathrm{B}_1}$] [We know that 3,000 teens start smoking each day,$^{\mathrm{C}_1}$] [although it is a fact that 90% of them once thought that smoking was something that they'd never do.$^{\mathrm{D}_1}$]

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.

$$(2.4) \qquad RR = \begin{cases} rhet\_rel(\textsc{justification}, \mathrm{A}_1, \mathrm{B}_1) \\ rhet\_rel(\textsc{justification}, \mathrm{D}_1, \mathrm{B}_1) \\ rhet\_rel(\textsc{evidence}, \mathrm{C}_1, \mathrm{B}_1) \\ rhet\_rel(\textsc{concession}, \mathrm{D}_1, \mathrm{C}_1) \\ rhet\_rel(\textsc{restatement}, \mathrm{D}_1, \mathrm{A}_1) \end{cases}$$

These relations hold because the understanding of both A_1 and D_1 will increase the reader's readiness to accept the writer's right to present B_1; the understanding of C_1 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 thesis, I use the convention that rhetorical relations are represented as sorted, first-order predicates having the form $rhet\_rel(name, satellite, nucleus)$ where $name$, $satellite$, and $nucleus$ represent the name, satellite, and nucleus of a rhetorical relation, respectively. Multinuclear relations are represented as predicates having the form $rhet\_rel(name, nucleus_1, nucleus_2)$.

Assume now that one is given the task of building an RS-tree for text (2.3) and that one produces the candidates in figure 2.4. Any student in RST would notice from the beginning

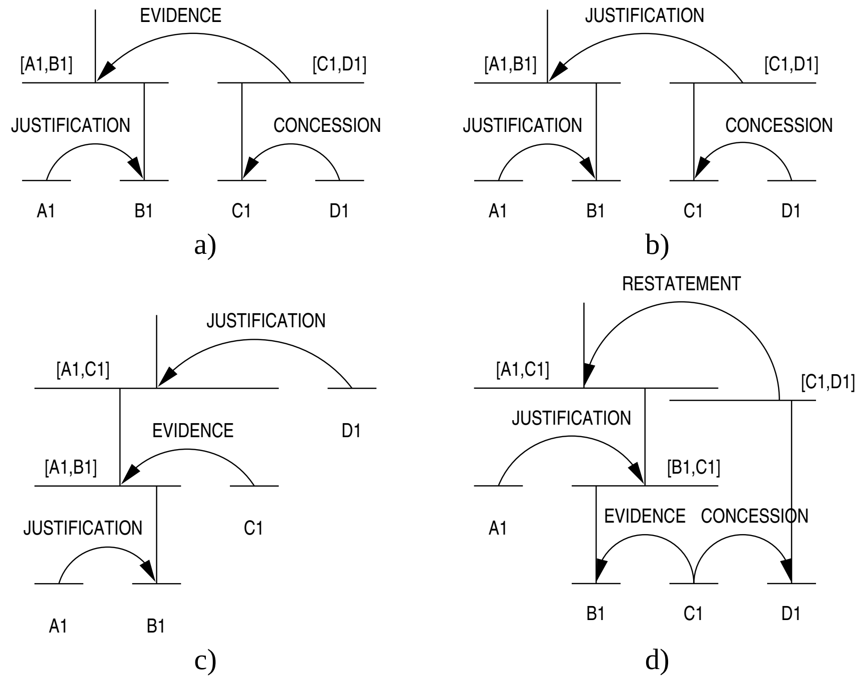

Figure 2.4: A set of possible rhetorical analyses of text (2.3).

that the tree in figure 2.4.d 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 2.4.a–c represent *all* the possible ways in which a rhetorical structure could be assigned to text (2.3), 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.

I believe that the explanation for the current lack of algorithms capable of automatically building the RS-trees that pertain to a given discourse can be found not only in the ambiguous definition of the rhetorical relations but also in the incomplete description of RS-trees that is provided in the original theory. A careful analysis of the constraints provided by Mann and Thompson [1988, p. 248] shows that their specification for RS-trees is not complete 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 (2.3) and the set of rhetorical relations that pertain to the minimal units (2.4), 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 RS-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 she considers the

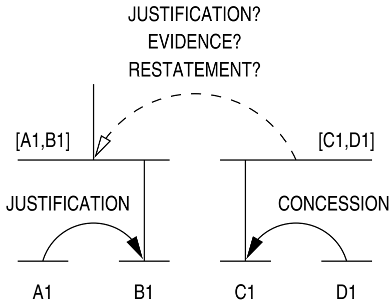

Figure 2.5: An example of the ambiguity that pertains to the construction of RS-trees.

elementary relations (2.4) that hold across the two spans, she has three choices, which correspond to the relations $rhet\_rel(\text{JUSTIFICATION}, D_1, B_1)$, $rhet\_rel(\text{EVIDENCE}, C_1, B_1)$, and $rhet\_rel(\text{RESTATEMENT}, D_1, A_1)$. Which is the correct one to choose?

More generally, suppose that the analyst has already built two partial RS-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 RS-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 RS-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? Mann and Thompson [1987, 1988] provide no precise answer for these questions.

2.4 Compositionality in other discourse theories

The lack of a compositionality criterion of the kind mentioned in the previous section is not specific only to RST, but rather to the majority of discourse theories. In what follows, I discuss a few.

2.4.1 Compositionality in Grosz and Sidner's theory

Grosz and Sidner's Theory (GST) [1986] proposes a discourse structure that is also compatible with the essential features discussed in section 2.2.1. In GST, the elementary textual units are called *discourse segments* (DS) and the discourse structure is explicitly stated to be a tree. Each discourse segment is characterized by a primary intention, which is called the *discourse segment purpose* (*DSP*). GST identifies only two kinds of intention-based relations that hold between two discourse segments: *dominance*, and *satisfaction precedence*.

When the text of a discourse segment DS_1 satisfies the discourse segment purpose DSP_1 and provides part of the satisfaction of a discourse segment DS_2 that includes DS_1, it is said that there exists a dominance relation between DS_2 and DS_1, i.e., DS_2 *dominates* DS_1. If the satisfaction of DSP_2 is conditioned by the satisfaction of DSP_1, it is said that DSP_1 *satisfaction-precedes* DSP_2.

Reconsider now text (2.3) from the perspective of GST. In order to build the discourse structure for this text, we need to have a clear criterion for determining the discourse segment boundaries and we also need a clear procedure for determining the primary intentions that pertain to each of these segments. GST provides no unambiguous solutions for any of these problems [Grosz and Hirschberg, 1992, Passonneau and Litman, 1993, Hirschberg and Nakatani, 1996, Passonneau and Litman, 1997a], but for the sake of the argument, let us assume that it does. An informal analysis of text (2.3) could produce at least three discourse segments:

1. The first segment, DS_1, contains units $\text{A}_1 - \text{B}_1$ and its primary intention is (*Intend writer* (*Believe reader* B_1)), i.e., the writer intends to make the reader believe that the pressure to smoke in junior high is greater than it will be any other time of one's life.

2. The second segment, DS_2, contains unit C_1 and its primary intention is (*Intend writer* (*Believe reader* C_1)), i.e., the writer intends to make the reader believe that 3000 teens start smoking each day.

3. The third segment, DS_3, contains unit D_1 and its primary intention is (*Intend writer* (*Believe reader* D_1)), i.e., the writer intends to make the reader believe that 90% of the teens once thought that smoking was something that they'd never do.

In order to build the discourse structure of this text, we would need now to consider larger segments. A reasonable candidate is the segment that dominates segments DS_2 and DS_3 — let us call this segment DS_{23}. The problem that we have when we create this segment is isomorphic with the problem that we had when we tried to put text spans together in RST because it is not clear what the primary intention of segment DS_{23} should be. One choice is to take this intention to be that associated with segment DS_2. Another choice is to take it to be that associated with segment DS_3. And an equally valid choice is to take the intention to be that the writer intends to make the reader aware of the contrast between the teens' behavior (3000 of them start smoking each day) and the beliefs that they held when they were younger (90% of them once thought that smoking was something that they'd never do). As in the case of RST, where we did not know how the rhetorical relations that pertain to large text spans are related to those between the subordinated spans, in GST we do not know how the primary intentions of large discourse segments are related to those of the subordinated segments.

2.4.2 Compositionality in Hobbs's theory

Hobbs's theory [1990, 1995] is part of a larger theory that attempts to make explicit the relation between the interpretation of text, events in the real world, and the knowledge and beliefs of the speaker and hearer. The main difference between Hobbs's theory and the discourse theories proposed by others is in the nature of the taxonomy of coherence relations. According to Hobbs, a discourse is coherent when it talks about coherent events in the world; when it reflects some rational structure of goals; when it relates discourse segments to the reader's prior knowledge; or when it helps the reader derive inferential relations between discourse segments, thus enabling her to create a high-level structure of text.

Hobbs's theory is consistent with the essential features of discourse that I discussed in section 2.2: elementary units are contiguous spans of text, coherence relations are hypotactic and paratactic, and discourse structures are trees.[2] However, as in Mann and Thompson's and Grosz and Sidner's theories, Hobbs does not provide a compositionality criterion for the discourse structures of texts. The algorithm that he proposes for analyzing discourse is a top-down one. In the first step, a human analyst is supposed to identify *intuitively* one or two major breaks in the text and then apply the same process recursively, on the resulting subtexts, until a tree-like structure is obtained. It is only then that the analyst proceeds in a bottom-up fashion with labelling the nonterminal nodes with coherence relations and with making explicit the knowledge and beliefs that support the assignment of coherence relations to nodes. Obviously, the intuitive nature of Hobbs's algorithm does not answer the compositionality-related questions that we raised in connection with RST and GST.

In spite of this, Hobbs is closer than Grosz and Sidner and Mann and Thompson to providing a compositionality criterion for discourse structures, as he explicitly acknowledges the need for it:

> If the definitions of the coherence relations are to be applied to segments of discourse larger than a single clause, we need to be able to say what is asserted by those segments. We can do so if, in the composition process, when two segments S_0 and S_1 are joined by a coherence relation into a larger segment S, we have a way of assigning an assertion to S in terms of the assertions of S_0 and S_1. The assertion of S will constitute a kind of summary of the segment S. [Hobbs, 1990, p. 104]

Although Hobbs discusses how the assertion of S might be constructed depending on the nature of the relation that holds between segments S_0 and S_1, he does not discuss the

[2] An in-order traversal of the leaves of the discourse trees built by Hobbs yields, in some cases, a sequence of units that differs from that of the original text (see for example the tree in figure 6.1 in [Hobbs, 1990, p. 117]). In contrast, an in-order traversal of the leaves of the discourse trees built in RST and GST always yields a sequence of units that reflects the original text.

relationship between coherence relations that hold between elementary textual units and coherence relations that hold between larger textual spans.

2.4.3 Compositionality in Polanyi's theory

Polanyi's theory [1988, 1996] (PT) is also compatible with the essential features of discourse that were discussed in section 2.2: Polanyi explicitly mentions that discourse structures are trees; that the elementary units are sentences (or discourse constituent units); and therefore, that the elementary units are non-overlapping pieces of text. Although Polanyi rejects the approaches to discourse that rely on coherence relations, the valid structures of her discourse parse trees can be interpreted as a direct expression of such relations: the coordination, subordination, and binary structures are nothing but the structural consequence of the relations that hold between the constituent units.

One of the main interests of Polanyi is to explain how the incremental processing of discourse constituent units yields a discourse parse tree. To do this, Polanyi assumes that each discourse constituent unit "comes with" a context frame that encodes all the information that might be needed during the parsing process. The information in these frames is used to determine unambiguously the node on the right frontier of the partial discourse tree to which the discourse unit will be attached, and also, the type of attachment. In addition, Polanyi assumes that the attachment process modifies the frame of the immediate mother node so that the mother node will reflect the extra information that has been added to the overall structure. The existence of such an oracle, which determines unambiguously the attachment nodes and the information that is inherited by the immediate mother nodes whenever such an attachment occurs, obviates a compositionality principle.

2.5 The formulation of a compositionality criterion of valid text structures

2.5.1 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 theories that I have discussed so far 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, 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" [1988, p. 290]. The discourse analyses that were built by Grosz and Sidner [1986] exhibit a similar property: the intentions of some discourse segments are more important than the intentions of other discourse segments.

A careful analysis of the discourse structures that Mann, Thompson, Grosz, Sidner, Hobbs, and many others built and my own discourse analyses of more than 2100 texts (see chapter 4) has 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, that relation also holds between two or more linguistic or nonlinguistic constructs that pertain to the most important constituents of those nodes.*

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 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 2.4.a, 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. Similarly, if we examine text (2.1), we can notice, for example, that the JOINT relation that holds between span [1,2] and unit 3, also holds between unit 1, which is the most important unit in span [1,2], and unit 3.

In the general case, the constructs that the compositionality criterion refers to need not be clauses. Consider the following example:

(2.5) [He wanted to play squash with Janet,^{A_2}] [but he also wanted to have dinner with Suzanne.^{B_2}] [This indecisiveness drove him crazy.^{C_2}]

The RS-tree in figure 2.6 shows the RST analysis of text (2.5), in which units A_2 and B_2

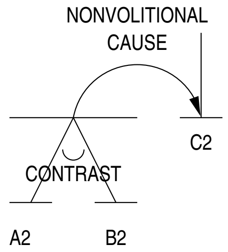

Figure 2.6: A rhetorical analysis of text (2.5).

are connected through a CONTRAST relation. The text span that results, $[A_2, B_2]$, is further connected with textual unit C_2 through a NONVOLITIONAL CAUSE relation. Note, however, that in this case, the NONVOLITIONAL CAUSE relation holds neither between A_2 and C_2, nor between B_2 and C_2. Rather, the relation shows that the CONTRAST between A_2 and B_2, i.e., the incompatibility between the two plans, caused the situation presented in C_2. In this case, the constructs that the compositionality criterion refers to are the textual unit C_2, and the CONTRAST relation that holds between units A_2 and B_2. The phrase "This indecisiveness" in textual unit C_2 makes reference precisely to the CONTRAST relation. Note also that the CONTRAST relation is a multinuclear (or paratactic) relation that assigns the rhetorical status of NUCLEUS to both units A_2 and B_2. Since both A_2 and B_2 are the most important units of span $[A_2, B_2]$, it follows that the rhetorical relation between them 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 following example, which was first used by Webber [1988a, p. 115]:

(2.6) [There are two houses you might be interested in:^{A_3}]

[House A is in Palo Alto.^{B_3}] [It's got 3 bedrooms and 2 baths,^{C_3}] [and was built in 1950.^{D_3}] [It's on a quarter acre, with a lovely garden,^{E_3}] [and the owner is asking \$425K.^{F_3}] [But **that**'s all I know about it.^{G_3}]

[House B is in Portola Valley.^{H_3}] [It's got 3 bedrooms, 4 baths and a kidney-shaped pool,^{I_3}] [and was also built in 1950.^{J_3}] [It's on 4 acres of steep wooded slope, with a view of the mountains.^{K_3}] [The owner is asking \$600K.^{L_3}] [I heard all **this** from a friend,^{M_3}] [who saw the house yesterday.^{N_3}]

[Is **that** enough information for you to decide which to look at?^{P_3}]

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 text (2.6) refers not to house A, an

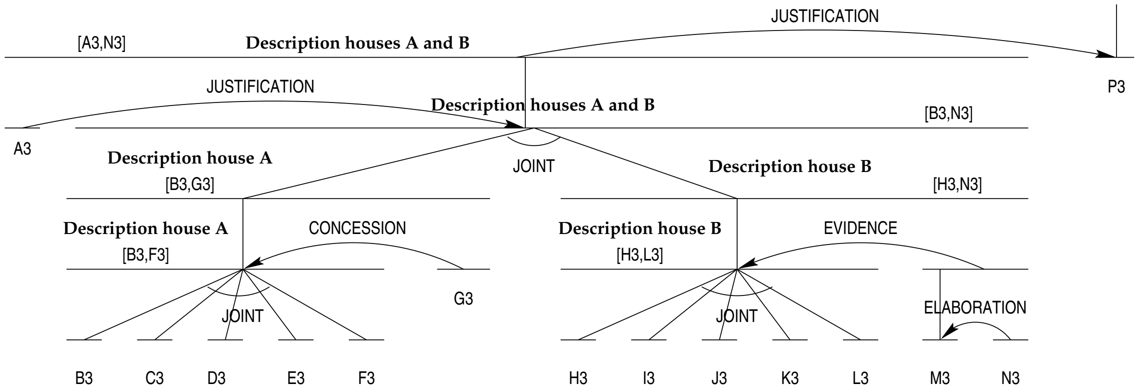

Figure 2.7: A rhetorical analysis of text (2.6).

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 2.7 shows the RST analysis of text (2.6). 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 [B3, F3] and [H3, L3] contains a set of elementary units that are connected by a JOINT relation. The linguistic constructs that these sets of units induce are the descriptions of the two houses; these constructs are shown in boldface fonts in figure 2.7. Text span G3 specifies only that the content presented in units B3–F3 is all that the writer knows. At the time unit G3 is produced, the construct **Description house A** is already available for reference, so this explains why the first boldfaced "that" in text (2.6) makes sense. Because **Description house A** is an important construct of span [B3, F3], and because [B3, F3] is the nucleus of the span [B3, G3], it is natural to consider that **Description house A** is an important construct for span [B3, G3] as well. Reasoning similarly, we can explain why the boldfaced "this" makes sense and why **Description house B** is an important construct for span [H3, N3]. Because spans [B3, G3] and [H3, N3] are connected through a JOINT relation, i.e., a multinuclear relation, the important constructs of each of them could be promoted to the higher level span, [B3, N3]. This explains why **Description houses A and B** is an important construct of span [B3, N3]. Following the same procedure, **Description houses A and B** becomes an important construct for span [A3, N3], which explains why the second boldfaced "that" in text 2.6 makes sense.

Again, as in the previous cases, the interpretation given above is consistent with the compositionality criterion. For example, the CONCESSION relation between span [B3, F3] and unit G3 also holds between the the construct **Description house A** and unit G3. The JOINT relation between spans [B3, G3] and [H3, N3] also holds between the descriptions of the two houses.

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 representation. 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 theory of the structure of unrestricted texts, I cannot take compositionality criterion 2.1 as foundational because is too underspecified.

2.5.2 A strong compositionality criterion

Although compositionality criterion 2.1 is too weak to be useful, I believe that we can still contribute to the general understanding of text by constructing a theory that takes as foundational a weaker criterion. The intuition behind the weaker 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, that relation also holds between the most important units of the constituent spans.*

If we reconsider text (2.3) and the tree in figure 2.4.a from the perspective of the strong compositionality criterion, we get the same interpretation as in the case of the weak compositionality criterion: the EVIDENCE relation that holds between text spans $[\text{C}_1, \text{D}_1]$ and $[\text{A}_1, \text{B}_1]$ also holds between their most important subspans, i.e., between the spans C_1 and B_1.

In the case of text (2.5), whose RS-tree is given in figure 2.6, the strong compositionality criterion is tautological because it specifies that the NONVOLITIONAL CAUSE relation that holds between spans $[\text{A}_2, \text{B}_2]$ and C_2 also holds between A_2,B_2 and C_2 — the most important subspans of span $[\text{A}_2, \text{B}_2]$ are both A_2 and B_2. 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 2.6 could identify that by herself because both units A_2 and B_2 are considered important for span $[\text{A}_2, \text{B}_2]$ and therefore, that span represents the relation between the CONTRAST relation and textual unit C_2 implicitly.

In the case of text (2.6), whose RS-tree is shown in figure 2.7, the strong compositionality criterion specifies, for example, that the rhetorical relation between spans $[B_3, G_3]$ and $[H_3, N_3]$ also holds between their most important subspans, i.e., between spans $[B_3, F_3]$ and $[H_3, L_3]$. 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 spans $[B_3, F_3]$ and $[H_3, L_3]$, respectively.

2.6 The formalization of text structures

2.6.1 A concrete formulation of the text structure formalization problem

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 thesis 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 2.2.1 and the strong compositionality criterion given in proposition 2.2 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 if and only if that relation holds also between the most salient units of those spans. Obviously, the formalization will also specify the rules according to which the most salient units of a text are determined. Formally, the problem that I want to solve is that given in definition 2.1, below.

Definition 2.1. The problem of text structure derivation: *Given a sequence of textual units $U = u_1, u_2, \dots, u_n$ and a set RR of rhetorical relations that hold among these units, find all valid text structures (trees) of the linear sequence $u_1, u_2, \dots, u_n$.*

The problem of text structure derivation given above is consistent with a position that assumes that rhetorical relations that hold between large textual spans should be derived only from rhetorical relations that hold between elementary units. Nevertheless, psycholinguistic experiments suggest that humans are able to determine rhetorical relations that hold between large textual spans as well. I call such relations *extended rhetorical relations.* Although humans are not consistent at determining the boundaries of large textual spans [Grosz and Hirschberg, 1992, Passonneau and Litman, 1993, Hirschberg and Nakatani, 1996, Passonneau and Litman, 1997a, Moser and Moore, 1997], I believe that a theory of text structures should accommodate judgements that pertain to large textual spans as well. Definition 2.2, which is given below, accounts for this case.

Definition 2.2. An extended formulation of 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 and extended 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.*

In this section, I provide a formalization for the extended formulation of the problem of text structure derivation. The formalization of the formulation given in definition 2.1 can be obtained from the formalization given here by taking the set of extended rhetorical relations that hold among non-elementary spans of a text to be empty.

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

- Predicate $position(u, i)$ is true for a textual unit u in sequence U if and only if u is the i-th element in the sequence.[3]

- 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 (2.3) 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 (2.4):

$$(2.7) \qquad \begin{cases} rhet\_rel(\text{JUSTIFICATION}, \text{A}_1, \text{B}_1) \\ rhet\_rel(\text{JUSTIFICATION}, \text{D}_1, \text{B}_1) \\ rhet\_rel(\text{EVIDENCE}, \text{C}_1, \text{B}_1) \\ rhet\_rel(\text{CONCESSION}, \text{D}_1, \text{C}_1) \\ rhet\_rel(\text{RESTATEMENT}, \text{D}_1, \text{A}_1) \\ position(\text{A}_1, 1), position(\text{B}_1, 2) \\ position(\text{C}_1, 3), position(\text{D}_1, 4) \end{cases}$$

[3]Instead of using the predicate *position*, we could have assumed that the textual units of a text are always labelled with numbers that reflect their index in the text they occur. However, since the formalization of text structures will be also used in natural language generation in order to produce sequences of units that are most likely to be coherent, such an approach would be misleading. To avoid confusion, I prefer to use an explicit predicate.

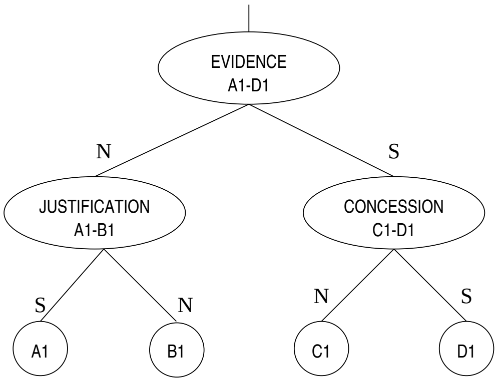

Figure 2.8: A binary representation isomorphic to the RS-tree shown in figure 2.4a.

- Predicate $rhet\_rel\_ext(name, s_s, s_e, n_s, n_e)$ 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 $rhet\_rel\_ext(\text{JUSTIFICATION}, \text{A}_1, \text{A}_1, \text{B}_1, \text{D}_1)$ holds between unit A_1 and span $[\text{B}_1, \text{D}_1]$.

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

Features of the formalization. To simplify my formalization, I follow the traditional approach and assume without restricting the generality of the problem that text trees are binary trees. 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

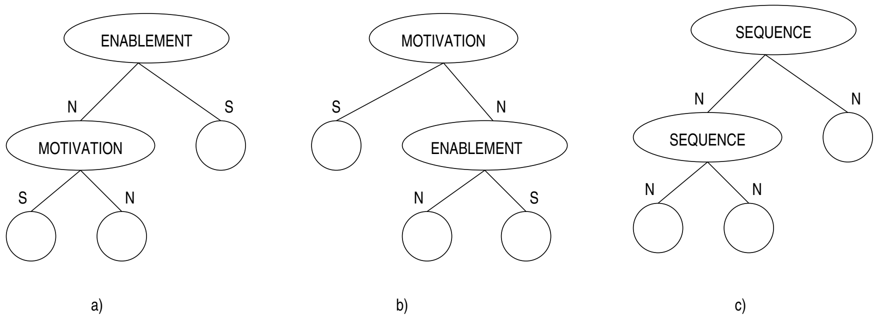

Figure 2.9: Binary trees isomorphic to the non-binary trees shown in figure 2.3(d,e)

---

which that rhetorical relation holds. The mapping preserves the labelling associated with the nuclear status of each node. For example, a binary representation of the RS-tree in figure 2.4.a is given in figure 2.8.

In fact, we can interpret non-binary trees, such as those shown in figure 2.3.(d,e), as being collapsed versions of binary trees. For example, the tree in figure 2.3.d can be derived either from the tree in figure 2.9.a or that in 2.9.b; and the tree in figure 2.3.e can be derived from the tree in figure 2.9.c. 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].

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 2.4.a, which reflects these characteristics, is given in figure 2.10. The status, type, and salience unit that are associated with each leaf follows directly from the convention that I have given above. The status and the type of each internal node is a one-to-one mapping of the status and rhetorical relation that are associated with each non-minimal text span from the original representation. 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.

The most significant differences between the tree in figure 2.10 and the tree in figure 2.4.a pertain to the promotion sets that are associated with every internal node. These promotion

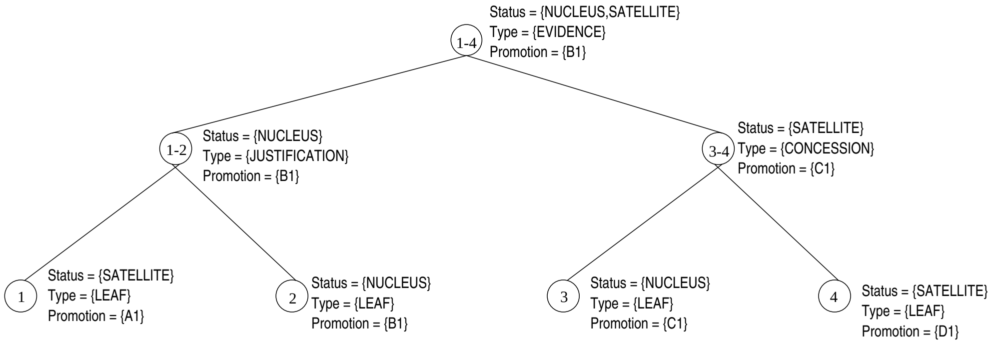

Figure 2.10: An isomorphic representation of tree in figure 2.4.a according to the status, type, and promotion features that characterize every node. The numbers associated with each node denote the limits of the text span that that node characterizes.

---

sets play a major role in determining the validity of a text tree. The tree in figure 2.10 is valid, because the EVIDENCE relation that holds between spans $[\mathrm{C}_1, \mathrm{D}_1]$ and $[\mathrm{A}_1, \mathrm{B}_1]$ also holds between their most salient units, i.e., C_1 and B_1.

The status, type, and 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.[4]

Let $span_{i,j}$, or simply $[i, j]$, denote a text span that includes all the textual units between position 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, $span_{1,3}, span_{2,4}, \ldots, span_{n-2,n}$; $\ldots$; 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 correct text trees. In fact, my intent is to determine from the set of $n+(n-1)+(n-2)+\ldots+1 = n(n+1)/2$ potential text spans that

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

pertain to a sequence of n textual units, the subset that adheres to the constraints that I have mentioned above. For example, for text 2.3, 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 2.10, 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, $span_{l,h}$,[5] which will or will not eventually become a node in the final discourse tree, is characterized by the following relations:

- $S(l, h, status)$ denotes 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 2.10, the following relations hold: $S(1, 2, \text{NUCLEUS}), S(3, 4, \text{SATELLITE})$, $S(1, 3, \text{NONE})$.

- $T(l, h, relation\_name)$ denotes the name of the rhetorical relation that holds between the text spans that are immediate subordinates of $span_{l,h}$ in the text tree.[6] 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 2.10, the following relations hold: $T(1, 1, \text{LEAF}), T(1, 2, \text{JUSTIFICATION}), T(3, 4, \text{CONCESSION})$, $T(1, 3, \text{NONE})$.

- $P(l, h, unit\_name)$ denotes 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. 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 2.10, the following relations hold: $P(1, 1, \text{A}_1), P(1, 2, \text{B}_1), P(1, 3, \text{NONE}), P(3, 4, \text{C}_1)$.

2.6.2 A complete formalization of text trees

Using the conventions that I have discussed in the previous subsection, I present now a complete first-order formalization of 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; the set of unit names that are associated with each textual unit; and four extra constants: NUCLEUS, SATELLITE, NONE, and LEAF. 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 five

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

[6] The names of the rhetorical relations are dependent on the set of relations that one uses.

other predicate symbols: S, T, and P to account for the status, type, and salient units that are associated with every text span; *rhet_rel* to account for the rhetorical relations that hold between different textual units; and *position* to account for the index of the textual units in the text that one considers. I use 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 *RS-tree* whenever I refer to a text structure that uses the taxonomy of relations defined by Mann and Thompson [1988].

Throughout this thesis, I apply the convention that all unbound variables are universally quantified and that variables are represented in lower case letters while constants 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 $span_{l,h}$, $relevant\_rel(l, h, name)$ (2.8) describes the set of simple 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 $[l, h]$.

$$
\begin{aligned}
(2.8)\qquad & relevant\_rel(l, h, name) \equiv \\
& \quad (\exists s, n, sp, np)[position(s, sp) \wedge position(n, np) \wedge \\
& \quad\quad (l \leq sp \leq h) \wedge (l \leq np \leq h) \wedge rhet\_rel(name, s, n)] \vee \\
& \quad (\exists s_s, s_e, n_s, n_e, l_1, h_1, l_2, h_2,)[position(s_s, l_1) \wedge position(s_e, h_1) \wedge \\
& \quad\quad position(n_s, l_2) \wedge position(n_e, h_2) \wedge (l \leq l_1 \leq h_1 \leq h) \wedge \\
& \quad\quad (l \leq l_2 \leq h_2 \leq h) \wedge rhet\_rel\_ext(name, s_s, s_e, n_s, n_e)]
\end{aligned}
$$

For every text span $span_{l,h}$, $relevant\_unit(l, h, u)$ (2.9) 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 $[l, h]$.

$$
(2.9)\qquad relevant\_unit(l, h, u) \equiv \ (\exists x)[position(u, x) \wedge (l \leq x \leq h)]
$$

For example, for text (2.3), which is described formally in (2.7), 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:

$$
\begin{aligned}
\{&relevant\_rel(1, 3, \textsc{justification}), relevant\_rel(1, 3, \textsc{evidence}), \\
&\quad relevant\_unit(1, 3, \textsc{a}_1), relevant\_unit(1, 3, \textsc{b}_1), relevant\_unit(1, 3, \textsc{c}_1)\}
\end{aligned}
$$

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.

$$\begin{aligned}(2.10)\qquad & [(1 \leq h \leq \text{N}) \wedge (1 \leq l \leq h)] \rightarrow \\ & \{[l = h \rightarrow (S(l,h,\text{NUCLEUS}) \vee S(l,h,\text{SATELLITE}))] \wedge \\ & [l \neq h \rightarrow (S(l,h,\text{NUCLEUS}) \vee S(l,h,\text{SATELLITE}) \vee S(l,h,\text{NONE}))]\}\end{aligned}$$

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

$$\begin{aligned}(2.11)\qquad & [(1 \leq h \leq \text{N}) \wedge (1 \leq l \leq h)] \rightarrow \\ & [(S(l,h,status_1) \wedge S(l,h,status_2)) \rightarrow status_1 = status_2]\end{aligned}$$

• **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.

$$\begin{aligned}(2.12)\qquad & [(1 \leq h \leq \text{N}) \wedge (1 \leq l \leq h)] \rightarrow \\ & \{[l = h \rightarrow T(l,h,\text{LEAF})] \wedge \\ & [l \neq h \rightarrow (T(l,h,\text{NONE}) \vee \\ & \qquad (T(l,h,name) \rightarrow relevant\_rel(l,h,name)))]\}\end{aligned}$$

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

$$\begin{aligned}(2.13)\qquad & [(1 \leq h \leq \text{N}) \wedge (1 \leq l < h)] \rightarrow \\ & [(T(l,h,name_1) \wedge T(l,h,name_2)) \rightarrow name_1 = name_2]\end{aligned}$$

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

$$\begin{aligned}(2.14)\qquad & [(1 \leq h \leq \text{N}) \wedge (1 \leq l \leq h)] \rightarrow \\ & [P(l,h,\text{NONE}) \vee (P(l,h,u) \rightarrow relevant\_unit(l,h,u))]\end{aligned}$$

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 2.2.1 and from the strong compositionality criterion given in proposition 2.2.

- **Text spans do not overlap.**

(2.15)
$$[(1 \leq h_1 \leq \textsc{n}) \wedge (1 \leq l_1 \leq h_1) \wedge (1 \leq h_2 \leq \textsc{n}) \wedge (1 \leq l_2 \leq h_2) \wedge (l_1 < l_2) \wedge (h_1 < h_2) \wedge (l_2 \leq h_1)] \rightarrow [\neg S(l_1, h_1, \textsc{none}) \rightarrow S(l_2, h_2, \textsc{none})]$$

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

(2.16)
$$[(1 \leq h \leq \textsc{n}) \wedge (1 \leq l < h)] \rightarrow [(S(l, h, \textsc{none}) \wedge P(l, h, \textsc{none}) \wedge T(l, h, \textsc{none})) \vee (\neg S(l, h, \textsc{none}) \wedge \neg P(l, h, \textsc{none}) \wedge \neg T(l, h, \textsc{none}))]$$

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

(2.17)
$$\neg S(1, \textsc{n}, \textsc{none}) \wedge \neg P(1, \textsc{n}, \textsc{none}) \wedge \neg T(1, \textsc{n}, \textsc{none})$$

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

(2.18)
$$\begin{aligned}
&[(1 \leq h \leq \textsc{n}) \wedge (1 \leq l < h) \wedge \neg S(l, h, \textsc{none})] \rightarrow \\
&\quad (\exists name, split_point, s, n)[(l \leq split_point \leq h) \\
&\qquad \wedge (Nucleus_first(name, split_point, s, n) \vee \\
&\qquad\quad Satellite_first(name, split_point, s, n))] \vee \\
&\quad (\exists name, split_point, s_s, s_e, n_s, n_e)[(l \leq split_point \leq h) \\
&\qquad \wedge (Nucleus_first_ext(name, split_point, s_s, s_e, n_s, n_e) \vee \\
&\qquad\quad Satellite_first_ext(name, split_point, s_s, s_e, n_s, n_e))]
\end{aligned}$$

$$(2.19)\qquad \begin{aligned}&\mathit{Nucleus_first}(\mathit{name}, \mathit{split_point}, s, n) \equiv\\
&\mathit{rhet_rel}(\mathit{name}, s, n) \wedge T(l, h, \mathit{name}) \wedge \mathit{position}(s, sp) \wedge \mathit{position}(n, np) \wedge\\
&l \leq np \leq \mathit{split_point} \wedge \mathit{split_point} < sp \leq h \ \wedge\\
&P(l, \mathit{split_point}, n) \wedge P(\mathit{split_point}+1, h, s) \wedge\\
&\{(\mathit{name} \in \mathit{Rels}_{\mathit{paratactic}}) \rightarrow\\
&\quad S(l, \mathit{split_point}, \textsc{nucleus}) \wedge S(\mathit{split_point}+1, h, \textsc{nucleus}) \wedge\\
&\quad (\forall p)[P(l, h, p) \equiv (P(l, \mathit{split_point}, p) \vee P(\mathit{split_point}+1, h, p))]\} \wedge\\
&\{(\mathit{name} \in \mathit{Rels}_{\mathit{hypotactic}}) \rightarrow\\
&\quad S(l, \mathit{split_point}, \textsc{nucleus}) \wedge S(\mathit{split_point}+1, h, \textsc{satellite}) \wedge\\
&\quad (\forall p)(P(l, h, p) \equiv P(l, \mathit{split_point}, p))\}\end{aligned}$$

$$(2.20)\qquad \begin{aligned}&\mathit{Satellite_first}(\mathit{name}, \mathit{split_point}, s, n) \equiv\\
&\mathit{rhet_rel}(\mathit{name}, s, n) \wedge T(l, h, \mathit{name}) \wedge \mathit{position}(s, sp) \wedge \mathit{position}(n, np) \wedge\\
&l \leq sp \leq \mathit{split_point} \wedge \mathit{split_point} < np \leq h \ \wedge\\
&P(l, \mathit{split_point}, s) \wedge P(\mathit{split_point}+1, h, n) \wedge\\
&\{(\mathit{name} \in \mathit{Rels}_{\mathit{paratactic}}) \rightarrow\\
&\quad S(l, \mathit{split_point}, \textsc{nucleus}) \wedge S(\mathit{split_point}+1, h, \textsc{nucleus}) \wedge\\
&\quad (\forall p)[P(l, h, p) \equiv (P(l, \mathit{split_point}, p) \vee P(\mathit{split_point}+1, h, p))]\} \wedge\\
&\{(\mathit{name} \in \mathit{Rels}_{\mathit{hypotactic}}) \rightarrow\\
&\quad S(l, \mathit{split_point}, \textsc{satellite}) \wedge S(\mathit{split_point}+1, h, \textsc{nucleus}) \wedge\\
&\quad (\forall p)(P(l, h, p) \equiv P(\mathit{split_point}+1, h, p))\}\end{aligned}$$

$$(2.21)\qquad \begin{aligned}&\mathit{Nucleus_first_ext}(\mathit{name}, \mathit{split_point}, s_s, s_e, n_s, n_e) \equiv\\
&\{[\mathit{rhet_rel_ext}(\mathit{name}, s_s, s_e, n_s, n_e) \wedge T(l, h, \mathit{name}) \wedge\\
&\mathit{position}(s_s, \mathit{split_point}+1) \wedge \mathit{position}(s_e, h) \wedge\\
&\mathit{position}(n_s, l) \wedge \mathit{position}(n_e, \mathit{split_point}) \wedge\\
&\{(\mathit{name} \in \mathit{Rels}_{\mathit{paratactic}}) \rightarrow\\
&\quad S(l, \mathit{split_point}, \textsc{nucleus}) \wedge S(\mathit{split_point}+1, h, \textsc{nucleus}) \wedge\\
&\quad (\forall p)[P(l, h, p) \equiv (P(l, \mathit{split_point}, p) \vee P(\mathit{split_point}+1, h, p))]\} \wedge\\
&\{(\mathit{name} \in \mathit{Rels}_{\mathit{hypotactic}}) \rightarrow\\
&\quad S(l, \mathit{split_point}, \textsc{nucleus}) \wedge S(\mathit{split_point}+1, h, \textsc{satellite}) \wedge\\
&\quad (\forall p)(P(l, h, p) \equiv P(l, \mathit{split_point}, p))\}\end{aligned}$$

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

Formula (2.18) 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 (2.19) specifies that there is a rhetorical relation with name *name*, from a unit s (in most cases a satellite) that belongs to span $[\mathit{split\_point}+1, h]$ to a unit n, the nucleus, that belongs to span $[l, \mathit{split\_point}]$; that unit n is salient with respect to text span $[l, \mathit{split\_point}]$ and unit s is salient with respect to text span $[\mathit{split\_point}+1, h]$; 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 sub-spans 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 sub-spans. If the relation is hypotactic, the status of text span $[l, \mathit{split\_point}]$ is NUCLEUS, the status of text span $[\mathit{split\_point}+1, h]$ is SATELLITE and the set of salient units for text span $[l, h]$ are given by the salient units that are associated with the subordinate nucleus span. The $\in$ symbol in formulas (2.19) and (2.22) 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) (2.20) is a mirror image of (2.19) and it describes the case when the satellite that pertains to rhetorical relation $\mathit{rhet\_rel}(\mathit{name}, s, n)$ belongs to text span $[l, \mathit{split\_point}]$, i.e., when the satellite goes before the nucleus.

Formula (2.21) 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 sub-spans 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 sub-spans. If the relation is hypotactic, the status of text span $[l, \mathit{split\_point}]$ is NUCLEUS, the status of text span $[\mathit{split\_point} + 1, h]$ is SATELLITE and the set of salient units for text span $[l, h]$ are given by the salient units that are associated with the subordinate nucleus span. Formula (2.22) is a mirror image of (2.21) and it describes the case when the units of the satellite span s_s–s_e that pertains to the extended rhetorical relation $rhet\_rel\_ext(name, s_s, s_e, n_s, n_e)$ belongs to text span $[l, \mathit{split\_point}]$, i.e., when the satellite goes before the nucleus.

For the rest of the thesis, the set of axioms (2.8)–(2.22) will be referred to as *the axiomatization of valid text structures.*

2.6.3 A formalization of RST

The axiomatization of valid text structures given in section 2.6.2 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 (2.19), (2.20), (2.21), and (2.22) the sets of hypotactic and paratactic relations that are proposed in RST. For example, axiom (2.23) is the RST instantiation of axiom (2.19).

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

In a similar manner, we can instantiate axioms (2.20), (2.21), and (2.22) as well. For the rest of the thesis, axioms (2.8)–(2.18) and the set of axioms that are derived from axioms (2.19)–(2.22) by instantiating the taxonomy 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 2.4.a and 2.4.c satisfy all

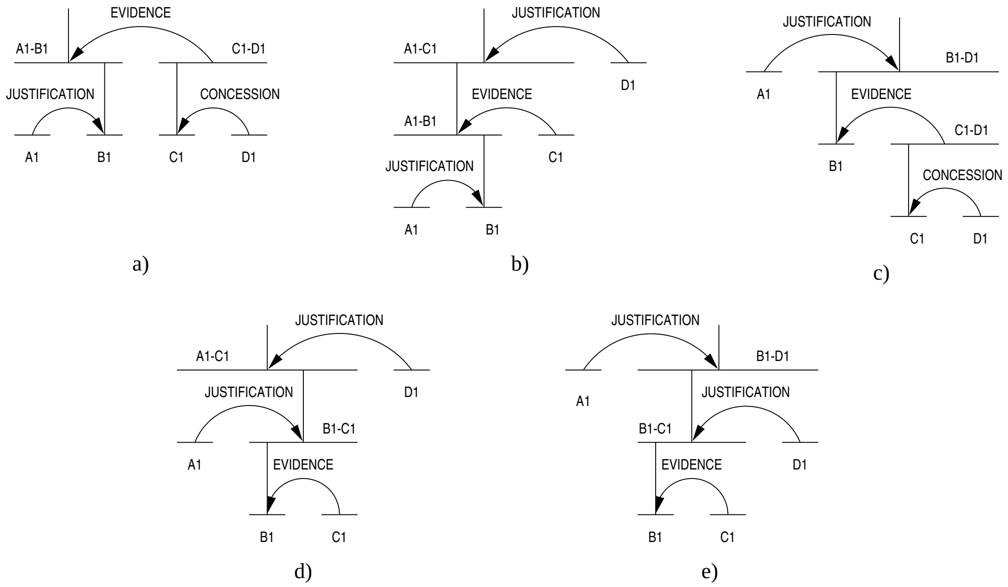

Figure 2.11: The set of all RS-trees that could be built for text (2.3).

---

the axioms, while the structure of the tree in figure 2.4.b does not satisfy axiom (2.18). More precisely, the rhetorical relation of CONCESSION between D_1 and C_1 projects C_1 as the salient unit for text span $[C_1, D_1]$. The initial set of rhetorical relations (2.7) 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 to play both a satellite role in the span $[C_1, D_1]$, and to be, at the same time, a salient unit for it, it follows that tree 2.4.b is incorrect.

If we determine all the ways in which the logical theory that pertains to the formal representation of text (2.3) (axioms (2.7)) and the axiomatization of RST can be satisfied, we obtain five models that correspond to the trees in figure 2.11. Among the set of trees in figure 2.11, trees 2.11.a and 2.11.b match the trees given earlier in figure 2.4.a and 2.4.c. Trees 2.11.c–e represent trees that are not given in figure 2.4.

If the relations to the same text were to consist of the relations given below in (2.24), then only one tree could correspond to text (2.3), the tree in figure 2.11.e.

$$(2.24) \qquad \begin{cases} rhet\_rel(\text{JUSTIFICATION}, D_1, B_1) \\ rhet\_rel(\text{EVIDENCE}, C_1, B_1) \\ rhet\_rel(\text{JUSTIFICATION}, A_1, [B_1\text{–}D_1]) \end{cases}$$

2.7 Towards formalizing the relationship between text trees and intentions

2.7.1 Preamble

In the last decade, the members of the computational linguistics community have adopted primarily either an RST- or a GST-based perspective on discourse. Only recently, researchers have started to investigate the relationship between the two perspectives [Moser and Moore, 1996]. In this section, I formalize the relationship between the structure of text and intentions. As in the rest of the chapter, I will take a more general perspective and assume only that rhetorical relations can be partitioned into paratactic and hypotactic relations. However, for exemplification, I will use the set of rhetorical relations that was defined by Mann and Thompson [1988]. To increase the understandability of the arguments that I am going to make in this section, I will rely on a text that was first used by Holmes and Gallagher [1917] and Cohen [1983], and then by Grosz and Sidner [1986, p. 183]. The text is given in (2.25), below.

(2.25) [The "movies" are so attractive to the great American public,^{A4}] [especially to young people,^{B4}] [that it is time to take careful thought about their effect on mind and morals.^{C4}] [Ought any parent to permit his children to attend a moving picture show often or without being quite certain of the show he permits them to see?^{D4}] [No one can deny, of, course, that great educational and ethical gains may be made through the movies^{E4}] [because of their astonishing vividness.^{F4}] [But the important fact to be determined is the total result of continuous and indiscriminate attendance of shows of this kind.^{G4}] [Can it be other than harmful?^{H4}] [In the first place the character of the plays is seldom of the best.^{I4}] [One has only to read the ever-present "movie" billboard to see how cheap, melodramatic and vulgar most of the photoplays are.^{J4}] [Even the best plays, moreover, are bound to be exciting and over-emotional.^{K4}] [Without spoken words, facial expression and gesture must carry the meaning:^{L4}] [but only strong emotion, or buffoonery, can be represented through facial expression and gesture.^{M4}] [The more reasonable and quiet aspects of life are necessarily neglected.^{N4}] [How can our young people drink in through their eyes a continuous spectacle of intense and strained activity and feeling without harmful effects?^{O4}] [Parents and teachers will do well to guard the young against overindulgence in the taste for the "movie".^{P4}]

The intention-based discourse structure that Grosz and Sidner built for text (2.25) is shown in figure 2.12: the leaves of the structure are labelled both with the literals that are used in example (2.25) and with numbers that correspond to the boundaries of those

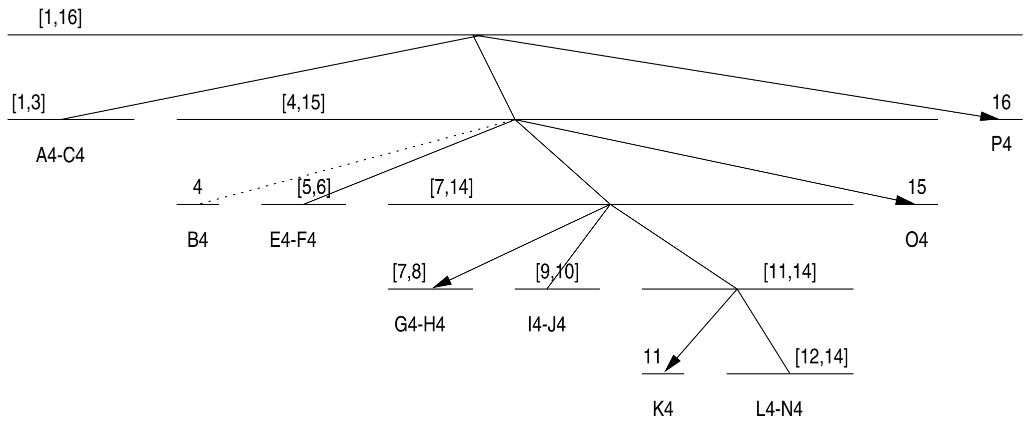

Figure 2.12: The intention-based discourse structure of text (2.25).

units in the text — as before, it is assumed that to each elementary textual unit there corresponds a natural number that reflects the position of that unit in the sequence of units that make up the text. For simplicity, the internal nodes are labelled using only the numbers that correspond to the boundaries of the corresponding discourse segments. The solid lines depict explicit dominance relations; the arrows depict the segments that induce the primary intentions of the immediately dominant discourse segments; and the dotted line depicts an implicit dominance relation that is not mentioned by Grosz and Sidner [1986, p. 184]. For example, discourse segment $[11, 14]$ dominates discourse segment $[12, 14]$, discourse segment $[7, 14]$ dominates discourse segments $[9, 10]$ and $[11, 14]$, etc. The primary intention of discourse segment $[11, 14]$ is that the writer intends the reader to believe proposition 11. The primary intention of discourse segment $[7, 14]$ is that the writer intends the reader to believe propositions $7, 8$, etc.

If we examine the structure that Grosz and Sidner propose and the relations between the discourse segments and their primary intentions, it is easy to notice that there is a clear correspondence between GST and RST. To highlight this correspondence, consider also an RST-like analysis of the same text (see figure 2.13). In addition to the classical conventions used to represent RS-trees, figure 2.13 also shows in bold the salient units that are associated with each internal node. For a better comparison, the spans that were considered elementary in Grosz and Sidner's analysis (figure 2.12) use horizontal lines that are thicker than the lines used for the other spans.

By inspecting figures 2.12 and 2.13, we can immediately notice that the structures that the two theories assign to text (2.25) are similar. The only difference pertains to their granularities: RST takes clause-like segments as being the elementary units of discourse, while GST puts no constraints on the size of the elementary units — in GST, elementary units can be clauses, sentences, groups of sentences, and even paragraphs. In addition, one can also see that there also exists a clear correspondence between the primary intentions

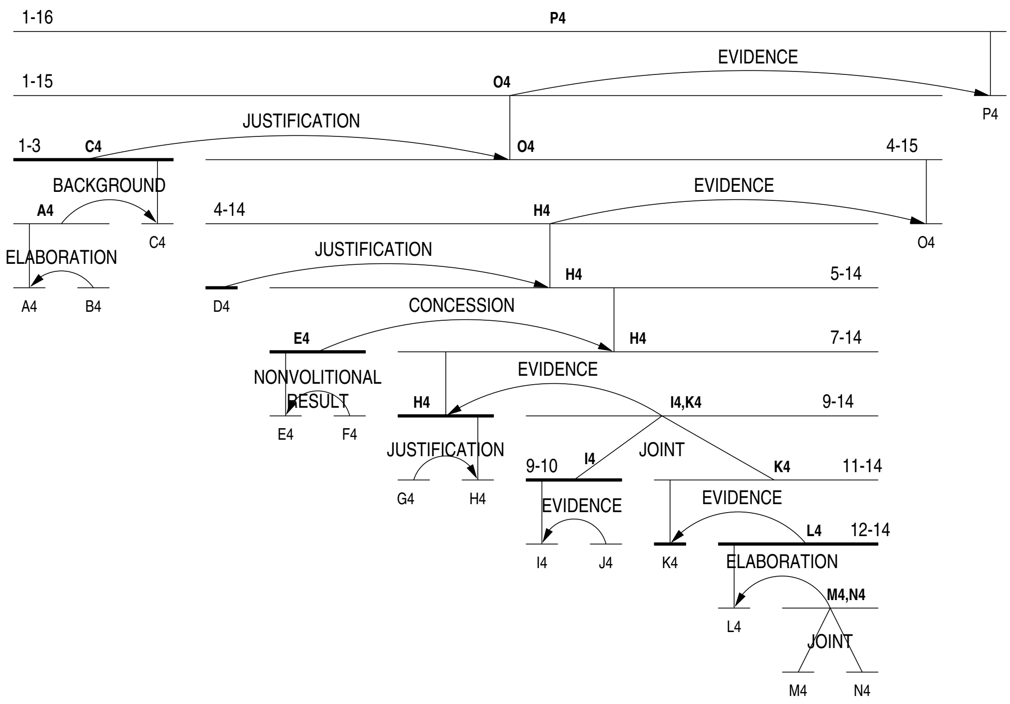

Figure 2.13: A rhetorical structure analysis of text (2.25).

associated with Grosz and Sidner's discourse segments and the salient units associated with the internal nodes of the RST representation. Table 2.1 makes this correspondence explicit: with the exception of textual unit H_4, all other salient units in RST correspond to the primary intentions associated with the discourse segments built by Grosz and Sidner. In fact, even the primary intention associated with span [7,14], which Grosz and Sidner take to be (Intend ICP (Believe OCP "the proposition that although there are gains, the total result of continuous and indiscriminate attendance at movies is harmful")), is mainly derived from unit H_4, which is the salient unit of the same span in the RST representation.[7]

In a recent proposal, Moser and Moore [1996] argued that the primary intentions in a GST representation can be associated with the nuclei of the corresponding RST representation. Although their proposal is consistent with the cases in which each textual span is characterized by an explicit nucleus that encodes the primary intention of that span, we believe that an adequate account of the correspondence between GST and RST can be given only if we consider the weak compositionality criterion 2.1. As we discussed in section 2.5, in some cases, the salient constructs of a textual span can be both of linguistic or nonlinguistic nature. For example, in the case of text (2.5), which we reproduce for convenience in (2.26) below, we can associate the primary intention of discourse segment

[7] I use the notation (Intend ICP (Believe OCP $f_I(\text{H}_4)$)) in order to distinguish between the cases in which the primary intention was given explicitly by a textual unit, and the special case that pertains to segment [7, 14], in which the primary intention is derived through some inferential mechanisms from unit H_4.

Span or Discourse Segment	Intention in GST	Salient units in RST
[1,16]	(Intend ICP (Believe OCP P_4))	P_4
[1,3]	(Intend ICP (Believe OCP C_4))	C_4
[4,15]	(Intend ICP (Believe OCP O_4))	O_4
[5,6]	(Intend ICP (Believe OCP E_4))	E_4
[7,14]	(Intend ICP (Believe OCP $f_I(\text{H}_4)$))	H_4
[9,10]	(Intend ICP (Believe OCP I_4))	I_4
[12,14]	(Intend ICP (Believe OCP M_4, N_4))	M_4, N_4

Table 2.1: The correspondence between the primary intentions of discourse segments in GST and the salient units of the text spans in RST. ICP and OCP denote the Initiating Conversational Participant (the writer) and the Other Conversational Participant (the reader) respectively; the terms x associated with the tuples (Believe OCP x) denote the corresponding propositions from text (2.25).

$[\text{A}_2, \text{B}_2]$ neither to unit A_2 nor to unit B_2. Rather, the primary intention pertains to the rhetorical relation between the two units. In Grosz and Sidner's terms, we can say that the primary intention of segment $[\text{A}_2, \text{B}_2]$ is (Intend ICP (Believe OCP "he wanted to do two things that were incompatible")). In other words, the intention associated with segment $[\text{A}_2, \text{B}_2]$ is a function both of its salient units, A_2 and B_2, and of the rhetorical relation that holds between these units.

(2.26) [He wanted to play squash with Janet,$^{\text{A}_2}$] [but he also wanted to have dinner with Suzanne.$^{\text{B}_2}$] [This indecisiveness drove him crazy.$^{\text{C}_2}$]

Similarly, in Webber's text (2.6), the primary intention of segment $[\text{B}_3, \text{F}_3]$, for example, — (Intend ICP (Inform OCP "description house A")) — arises from the juxtaposition of all the individual units in the segment. That is, the primary intention is a function both of the salient units of discourse segment $[\text{B}_3, \text{F}_3]$ and of the rhetorical relation of JOINT that holds among them. I now formalize this relationship between the primary intentions and the structure of text.

2.7.2 The melding of text structures and intentions

In formalizing the constraints that pertain both to RST-like structures and GST-like intentions, I use the same conventions that I used in section 2.6. Again, because I want to provide a formalization that is independent of the set of rhetorical relations that one uses, I will assume only that the set of rhetorical relations can be partitioned into two classes: paratactic and hypotactic. In addition to the relations discussed in section 2.6, I will also use the following predicates and functions:

- Predicate $I(l, h, intention)$ is true when *intention* denotes the primary intention of discourse span $[l, h]$. The term *intention* is represented using an oracle function f_I, which is discussed below. However, in order to simplify the exposition, let us assume for the moment that strings are first-order objects. When we do so, the following are some of the predicates that are true with respect to the discourse analysis given by Grosz and Sidner for text (2.25): $I(1, 16,$ "parents and teachers should guard the young against overindulgence in the movies"$)$ and $I(11, 14,$ "stories in movies are exciting and over-emotional"$)$.

- Predicate $dom(l_1, h_1, l_2, h_2)$ is true whenever a discourse span $[l_1, h_1]$ dominates a discourse span $[l_2, h_2]$. Some of the predicates that hold for text (2.25) are: $dom(1, 16, 1, 3)$ and $dom(11, 14, 12, 14)$. A dominance relation is well-formed if span $[l_2, h_2]$ is a proper subspan of span $[l_1, h_1]$, i.e., $l_1 \leq l_2 \leq h_2 \leq h_1 \wedge (l_1 \neq l_2 \vee h_1 \neq h_2)$.

- Predicate $satprec(l_1, h_1, l_2, h_2)$ is true whenever an intentional satisfaction-precedence relation holds between two segments $[l_1, h_1]$ and $[l_2, h_2]$. A satisfaction-precedence relation is well-formed if the spans do not overlap.

- Oracle function $f_I(r, x_1, \ldots, x_n)$ takes as arguments a rhetorical relation r and a set of textual units, and returns the primary intention that pertains to that relation and those units. For example, in the case of segment $[\text{A}_2, \text{B}_2]$ in text (2.26), the oracle function $f_I(\text{CONTRAST}, \text{A}_2, \text{B}_2)$ is assumed to return a first-order object whose meaning can be glossed as "inform the reader that the character of the story wanted to do two things that were incompatible". And the oracle function $f_I(\text{BACKGROUND}, \text{C}_4)$ associated with segment $[1, 3]$ in text (2.25) is assumed to return a first-order object whose meaning can be glossed as "inform the reader that it is time to consider the effects of movies on mind and morals"; in this case, the oracle function makes no use of the associated rhetorical relation.

The dominance and satisfaction-precedence relations that are used by Grosz and Sidner are relations that characterize a different level of abstraction than that characterized by rhetorical relations. On one hand, the dominance and satisfaction-precedence relations specify how the intentions of some discourse segments are related to the intentions of other segments. In this respect, their nature is semantic and pragmatic. On the other hand, they impose constraints on the overall discourse structure. In this respect, their nature is structural. Given the fact that the intention-based relations proposed by Grosz and Sidner are hence somewhat different from those proposed by Mann and Thompson and other discourse theorists, I will assign them a different status in the formalization.

In the formalization that I propose, each node of a discourse structure is characterized by four features: the status of the node, the rhetorical relation that holds between the nodes that are immediate children, the set of salient units, and the primary intention. For the sake of completeness, I specify here all the axioms that pertain to the axiomatization of valid text structures and GST. The axioms whose meaning was explained in the previous sections are reproduced with no further explanation.

The set of relevant relations for discourse segment $[l, h]$ is the set of rhetorical relations that span over text spans that have their boundaries within the interval $[l, h]$.

$$(2.27) \quad \begin{aligned} & relevant\_rel(l, h, name) \equiv \\ & \quad (\exists s, n, sp, np)[position(s, sp) \wedge position(n, np) \wedge \\ & \qquad (l \leq sp \leq h) \wedge (l \leq np \leq h) \wedge rhet\_rel(name, s, n)] \vee \\ & \quad (\exists s_s, s_e, n_s, n_e, l_1, h_1, l_2, h_2,)[position(s_s, l_1) \wedge position(s_e, h_1) \wedge \\ & \qquad position(n_s, l_2) \wedge position(n_e, h_2) \wedge (l \leq l_1 \leq h_1 \leq h) \wedge \\ & \qquad (l \leq l_2 \leq h_2 \leq h) \wedge rhet\_rel\_ext(name, s_s, s_e, n_s, n_e)] \end{aligned}$$

The set of relevant units for segment $[l, h]$ is given by the units whose positions in the initial sequence are numbers in the interval $[l, h]$.

$$(2.28) \quad relevant\_unit(l, h, u) \equiv \ (\exists x)[position(u, x) \wedge (l \leq x \leq h)]$$

Constraints that concern the objects over which the predicates that describe every segment $[l, h]$ of a text structure range

• **For every segment $[l, h]$, the set of objects over which predicate S ranges is the set** NUCLEUS, SATELLITE, NONE.

$$(2.29) \quad \begin{aligned} & [(1 \leq h \leq \text{N}) \wedge (1 \leq l \leq h)] \rightarrow \\ & \quad \{[l = h \rightarrow (S(l, h, \text{NUCLEUS}) \vee S(l, h, \text{SATELLITE}))] \wedge \\ & \qquad [l \neq h \rightarrow (S(l, h, \text{NUCLEUS}) \vee S(l, h, \text{SATELLITE}) \vee S(l, h, \text{NONE}))]\} \end{aligned}$$

• **The status of any discourse segment is unique**

$$(2.30) \quad \begin{aligned} & [(1 \leq h \leq \text{N}) \wedge (1 \leq l \leq h)] \rightarrow \\ & \quad [(S(l, h, status_1) \wedge S(l, h, status_2)) \rightarrow status_1 = status_2] \end{aligned}$$

• **For every segment $[l, h]$, the set of objects over which predicate T ranges is the set of rhetorical relations that are relevant to that span.**

$$(2.31) \qquad \begin{aligned} &[(1 \leq h \leq \textsc{n}) \wedge (1 \leq l \leq h)] \rightarrow \\ &\quad \{[l = h \rightarrow T(l, h, \textsc{leaf})] \wedge \\ &\quad [l \neq h \rightarrow (T(l, h, \textsc{none}) \vee \\ &\qquad\qquad (T(l, h, name) \rightarrow relevant\_rel(l, h, name)))]\} \end{aligned}$$

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

$$(2.32) \qquad \begin{aligned} &[(1 \leq h \leq \textsc{n}) \wedge (1 \leq l < h)] \rightarrow \\ &\quad [(T(l, h, name_1) \wedge T(l, h, name_2)) \rightarrow name_1 = name_2] \end{aligned}$$

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

$$(2.33) \qquad \begin{aligned} &[(1 \leq h \leq \textsc{n}) \wedge (1 \leq l \leq h)] \rightarrow \\ &\quad [P(l, h, \textsc{none}) \vee (P(l, h, u) \rightarrow relevant\_unit(l, h, u))] \end{aligned}$$

• **The primary intention of a discourse segment is either NONE or is a function of the salient units that pertain to that segment and of the rhetorical relation that holds between the immediate subordinated segments.** Since we want to stay within the boundaries of first-order logic, we express this by means of a disjunction of at most N subformulas, which correspond to the cases in which the span has 1, 2, ... , or N salient units. Formula (2.34) specifies that the intention $intention_{lh}$ associated with each node is either NONE or is a function of the salient units of the node and of the rhetorical relation that characterizes that node.

$$
\begin{aligned}
(2.34) \quad & [(1 \le h \le \text{N}) \wedge (1 \le l \le h)] \rightarrow \\
& \{I(l, h, intention_{lh}) \rightarrow \\
& intention_{lh} = \text{NONE} \vee \\
& (\exists r, x)[T(l, h, r) \wedge r \neq \text{NONE} \wedge \\
& \qquad P(l, h, x) \wedge (\forall y)(P(l, h, y) \rightarrow x = y) \wedge \\
& \qquad intention_{lh} = f_I(r, x)] \vee \\
& (\exists r, x_1, x_2)[T(l, h, r) \wedge r \neq \text{NONE} \wedge \\
& \qquad P(l, h, x_1) \wedge P(l, h, x_2) \wedge x_1 \neq x_2 \wedge \\
& \qquad (\forall y)(P(l, h, y) \rightarrow (y = x_1 \vee y = x_2)) \wedge \\
& \qquad intention_{lh} = f_I(r, x_1, x_2)] \vee \\
& \vdots \\
& (\exists r, x_1, x_2, \ldots, x_{\text{N}})[T(l, h, r) \wedge r \neq \text{NONE} \wedge \\
& \qquad x_1 \neq x_2 \wedge x_1 \neq x_3 \wedge \ldots \wedge x_1 \neq x_{\text{N}} \wedge \\
& \qquad \qquad x_2 \neq x_3 \wedge \ldots \wedge x_2 \neq x_{\text{N}} \wedge \\
& \qquad \vdots \\
& \qquad \qquad \qquad x_{\text{N}-1} \neq x_{\text{N}} \wedge \\
& \qquad P(l, h, x_1) \wedge P(l, h, x_2) \wedge \ldots \wedge P(l, h, x_n) \wedge \\
& \qquad (\forall y)(P(l, h, y) \rightarrow (y = x_1 \vee y = x_2 \vee \ldots \vee y = x_n)) \wedge \\
& \qquad intention_{lh} = f_I(r, x_1, x_2, \ldots, x_n)]\}
\end{aligned}
$$

- **The primary intention of any discourse segment is unique.**

$$
\begin{aligned}
(2.35) \quad & [(1 \le h \le \text{N}) \wedge (1 \le l < h)] \rightarrow \\
& \quad [(I(l, h, intention_1) \wedge I(l, h, intention_2)) \rightarrow intention_1 = intention_2]
\end{aligned}
$$

Constraints that concern the structure of the discourse trees

- **Discourse segments do not overlap.**

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

• **A discourse segment with status** NONE **does not participate in the tree at all.**

(2.37)
$$[(1 \leq h \leq \text{N}) \wedge (1 \leq l < h)] \rightarrow$$
$$[(S(l, h, \text{NONE}) \wedge P(l, h, \text{NONE}) \wedge T(l, h, \text{NONE}) \wedge I(l, h, \text{NONE})) \vee$$
$$(\neg S(l, h, \text{NONE}) \wedge \neg P(l, h, \text{NONE}) \wedge \neg T(l, h, \text{NONE}) \wedge \neg I(l, h, \text{NONE}))]$$

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

(2.38)
$$\neg S(1, \text{N}, \text{NONE}) \wedge \neg P(1, \text{N}, \text{NONE}) \wedge \neg T(1, \text{N}, \text{NONE}) \wedge \neg I(1, \text{N}, \text{NONE})$$

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

(2.39)
$$[(1 \leq h \leq \text{N}) \wedge (1 \leq l < h) \wedge \neg S(l, h, \text{NONE})] \rightarrow$$
$$(\exists name, split\_point, s, n)[(l \leq split\_point \leq h)$$
$$\wedge (Nucleus\_first(name, split\_point, s, n) \vee$$
$$Satellite\_first(name, split\_point, s, n))] \vee$$
$$(\exists name, split\_point, s_s, s_e, n_s, n_e)[(l \leq split\_point \leq h)$$
$$\wedge (Nucleus\_first\_ext(name, split\_point, s_s, s_e, n_s, n_e) \vee$$
$$Satellite\_first\_ext(name, split\_point, s_s, s_e, n_s, n_e))]$$

(2.40)
$$Nucleus\_first(name, split\_point, s, n) \equiv$$
$$rhet\_rel(name, s, n) \wedge T(l, h, name) \wedge position(s, sp) \wedge position(n, np) \wedge$$
$$l \leq np \leq split\_point \wedge split\_point < sp \leq h \wedge$$
$$P(l, split\_point, n) \wedge P(split\_point + 1, h, s) \wedge$$
$$\{(name \in Rels_{paratactic}) \rightarrow$$
$$S(l, split\_point, \text{NUCLEUS}) \wedge S(split\_point + 1, h, \text{NUCLEUS}) \wedge$$
$$(\forall p)[P(l, h, p) \equiv (P(l, split\_point, p) \vee P(split\_point + 1, h, p))]\} \wedge$$
$$\{(name \in Rels_{hypotactic}) \rightarrow$$
$$S(l, split\_point, \text{NUCLEUS}) \wedge S(split\_point + 1, h, \text{SATELLITE}) \wedge$$
$$(\forall p)(P(l, h, p) \equiv P(l, split\_point, p))\}$$

(2.41)
$$
\begin{aligned}
&\mathit{Satellite\_first}(name, split\_point, s, n) \equiv \\
&\quad rhet\_rel(name, s, n) \wedge T(l, h, name) \wedge position(s, sp) \wedge position(n, np) \wedge \\
&\quad l \leq sp \leq split\_point \wedge split\_point < np \leq h \ \wedge \\
&\quad P(l, split\_point, s) \wedge P(split\_point + 1, h, n) \wedge \\
&\quad \{(name \in Rels_{paratactic}) \rightarrow \\
&\qquad S(l, split\_point, \textsc{nucleus}) \wedge S(split\_point + 1, h, \textsc{nucleus}) \wedge \\
&\qquad (\forall p)[P(l, h, p) \equiv (P(l, split\_point, p) \vee P(split\_point + 1, h, p))]\} \wedge \\
&\quad \{(name \in Rels_{hypotactic}) \rightarrow \\
&\qquad S(l, split\_point, \textsc{satellite}) \wedge S(split\_point + 1, h, \textsc{nucleus}) \wedge \\
&\qquad (\forall p)(P(l, h, p) \equiv P(split\_point + 1, h, p))\}
\end{aligned}
$$

(2.42)
$$
\begin{aligned}
&\mathit{Nucleus\_first\_ext}(name, split\_point, s_s, s_e, n_s, n_e) \equiv \\
&\quad \{[rhet\_rel\_ext(name, s_s, s_e, n_s, n_e) \wedge T(l, h, name) \wedge \\
&\quad position(s_s, split\_point + 1) \wedge position(s_e, h) \wedge \\
&\quad position(n_s, l) \wedge position(n_e, split\_point) \wedge \\
&\quad \{(name \in Rels_{paratactic}) \rightarrow \\
&\qquad S(l, split\_point, \textsc{nucleus}) \wedge S(split\_point + 1, h, \textsc{nucleus}) \wedge \\
&\qquad (\forall p)[P(l, h, p) \equiv (P(l, split\_point, p) \vee P(split\_point + 1, h, p))]\} \wedge \\
&\quad \{(name \in Rels_{hypotactic}) \rightarrow \\
&\qquad S(l, split\_point, \textsc{nucleus}) \wedge S(split\_point + 1, h, \textsc{satellite}) \wedge \\
&\qquad (\forall p)(P(l, h, p) \equiv P(l, split\_point, p))\}
\end{aligned}
$$

(2.43)
$$
\begin{aligned}
&\mathit{Satellite\_first\_ext}(name, split\_point, s_s, s_e, n_s, n_e) \equiv \\
&\quad \{[rhet\_rel\_ext(name, s_s, s_e, n_s, n_e) \wedge T(l, h, name) \wedge \\
&\quad position(n_s, split\_point + 1) \wedge position(n_e, h) \wedge \\
&\quad position(s_s, l) \wedge position(s_e, split\_point) \wedge \\
&\quad \{(name \in Rels_{paratactic}) \rightarrow \\
&\qquad S(l, split\_point, \textsc{nucleus}) \wedge S(split\_point + 1, h, \textsc{nucleus}) \wedge \\
&\qquad (\forall p)[P(l, h, p) \equiv (P(l, split\_point, p) \vee P(split\_point + 1, h, p))]\} \wedge \\
&\quad \{(name \in Rels_{hypotactic}) \rightarrow \\
&\qquad S(l, split\_point, \textsc{satellite}) \wedge S(split\_point + 1, h, \textsc{nucleus}) \wedge \\
&\qquad (\forall p)(P(l, h, p) \equiv P(split\_point + 1, h, p))\}
\end{aligned}
$$

[1,16]	dominates	[1,3]
[1,16]	dominates	[4,15]
[4,15]	dominates	[5,6]
[4,15]	dominates	[7,14]
[7,14]	dominates	[9,10]
[7,14]	dominates	[11,14]
[11,14]	dominates	[12,14]

Table 2.2: The dominance relations given by Grosz and Sidner with respect to text (2.25).

• **The dominance relations described by Grosz and Sidner hold between a discourse segment and the subordinated satellite.** The dominance relations that are given by Grosz and Sidner with respect to text (2.25) are shown in table 2.2. If we inspect closely the GST representation in figure 2.12, the RST representation in figure 2.13, table 2.1, and table 2.2, we notice that the dominated discourse segments in Grosz and Sidner's enumeration of dominance relations corresponds *always* to the satellite of the RST representation. This is not surprising if we examine the definitions of dominance relation given by Grosz and Sidner and satellite given by Mann and Thompson: a segment DSP_2 dominates a segment DSP_1 if the intention associated with DSP_1 provides part of the satisfaction of the intention associated with DSP_2. In other words, the intention of DSP_1 contributes to the satisfaction of the intention associated with DSP_2. But this is exactly the role that satellites play in Mann and Thompson's theory: they do not express what is most essential for the writer's purpose, but rather, provide supporting information that contributes to the understanding of the nucleus.

The relationship between Grosz and Sidner's dominance relations and the general distinction between nuclei and satellites is formalized by axioms (2.44) and (2.45).

(2.44)
$$\begin{aligned}&[(1 \leq h_1 \leq \textsc{n}) \wedge (1 \leq l_1 \leq h_1) \wedge (1 \leq h_2 \leq \textsc{n}) \wedge (1 \leq l_2 \leq h_2)] \rightarrow\\ &\quad \{[\neg S(l_1, h_1, \textsc{none}) \wedge S(l_2, h_2, \textsc{satellite}) \wedge l_1 \leq l_2 \leq h_2 \leq h_1 \wedge\\ &\quad \neg(\exists l_3, h_3)(l_1 \leq l_3 \leq l_2 \leq h_2 \leq h_3 \leq h_1 \wedge\\ &\qquad\qquad (l_3 \neq l_2 \vee h_3 \neq h_2) \wedge S(l_3, h_3, \textsc{satellite}))] \rightarrow\\ &\qquad dom(l_1, h_1, l_2, h_2)\}\end{aligned}$$

(2.45)
$$\begin{aligned}&[(1 \leq h_1 \leq \textsc{n}) \wedge (1 \leq l_1 \leq h_1) \wedge (1 \leq h_2 \leq \textsc{n}) \wedge\\ &\quad (1 \leq l_2 \leq h_2) \wedge dom(l_1, h_1, l_2, h_2)] \rightarrow\\ &\qquad \neg S(l_1, h_1, \textsc{none}) \wedge S(l_2, h_2, \textsc{satellite})\end{aligned}$$

Axiom (2.44) specifies that if segment $[l_2, h_2]$ is the immediate satellite of segment $[l_1, h_1]$, then there exists a dominance relation between segment $[l_1, h_1]$ and segment $[l_2, h_2]$. Hence, axiom (2.44) explicates the relationship between the structure of discourse and the intentional dominance. In contrast, axiom (2.45) explicates the relationship between intentional dominance and the structure of discourse. That is, if we know that the intention associated with span $[l_1, h_1]$ dominates the intention associated with span $[l_2, h_2]$, then both these spans play an active role in the representation and, moreover, the segment $[l_2, h_2]$ plays a SATELLITE role.

• **The satisfaction-precedence relations described by Grosz and Sidner could be interpreted as paratactic relations that hold between arbitrarily large textual spans.** Nevertheless, as we have seen in the examples discussed in this chapter, the fact that a paratactic relation holds between spans does not imply that there exists a satisfaction-precedence relation at the intentional level between those spans. Therefore, for satisfaction-precedence relations, we will have only one axiom, that shown in (2.46) below.

(2.46)
$$\begin{aligned}&[(1 \leq h_1 \leq \text{N}) \wedge (1 \leq l_1 \leq h_1) \wedge (1 \leq h_2 \leq \text{N}) \wedge \\ &(1 \leq l_2 \leq h_2) \wedge satprec(l_1, h_1, l_2, h_2)] \rightarrow \\ &S(l_1, h_1, \text{NUCLEUS}) \wedge S(l_2, h_2, \text{NUCLEUS})\end{aligned}$$

It specifies that the spans that are arguments of a satisfaction-precedence relation have a NUCLEUS status in the final representation.

2.7.3 Applications of the formalization of text structures and intentions

Consider again the example text (2.3) that we have used through this chapter, which we reproduce in (2.47) for convenience. As we discussed in section 2.6.3, if we assume that an analyst determines that the rhetorical relations given in (2.48) hold between the elementary units of the text, there are five valid RS-trees that correspond to text (2.47). The valid trees were shown in figure 2.11.

(2.47) [No matter how much one wants to stay a non-smoker,[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]]

$$(2.48)\quad \begin{cases} rhet\_rel(\text{JUSTIFICATION}, \text{A}_1, \text{B}_1) \\ rhet\_rel(\text{JUSTIFICATION}, \text{D}_1, \text{B}_1) \\ rhet\_rel(\text{EVIDENCE}, \text{C}_1, \text{B}_1) \\ rhet\_rel(\text{CONCESSION}, \text{D}_1, \text{C}_1) \\ rhet\_rel(\text{RESTATEMENT}, \text{D}_1, \text{A}_1) \end{cases}$$

If we consider now the axioms that describe the relationship between text structures and intentions, we can derive, for example, that, for the tree 2.11.a, the span $[\text{A}_1, \text{D}_1]$ dominates the span $[\text{C}_1, \text{D}_1]$; and that the primary intention of the whole text depends on unit B_1 and on the rhetorical relation of JUSTIFICATION. In such a case, the axiomatization provides the means for drawing intentional inferences on the basis of the discourse structure.

Assume now that besides providing judgements concerning the rhetorical relations that hold between various units, an analyst provides intention-based judgements as well. If, for example, besides the relations given in (2.48) an analyst determines that span $[\text{A}_1, \text{D}_1]$ dominates unit D_1, the theory that corresponds to these judgements (2.49) and the axioms given in section 2.7.2 yields only two valid text structures, those presented in figure 2.11.b and 2.11.d. Therefore, in this case, the axiomatization provides the means of using intentional judgements for reducing the ambiguity that characterizes text structures.

$$(2.49)\quad \begin{cases} rhet\_rel(\text{JUSTIFICATION}, \text{A}_1, \text{B}_1) \\ rhet\_rel(\text{JUSTIFICATION}, \text{D}_1, \text{B}_1) \\ rhet\_rel(\text{EVIDENCE}, \text{C}_1, \text{B}_1) \\ rhet\_rel(\text{CONCESSION}, \text{D}_1, \text{C}_1) \\ rhet\_rel(\text{RESTATEMENT}, \text{D}_1, \text{A}_1) \\ dom(\text{A}_1, \text{D}_1, \text{D}_1, \text{D}_1) \end{cases}$$

2.8 Related work

The formalization that I have presented in this chapter 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 2.2.1 and by the strong compositionality criterion 2.2. As such, the formalization in chapter 2 can be interpreted as a sibling of model-theoretic frameworks that characterize the properties of the syntactic structures of sentences [Keller, 1992, Keller, 1993, Blackburn *et al.*, 1995, Rogers, 1994, Rogers, 1996]. In contrast to 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 our formalization 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, our formalization can be couched in the language of first-order logic.

To my knowledge, the formalization of text structures provided in this chapter is the first attempt to provide a model-theoretic framework for the study of discourse in general and the study of RST, GST, and the relationship between the two. In contrast to the model-theoretic framework that was developed here, most of the current approaches to discourse do not address so much the problem of what discourse structures are, but of how discourse structures can be derived from a given text in the context of discourse analysis [van Dijk, 1972, Polanyi, 1988, Scha and Polanyi, 1988, Lascarides and Asher, 1991, Lascarides *et al.*, 1992, Lascarides and Asher, 1993, Asher and Lascarides, 1994, Gardent, 1994, Polanyi and van den Berg, 1996, van den Berg, 1996, Gardent, 1997, Schilder, 1997, Cristea and Webber, 1997] and from a knowledge base, in the context of natural language generation [Hovy, 1988b, Moore and Swartout, 1991, Moore and Paris, 1993, Maybury, 1993]. I discuss in detail these lines of research in chapters 3 and 7 respectively.

2.9 Summary

In this chapter, I have provided a first-order formalization of valid text structures and a characterization of the relationship between text structures and intentions. The formalization relies on six essential features:

1. The elementary units of complex text structures are non-overlapping spans of text.

2. Rhetorical, coherence, and cohesive relations hold between textual units of various sizes.

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 also holds between the most important units of the constituent spans. The most important units are determined recursively: they correspond to 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.

6. The primary intention of a text span depends on the most salient units of that span and the rhetorical relation that introduced them.

Chapter 3

The automatic derivation of text structures: an algorithmic perspective

3.1 Preamble

The formalization in chapter 2 focuses on the mathematical properties of valid text structures, and not on the mechanisms that can be used to derive such structures. The idea of providing algorithms that derive the valid discourse structures of texts gives rise to two alternatives.

- The first alternative is to take advantage of the declarative formalization and equate 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 amounts to applying 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 these structures.

- The second alternative is to specify rewriting rules that can map a sequence of textual units into valid text structures. This alternative amounts to applying theorem-proving techniques.

  The major benefit of this alternative is that it enables one to control directly the process of text structure derivation. As we will see in section 3.5, such an approach can lead to substantial improvements with respect to the time that is needed to derive the valid structures of a text.

In this chapter, I study both alternatives: I propose and compare empirically four different paradigms for solving the problem of text structure derivation given in 2.2. In two of these paradigms I use model-theoretic techniques, i.e., I show how the problem of text structure derivation can be encoded as a classical constraint-satisfaction problem (section 3.2) and as a propositional, satisfiability problem (section 3.3). In the other two paradigms I apply proof-theoretic techniques, i.e., I show how the problem of text structure derivation can be encoded as a theorem-proving problem (section 3.4) and how it can be compiled into a parsing problem using a grammar in Chomsky normal form (section 3.5). The last paradigm yields the fastest algorithm, which derives text structures in polynomial time.

The empirical comparison of the four paradigms was done on a Sparc Ultra 2–2170 machine that was running in network mode. The implementations of the four paradigms were written in Lisp, C, and C++. As a consequence, it is obvious that the results have little meaning if they are taken in isolation. However, as will become apparent in the following sections, the differences in performance of the four implementations are large enough to provide clear-cut evidence with respect to the paradigm that is best suited for deriving valid text structures.

An adequate account of the relationship between text structures and intentions would require a sophisticated description of the oracle function f_I (see section 2.7.2). Such a description is beyond the scope of this thesis. Therefore, in what follows, I will investigate only the structural properties of discourse. I will rely on the set of rhetorical relations proposed by Mann and Thompson [1988] and consider text structures to be completely described by the axiomatization of RST (see section 2.6.3).

The work presented in this chapter is of primary interest for computer scientists and not for engineers of language. A reader whose interest is only to find out how discourse structures can be derived automatically from unrestricted texts can skip this chapter. All such a reader needs to bear in mind is that the problem of text structure derivation that was given in 2.2 has an algorithmic solution. Hence, in order to derive text structures of unrestricted texts we need only determine the elementary textual units and the rhetorical relations that hold among them.

3.2 Deriving text structures — a constraint-satisfaction approach

The formalization in chapter 2 naturally suggests that text structures can be automatically derived using constraint-satisfaction techniques. As we discussed in section 2.6, if we consider a sequence of textual units $u_1, u_2, \dots, u_{\mathrm{N}}$, there are N ways in which spans of length one could be built, $span_{1,1}, span_{2,2}, \dots, span_{\mathrm{N,N}}$; $\mathrm{N}-1$ ways in which spans of length two could be built, $span_{1,2}, span_{2,3}, \ldots, span_{\text{N}-1,\text{N}}$; $\text{N} - 2$ ways in which spans of length three could be built, $span_{1,3}, span_{2,4}, \ldots, span_{\text{N}-2,\text{N}}$; $\ldots$; and one way in which a span of length N could be built, $span_{1,\text{N}}$. Each of these spans has the potential of playing an active role in the final representation. An algorithm that constructs valid text structures for the sequence $u_1, u_2, \ldots, u_{\text{N}}$ will have to determine from the set of $\text{N}+(\text{N}-1)+(\text{N}-2)+\ldots+1 = \text{N}(\text{N}+1)/2$ potential text spans that pertain to the sequence of N textual units, the subset that adheres to the constraints that characterize valid structures.

As we have seen, the status, type, and promotion set associated with each span provides a complete characterization of the text structure. Following the axiomatization of RST, we can take a sequence of N textual units and the set of rhetorical relations that hold between them and automatically derive a constraint-satisfaction problem with $3\text{N}(\text{N}+1)/2$ variables — a status, a type, and a promotion variable for each of the $\text{N}(\text{N}+1)/2$ potential spans. The algorithm that creates the $3\text{N}(\text{N}+1)/2$ variables and asserts the constraints that pertain to the variables is shown in figure 3.1. In the following two subsections, I will explain it piece by piece.

3.2.1 The constraint variables

To begin with, the algorithm creates the status, type, and promotion constraint variables that are associated with each of the possible $\text{N}(\text{N}+1)/2$ spans of a text structure. In figure 3.1, the constraint variables are represented using the symbols S, T, and P, respectively. The constraint variables are indexed according to the lower and upper bounds of the spans that they correspond to. For example, the variable $S[l, h]$ corresponds to the status of the textual span that ranges between positions l and h.

Lines 1–9 of the algorithm correspond to the creation of the constraint variables and the specification of their associated domains. For each leaf, the domain of a status variable is the set $\{N, S\}$ (NUCLEUS or SATELLITE); the domain of a type variable is $\{\mathit{LEAF}\}$; and the domain of a promotion variable is the unit itself, $\{u_l\}$. For each non-elementary textual span, $l < h$, the domain of a status variable is the set $\{N, S, \mathit{NONE}\}$ (NUCLEUS, SATELLITE, or NONE); the domain of a type variable is given by the names of the relations that are relevant for that span (see axiom (2.8)); and the domain of a promotion variable is the set of textual units that correspond to the span, $\{u_l, \ldots, u_h\}$.

Traditionally, a solution of a constraint-satisfaction problem that is characterized by n variables having domains $D_1, \ldots, D_n$ is a member of the Cartesian product $D_1 \times \ldots \times D_n$. Therefore, if we adopt a constraint-satisfaction perspective, there is no need to explicitly encode the unicity constraints that pertain to the status (axioms (2.11)) and type (axiom (2.13)) of each potential node. Although this is appropriate for status and type variables, the fact that a solution of a constraint-satisfaction problem associates only one value to each variable appears to create difficulties with respect to the promotion variables,

Input: A sequence of textual units $U = u_1, u_2, \ldots, u_N$ and a set RR of simple and extended rhetorical relations that hold between units and spans in U.
Output: One or all valid text structures of U.

```
% Create N(N + 1)/2 status, type, and promotion variables whose domains range over
% the set of values described by axioms (2.10), (2.12) and (2.14) respectively.
for h := 1 to N
    for l := 1 to h
        if (l = h)
          domain(S[l, h]) = {N, S}; domain(T[l, h]) = {LEAF}; domain(P[l, h]) = {u_l};
        else {
            domain(S[l, h]) = {N, S, NONE};
            domain(T[l, h]) = {name(r) | r ∈ relevant_relations(RR, l, h)};
            domain(P[l, h]) = {u_l, ... , u_h};
            }
% Text spans do not overlap (axiom (2.15)).
for h_1 := 1 to N
    for l_1 := 1 to h_1
        for h_2 := 1 to N
            for l_2 := 1 to h_2
                if (l_1 < l_2 ∧ l_2 ≤ h_1 ∧ h_1 < h_2)
                  assert(S[l_1, h_1] = NONE ∨ S[l_2, h_2] = NONE))
% A span with status NONE does not play an active role (axiom (2.16)).
for h := 1 to N
    for l := 1 to h
        assert([S[l, h] ≠ NONE ∧ T[l, h] ≠ NONE ∧ P(l, h) ≠ NONE]∨
               [S[l, h] = NONE ∧ T[l, h] = NONE ∧ P(l, h) = NONE]);
% There exists a root node (axiom (2.17)).
assert(S[1, N] = N ∧ T[1, N] ≠ NONE ∧ P[1, N] ≠ NONE);
% Valid text structures obey the strong compositionality criterion (axioms (2.18),
% and (2.19)–(2.22)).
for size_of_span := 1 to N − 1
    for l := 1 to N− size_of_span
        h := l+ size_of_span;
        % for every span [l, h], 1 ≤ l < h ≤ N
        C := (S[l, h] = NONE);
        for  r ∈ relevant_relations(RR, l, h)
              for  sp from l to h
                  if valid_satellite_first(r, l, sp, h)
                    C := C ∨ {S[l, sp] = S ∧ S[sp + 1, h] = N ∧ T[l, h] = name(r)∧
                              P[l, sp] = sat(r) ∧ P[sp + 1, h] = nucl(r)∧
                              P[l, h] = P[sp + 1, h]};
                  if valid_nucleus_first(r, l, sp, h)
                    C := C ∨ {S[l, sp] = N ∧ S[sp + 1, h] = S ∧ T[l, h] = name(r)∧
                              P[l, sp] = nucl(r) ∧ P[sp + 1, h] = sat(r)∧
                              P[l, h] = P[l, sp]};
                  if valid_multinuclear(r, l, sp, h)
                    C := C ∨ {S[l, sp] = N ∧ S[sp + 1, h] = N ∧ T[l, h] = name(r)∧
                              P[l, sp] = nucl_1(r) ∧ P[sp + 1, h] = nucl_2(r)∧
                              (P[l, h] = P[l, sp] ∨ P[l, h] = P[sp + 1, h])};
        assert(C);
% solve the constraint satisfaction problem
find_solutions();
```

Figure 3.1: A constraint-satisfaction algorithm for deriving text structures

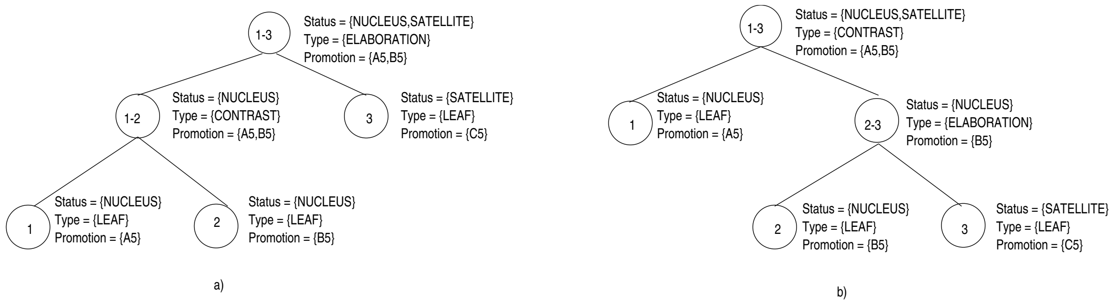

Figure 3.2: The valid text structures of text (3.1).

because a textual span may have more than one salient unit in the cases in which the textual structure is built using multinuclear relations. Fortunately, as I now show, this proves not to be problematic.

In the cases in which no multinuclear relation is used, each node in the final text structure will be characterized by one salient unit. In such a case, there exists a one-to-one mapping between a valid text structure and a solution of the corresponding constraint-satisfaction encoding. Assume now, however, that a text is characterized by multinuclear relations as well. For example, the set of relations that hold between the elementary units in text (3.1) [Mann and Thompson, 1988, p. 278] is shown in (3.2).

(3.1) [Animals heal,^{A_5}] [but trees compartmentalize.^{B_5}] [They endure a lifetime of injury and infection by setting boundaries that resist the spread of the invading microorganisms.^{C_5}]

$$(3.2) \quad \begin{cases} rhet\_rel(\textsc{constrast}, \textsc{a}_5, \textsc{b}_5) \\ rhet\_rel(\textsc{elaboration}, \textsc{c}_5, \textsc{b}_5) \end{cases}$$

There are two valid structures that can be built for text (3.1). In both of them (see figure 3.2), the promotion set of the root node has cardinality two. Let us focus, for the moment, on tree 3.2.a, which has two nodes that are characterized by promotion sets with cardinality larger than one. In a first approximation, it may appear that it is necessary to associate with each node of a text structure all the units that are salient. However, if we examine the definition of the problem of text structure derivation closely (see definition 2.2), it is easy to notice that the rhetorical relations that are given as input hold either between elementary units or between textual spans. The strong compositionality criterion specifies that two textual spans can be put together into a larger span when an elementary relation holds between two units that are salient in the spans, or when an extended relation holds between the spans. Therefore, in order to decide whether two spans can be joined by an elementary relation, we do not need to know all the units that are salient in the spans:

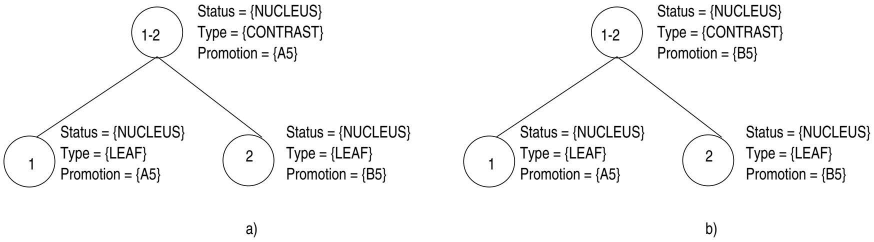

Figure 3.3: Representing multinuclear relations using promotion sets of cardinality one.

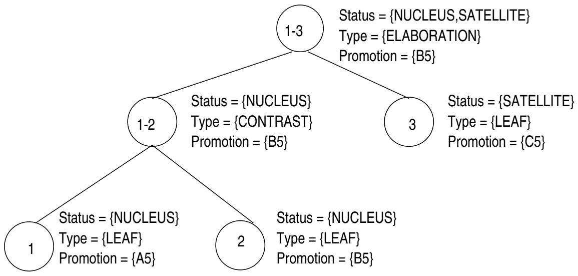

Figure 3.4: A textual structure of text (3.1) that uses only promotion sets of cardinality one.

rather, it is enough to know only whether the units that are arguments of the elementary relation are salient. Moreover, in order to decide whether two spans can be joined by an extended relation, we do not need any information about the salient units of the spans. Hence, if during the construction process we associate only one salient unit with each span — the one that is going to be used further in the tree-building process — we could still build a text structure. It is true that such a structure enforces only partially the strong compositionality criterion; but fortunately, it allows for the recovery of the full valid structure.

To understand better the claim above, let us reconstruct now tree 3.2.a using only promotion sets of cardinality one. To do this, we notice that when two spans are put together using a multinuclear relation, there exist two possible solutions; each solution corresponds to the promotion of only one salient unit. For example, if we put together the elementary units A_5 and B_5 using the CONTRAST relation and allowing the promotion sets of each span to have cardinality at most one, we have two choices (see figure 3.3). The choices correspond to promoting as salient either unit A_5 or unit B_5 for the span $[\mathrm{A}_5, \mathrm{B}_5]$. To complete the reconstruction of tree 3.2.a, we have to use the ELABORATION relation that holds between satellite C_5 and nucleus B_5. Tree 3.3.a cannot be extended into tree 3.2.a because it would violate the strong compositionality criterion (unit B_5 is not a salient unit for span $[\mathrm{A}_5, \mathrm{B}_5]$). However, tree 3.3.b can be extended, thus obtaining a version of tree 3.2.a that uses only promotion sets of cardinality one (see figure 3.4).

Input: A text structure, Tree, that obeys all the axioms of a valid text structure, with the exception of the strong compositionality criterion: the promotion set of each active node in this structure has cardinality one.
Output: A valid text structure.

```
function adjust(Tree)
    if(isleaf(Tree)) return Tree;
    Tree→left := adjust(Tree→left);
    Tree→right := adjust(Tree→right);
    if(type(Tree) = "paratactic")
      promotionSet(Tree) := promotionSet(Tree→left) ∪ promotionSet(Tree→right);
    else if(status(Tree→left) = NUCLEUS)
      promotionSet(Tree) := promotionSet(Tree→left);
    else
      promotionSet(Tree) := promotionSet(Tree→right);
    return Tree;
```

Figure 3.5: A recursive algorithm that maps "almost-valid" text structures into valid ones.

The tree in figure 3.4 is not valid because it obeys only a watered-down version of the strong compositionality criterion: the promotion set of span $[\text{A}_5, \text{B}_5]$ is not the set $\{\text{A}_5, \text{B}_5\}$, but its subset, $\{\text{B}_5\}$. Fortunately, the "almost-valid" tree in figure 3.4 enables recovery of the valid representation; if we traverse the tree bottom-up, we can update the promotion sets that characterize the nodes whose types are multinuclear relations such that the promotion sets become equal to the union of the promotion sets of the immediate subspans; and we can update the promotion sets that characterize the nodes whose types are mononuclear relations such that the promotion sets become equal to the promotion set of the nucleus subspan. If we apply this process to tree 3.4, we obtain tree 3.2.a.

The discussion above suggests that a constraint-satisfaction approach can be used to first build text structures of the kind shown in figure 3.4, i.e., structures that are characterized by promotion sets of cardinality one. These structures can then be mapped into valid ones using a simple bottom-up traversal. Figure 3.5 presents a recursive algorithm that maps a text structure that obeys only the watered-down version of the compositionality criterion into a valid one.

3.2.2 The constraints

Bearing in mind the fact that valid trees can be built using promotion sets of cardinality one, we return now to the algorithm in figure 3.1. Once the variables and their domains have been established, the algorithm asserts the structural constraints that correspond to axioms (2.15) (lines 10–15), (2.16) (lines 16-19), and (2.17) (line 20). Next, the algorithm asserts the constraints that pertain to the strong compositionality criterion (axioms (2.18), and (2.19)–(2.22)), using the assumption that the final solution will use promotion sets of cardinality one — see lines 21–40. The algorithm iterates over each non-elementary textual span $[l, h]$ and builds a constraint C that captures the watered-down version of the strong compositionality criterion. The constraint C rewrites axioms (2.18)–(2.22)) as a disjunction over all possible ways that can lead to that span having a non-NONE status. The algorithm iterates over all relations that are relevant to the span $[l, h]$ (lines 26–40) and over all ways in which span $[l, h]$ can be broken into two subspans: sp (split point) denotes the location between l and h where the span $[l, h]$ can be broken. For each relation r that is relevant to a span $[l, h]$, with respect to a splitting point sp, i.e., either r is a simple rhetorical relation that holds between two units found in the resulting subspans or an extended rhetorical relation that holds between the two immediate subspans, there exist four possibilities:

- The satellite of the relation r goes before the nucleus. In such a case, if r is used to join spans $[l, sp]$ and $[sp + 1, h]$ $(valid\_satellite\_first(r, l, sp, h))$, then the status of span $[l, sp]$ is satellite, the status of span $[sp + 1, h]$ is nucleus, the type of the span $[l, h]$ is given by the name of the relation r, the promotion set of span $[l, h]$ is given by the satellite of the relation, the promotion set of span $[sp+1, h]$ is given by the nucleus of the relation, and the promotion set of the span $[l, h]$ is given by the promotion set of the nucleus $[sp + 1, h]$ (see lines 28–31 in figure (3.1)).

- The nucleus of the relation r goes before the satellite. In such a case, if r is used to join spans $[l, sp]$ and $[sp+1, h]$ $(valid\_nucleus\_first(r, l, sp, h)$, then the status of span $[l, sp]$ is nucleus, the status of span $[sp + 1, h]$ is satellite, the type of the span $[l, h]$ is given by the name of the relation r, the promotion set of span $[l, h]$ is given by the nucleus of the relation, the promotion set of span $[sp + 1, h]$ is given by the satellite of the relation, and the promotion set of the span $[l, h]$ is given by the promotion set of the nucleus $[l, sp]$ (see lines 32–35 in figure (3.1)).

- The relation r is multinuclear. In such a case, if r is used to join spans $[l, sp]$ and $[sp+1, h]$ $(valid\_multinuclear(r, l, sp, h)$, then the status of spans $[l, sp]$ and $[sp+1, h]$ is nucleus, the type of the span $[l, h]$ is given by the name of the relation r, the promotion set of span $[l, h]$ is given by the first nucleus of the relation, the promotion set of span $[sp+1, h]$ is given by the second nucleus of the relation, and the promotion set of the span $[l, h]$ is given *either* by the promotion set of the first nucleus $[l, sp]$ or by the promotion set of the second nucleus $[sp + 1, h]$ (see lines 36–39 in figure (3.1)).

- The relation r does not hold across the splitting point sp, and, therefore, is irrelevant.

Once all the constraints have been asserted, one can apply any constraint-satisfaction algorithm in order to find one or all the solutions that pertain to the text that is considered, and hence one or all its valid text structures (see line 41 in figure 3.1).

The constraint-satisfaction problem that is generated by algorithm 3.1 has $3\text{N}(\text{N}+1)/2$ variables. In a text of N elementary textual units, for every span $[l,h]$, there are $(h-l-1)(\text{N}-h+l)$ spans that overlap that span. Therefore, the total number of constraints shown in line 15 of algorithm 3.1 is $\sum_{2\leq h<\text{N}}\sum_{1\leq l<h}(h-l-1)(\text{N}-h+l) = \text{N}(\text{N}-1)(\text{N}^2+5\text{N}-2)$. The number of constraints that have the form shown in line 19 of algorithm 3.1 is $\sum_{1\leq h\leq \text{N}}\sum_{1\leq l\leq h} 1 = \text{N}(\text{N}+1)/2$. In addition to these constraints, algorithm 3.1 derives one complex disjunctive constraint for each non-elementary span (lines 21–40). Since there are $\text{N}(\text{N}-1)/2$ non-elementary spans, it follows that there are $\text{N}(\text{N}-1)/2$ such constraints. The total number of constraints derived by algorithm 3.1 is, therefore, $1/12(\text{N}^4+4\text{N}^3+5\text{N}^2+2\text{N}+12)$.

3.2.3 Implementation and empirical results

It is well-known that finding solutions of constraint-satisfaction problems is NP-complete in the general case [Mackworth, 1977, Garey and Johnson, 1979]. In spite of this, CS algorithms seem to perform well for certain classes of problems. Determining whether the problem of text structure derivation falls into a class of problems for which CS algorithms perform well enough is an empirical question. To answer it, I used Lisp and Screamer [Siskind and McAllester, 1993a, Siskind and McAllester, 1993b], a macro package that provides constraint-satisfaction facilities, to fully implement a system that builds text structures by means of the algorithm shown in figure 3.1. The implementation takes as input a linear sequence of textual units $U = u_1, u_2, \ldots, u_{\text{N}}$ and the set of simple and extended rhetorical relations that hold among these units. The program follows the algorithm given in figure 3.1 in order to build the corresponding constraint-satisfaction problem. It then uses the built-in facilities of Screamer to find all the possible solutions, i.e., all the valid text structures. A simple procedure prints the text trees that pertain to each solution.

The program was run on eight texts: the simplest has three elementary units among which four rhetorical relations hold; the most complex has 19 elementary units among which 25 rhetorical relations hold. Appendix A contains these texts, their elementary units, and the rhetorical relations that characterize them.

Table 3.1 shows the amounts of time on a Sparc Ultra 2–2170 that were required by our implementation for determining all the valid text structures of these texts. The dashed lines in table 3.1 correspond to computations that did not terminate in less than three hours. Given the results in table 3.1, it is obvious that the performance of algorithm 3.1 is very poor. A close analysis of the behavior of our implementation showed that, in fact, the algorithm spent most of the time in asserting the constraints shown in line 40 in figure 3.1. As the text spans $[l,h]$ get bigger, more relations are relevant for them; as a consequence,

Text	Number of variables	Number of constraints	Time in seconds
A.1	18	21	0.3
A.2	30	51	38.0
A.3	45	106	–
A.4	84	337	–
A.5	84	337	–
A.6	360	5441	–
A.7	513	10831	–
A.8	570	13301	–

Table 3.1: Performance of the constraint-based implementation

the constraints that correspond to a straightforward encoding of the strong compositionality criterion contain more and more complex disjunctive constraints. The macro package that we used tries to reduce the domains of the variables every time a new constraint is asserted. As the spans grow bigger, the time that is taken by Screamer to assert these constraints increases exponentially. It is possible that different constraint-software packages behave better on the problems derived by the algorithm in figure 3.1. Still, I believe that the complexity of the constraints that correspond to the strong compositionality criterion could constitute a challenge for them.

3.3 Deriving text structures — a propositional logic, satisfiability approach

3.3.1 Preamble

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. In this section, I propose a propositional logic encoding of the problem of text structure derivation 2.2 and discuss a program that automatically generates such an encoding starting from the linear sequence of units that is subsumed by a text, and the simple and extended rhetorical relations that hold among these units. In presenting the propositional encoding, I will make use of text (2.3), which, for convenience, is reproduced below in (3.3). To simplify the discussion, the elementary textual units are labelled with natural numbers, from 1 to 4. The simple and extended rhetorical relations that I assume to hold among the textual units in (3.3) are listed in (3.4); rhetorical relations having the same name are given different subscripts in order to enable a clearer presentation of the propositional encoding.

(3.3) [No matter how much one wants to stay a non-smoker,[1]] [the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life.[2]] [We know that 3,000 teens start smoking each day,[3]] [although it is a fact that 90% of them once thought that smoking was something that they'd never do.[4]]

$$(3.4) \quad RR = \begin{cases} rhet\_rel(\text{JUSTIFICATION}_1, 1, 2) \\ rhet\_rel(\text{JUSTIFICATION}_2, 4, 2) \\ rhet\_rel(\text{EVIDENCE}, 3, 2) \\ rhet\_rel(\text{CONCESSION}, 4, 3) \\ rhet\_rel(\text{RESTATEMENT}, 4, 1) \\ rhet\_rel\_ext(\text{JUSTIFICATION}_3, 1, 1, 2, 4) \end{cases}$$

Because I want to estimate the size of the propositional encoding, I assume that at most k rhetorical relations hold between any pair of textual units. During my empirical experiments, I noticed that the number of elementary rhetorical relations that hold over the textual units of a text of size N was never bigger than 3N. Since there are $\binom{\text{N}}{2}$ distinct pairs of units in a text of size N, it follows that a good upper bound for the coefficient k is $3\text{N}/\binom{\text{N}}{2} = 3/[2(\text{N}-1)]$.

In order to fully specify a propositional encoding of the formalization of text structures, we need to specify a set of propositional variables and constraints (propositional formulas) that is logically equivalent with the axiomatization of text structures. I discuss each of these, in turn.

3.3.2 Variables of the propositional encoding

Status variables

As I discussed in section 3.2, there are $\text{N}(\text{N}+1)/2$ potential textual spans that can play an active role in the structure of a text made of N textual units, $u_1, u_2, \ldots, u_\text{N}$. Each potential textual span has a status that can be NUCLEUS, SATELLITE, or NONE. Two propositional variables suffice to encode the 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 a NUCLEUS and 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 $\text{N}(\text{N}+1)/2$ potential spans, it follows that a text of N units will yield $\text{N}(\text{N}+1)$ status variables.

Promotion variables

Each potential span is characterized by a promotion set whose members correspond to the elementary textual units that belong to that span. We associate with each potential span $[l,h]$, $h-l+1$ promotion variables. In order to refer to the promotion variables of a span $[l,h]$, we will use atomic formulas $P_{l,h,i}$, where $l \leq i \leq h$.

Since every span $[l,h]$ is characterized by $h-l+1$ promotion variables, it follows that a text of N units will be characterized by $\text{N} + \sum_{2 \leq h \leq \text{N}} \sum_{1 \leq l < h} (h-l+1) = \text{N}(\text{N}+1)(\text{N}+2)/6$ promotion variables. If a truth assignment assigns the value "true" to any of the promotion variables associated with a span $[l,h]$, the corresponding unit will be considered to be a member of the promotion set of that span. If a truth assignment assigns the value "false" to all the promotion variables associated with span $[l,h]$, we consider the span not to play an active role in the final representation (the status of the span is NONE).

Type variables

Each potential span $[l,h]$ has associated a set of type variables. By convention, the set has cardinality one for the leaves of the text structure. That is, we associate only one propositional variable, $T_{i,i,\text{LEAF}}$, to each elementary unit in the representation. For non-elementary spans $[l,h]$, $l < h$, we associate one propositional variable for each rhetorical relation that is relevant for that span (axiom (2.8)) and one propositional variable to reflect the case in which the span has type NONE. For example, there are three relations that are relevant to span $[2,4]$: $rhet\_rel(\text{JUSTIFICATION}_2, 4, 2)$, $rhet\_rel(\text{EVIDENCE}, 3, 2)$, and $rhet\_rel(\text{CONCESSION}, 4, 3)$. To span $[2,4]$, we will therefore associate four type variables, which we label $T_{2,4,\text{JUSTIFICATION}_2,4,2}$, $T_{2,4,\text{EVIDENCE},3,2}$, $T_{2,4,\text{CONCESSION},4,3}$, and $T_{2,4,\text{NONE}}$. The labelling $T_{l,h,\text{RELATION\_NAME},sat\_pos,nucl\_pos}$ provides a unique identification for each possible rhetorical relation that may end up being used in the text structure representation. We adopt the convention that extended rhetorical relations have associated one type variable, which is labelled $T_{l,h,\text{RELATION\_NAME},sp,sp}$, where sp represents the position at which the extended spans meet. For example, to span $[1,4]$, we will associate one extended type variable $T_{1,4,\text{JUSTIFICATION}_3,1,1}$, which is derived from the extended rhetorical relation $rhet\_rel\_ext(\text{JUSTIFICATION}_3, 1, 1, 2, 4)$. If a truth assignment assigns "true" to any of the non-NONE type variables, we consider the type of the corresponding span to be given by the name of the rhetorical relation that corresponds to the variable. If a truth assignment assigns value "true" to variable $T_{l,h,\text{NONE}}$, the type of the corresponding span is NONE.

In general, for a span $[l, h]$, $l < h$, the number of type variables is given by the sum of relations that are relevant to that span (see axiom (2.8)) and one — the extra variable accounts for the case in which the type is NONE. Essentially, a rhetorical relation is relevant when it holds between two textual units that are found within the boundaries of segment $[l, h]$. Since there are $\binom{h-l+1}{2}$ distinct pairs of elementary textual units within each segment $[l, h]$ and since at most k rhetorical relations hold between any pair, it follows that we associate at most $1 + k\binom{h-l+1}{2}$ variables for every span $[l, h]$. Overall, we associate at most $\sum_{2\leq h\leq \mathrm{N}} \sum_{1\leq l<h}(1 + k\binom{h-l+1}{2}) = \mathrm{N}(\mathrm{N}-1)/2 + k\mathrm{N}(\mathrm{N}-1)(\mathrm{N}+1)(\mathrm{N}+2)/24$ type variables with the non-elementary spans. Hence, the total number of type variables is at most $\mathrm{N}(\mathrm{N}+1)/2 + k\mathrm{N}(\mathrm{N}-1)(\mathrm{N}+1)(\mathrm{N}+2)/24$.

Active-span variables

We associate with every pair of adjacent spans, $[l, sp]$ and $[sp+1, h]$, one active-span variable $A(l, h, sp)$. If a truth assignment assigns the value "true" to a variable $A(l, h, sp)$, it means that both spans $[l, sp]$ and $[sp+1, h]$ play an active role in the text structure and, moreover, that they are the immediate subspans of the span $[l, h]$. If a truth assignment assigns the value "false" to a variable $A(l, h, sp)$, it means that spans $[l, sp]$ and $[sp + 1, h]$ are not the immediate subspans of the span $[l, h]$ in the text structure.

Since every span $[l, h]$ has $h - l$ possible locations at which it can be broken into two adjacent subspans, $l, l+1, \ldots, h-1$, it follows that the total number of active-span variables that characterize a text with N units is $\mathrm{N} + \sum_{2\leq h\leq \mathrm{N}} \sum_{1\leq l<h}(h - l) = \mathrm{N}(\mathrm{N}^2 + 5)/6$.

Discussion

It is possible to provide a propositional formulation of the problem of text structure derivation using only status, promotion, and type variables. The reason I use active-span variables is that they enable a simpler propositional encoding in conjunctive normal form than an encoding that uses only status, promotion, and type variables. If no active-span variables were used, a straightforward encoding of the strong compositionality criterion would yield an exponential number of conjunctive-normal-form formulas. By using active-span variables, the conjunctive-normal-form encoding is polynomial both in the number of variables and number of constraints. If we sum up all the propositional variables that are necessary to encode the text structure of a text with N units, we obtain at most $O(\mathrm{N}^3)$ variables. In what follows, we will see that the propositional encoding proposed here requires at most $O(\mathrm{N}^5)$ conjunctive-normal-form formulas.

3.3.3 Constraints on the variables

In presenting the constraints that pertain to a propositional encoding I adopt an approach similar to that used in section 2.6.2, i.e., I first present the constraints that pertain to the individual spans and variables and then the constraints that pertain to the overall structure of texts. Because most existing software packages that find solutions to propositional satisfiability problems assume that the input is given in conjunctive normal form, and because my intent is to evaluate empirically the suitability of these packages for finding valid discourse structures, I present the constraints as conjuncts of simple and negated disjuncts.

Constraints on the status variables

• **Each leaf of the final representation has either status "nucleus" or "satellite" — the status of a leaf cannot be "none".** 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 $S_{i,i,\textsc{nucleus}}$ and $S_{i,i,\textsc{satellite}}$. Because there are $\textsc{n}$ leaves, this constraint yields $\textsc{n}$ formulas that employ the schema shown in (3.5), where $i = 1, \ldots, \textsc{n}$, and $\textsc{n}$ formulas that employ the schema shown in (3.6), where $i = 1, \ldots, \textsc{n}$.

(3.5) $$S_{i,i,\textsc{nucleus}} \vee S_{i,i,\textsc{satellite}}$$

(3.6) $$\neg S_{i,i,\textsc{nucleus}} \vee \neg S_{i,i,\textsc{satellite}}$$

• **The status of each non-elementary span $[l, h]$, $l < h$, is "nucleus", "satellite", or "none".** For each non-elementary span $[l, h]$, this gives one constraint that employs the schema shown in (3.7). Because there are $\textsc{n}(\textsc{n}-1)/2$ non-elementary spans, it follows that there are $\textsc{n}(\textsc{n}-1)/2$ such constraints.

(3.7) $$\neg S_{l,h,\textsc{nucleus}} \vee \neg S_{l,h,\textsc{satellite}}$$

Constraints on the promotion variables

• **The promotion set associated with each leaf has cardinality one: it consists of the leaf under consideration.** This constraint is encoded by employing $\textsc{n}$ times the schema shown in (3.8), for $i = 1, \ldots, \textsc{n}$.

(3.8) $$P_{i,i,i}$$

Constraints on the active-span variables

• **By convention, in any model of the text structure, the active-span variable associated with each leaf is "true".** This constraint is encoded by employing $\textsc{n}$ times the schema shown in (3.9), for $i = 1, \ldots, \text{N}$.

$$A_{i,i,i} \tag{3.9}$$

Constraints on the type variables

• **The type associated with each leaf is "leaf".** The encoding of this constraint yields N formulas that employ schema (3.10), for $i = 1, \ldots, \text{N}$.

$$T_{i,i,\text{LEAF}} \tag{3.10}$$

• **The type associated with a non-elementary span $[l,h]$ is given either by the name of a relation that is relevant to that span** (2.8) **or is "none".** Since there are $\text{N}(\text{N}-1)/2$ non-elementary spans, this yields $\text{N}(\text{N}-1)/2$ formulas that have the schema given in (3.11), where $M = k\binom{h-l+1}{2}$ is the number of rhetorical relations that are relevant to span $[l,h]$.

$$T_{l,h,\text{NONE}} \vee T_{l,h,\text{NAME}_1,i_1,j_1} \vee \ldots \vee T_{l,h,\text{NAME}_M,i_M,j_M} \tag{3.11}$$

• **The type of each node is unique.** This constraint can be expressed as an exclusive "or" over the propositional variables in (3.11). When the exclusive "or" is written in conjunctive normal form, each non-elementary span $[l,h]$, yields $M(M+1)/2$ constraints that employ the schema given in (3.12), where $1 \le u \le M \wedge 1 \le v \le M \wedge u \ne v$, and M constraints that employ the schema given in (3.13), where $1 \le u \le M$.

$$\neg T_{l,h,\text{NAME}_u,i_u,j_u} \vee \neg T_{l,h,\text{NAME}_v,i_v,j_v} \tag{3.12}$$

$$\neg T_{l,h,\text{NONE}} \vee \neg T_{l,h,\text{NAME}_u,i_u,j_u} \tag{3.13}$$

The total number of binary constraints that employ schema (3.12) is given in (3.14), below.

$$\begin{aligned}\sum_{2\le h\le \text{N}} \sum_{1\le l<h} M(M+1)/2 &= \sum_{2\le h\le \text{N}} \sum_{1\le l<h} k\binom{h-l+1}{2}(k\binom{h-l+1}{2}+1)/2 \\ &= k\text{N}(\text{N}-1)(\text{N}+1)(\text{N}+2)(k\text{N}^2+k\text{N}+5-k)/120.\end{aligned} \tag{3.14}$$

The total number of binary constraints that employ schema (3.13) is given in (3.15), below.

$$\begin{aligned}\sum_{2\le h\le \text{N}} \sum_{1\le l<h} M &= \sum_{2\le h\le \text{N}} \sum_{1\le l<h} k\binom{h-l+1}{2} \\ &= k\text{N}(\text{N}-1)(\text{N}+1)(\text{N}+2)/24.\end{aligned} \tag{3.15}$$

• **Each rhetorical relation can be used to join at most two adjacent spans.** In the encoding that I proposed, the relations that are relevant to a span $[l_1, h_1]$ are also relevant to any span $[l_2, h_2]$ such that $[l_1, h_1]$ is a subspan of $[l_2, h_2]$, $1 \leq l_2 \leq l_1 < h_1 \leq h_2 \leq \textsc{n}$. When we construct a text structure, we do not want to use the same relation twice. To avoid this, for every two spans $[l_1, h_1]$ and $[l_2, h_2]$ that overlap, $1 \leq l_2 \leq l_1 < h_1 \leq h_2 \leq \textsc{n}$, if $rhet\_rel(\textsc{name}, i, j)$ is relevant to both spans, we specify that $T_{l_1,h_1,\textsc{name},i,j} \rightarrow \neg T_{l_2,h_2,\textsc{name},i,j}$. In conjunctive normal form, for each pairs of spans $[l_1, h_1]$ and $[l_2, h_2]$ such that $1 \leq l_2 \leq l_1 < h_1 \leq h_2 \leq \textsc{n}$ and for each relation that is common to them, we specify one constraint that employs schema (3.16).

(3.16)
$$\neg T_{l_1,h_1,\textsc{name},i,j} \vee \neg T_{l_2,h_2,\textsc{name},i,j}$$

For every span $[l_1, h_1]$, there exist $l_1(\textsc{n} - h_1) - 1$ spans $[l_2, h_2]$ such that $1 \leq l_2 \leq l_1 < h_1 \leq h_2 \leq \textsc{n}$. The average number of relations that are relevant to span $[l_1, h_1]$ is $k\binom{h_1-l_1+1}{2}$. Therefore, the average number of constraints that employ schema (3.16) is $k(l_1(\textsc{n} - h_1) - 1)\binom{h_1-l_1+1}{2}$. For the whole encoding, the total number of constraints is

(3.17)
$$\sum_{2 \leq h_1 \leq \textsc{n}} \sum_{1 \leq l_1 < h_1} k(l_1(\textsc{n} - h_1) - 1)\binom{h_1 - l_1 + 1}{2} =$$
$$k\textsc{n}(\textsc{n} - 1)(\textsc{n} + 1)(\textsc{n} + 2)(\textsc{n}^2 + \textsc{n} - 36)/720.$$

3.3.4 Constraints on the overall structure

• **Text spans do not overlap.** For each pair of spans $[l_1, h_1]$ and $[l_2, h_2]$ that overlap, i.e., $l_1 < l_2 \leq h_1 < h_2$, we need to specify a constraint having the form $(S_{l_1,h_1,\textsc{nucleus}} \vee S_{l_1,h_1,\textsc{satellite}}) \rightarrow (\neg S_{l_2,h_2,\textsc{nucleus}} \wedge \neg S_{l_2,h_2,\textsc{satellite}})$. The constraint specifies that when span $[l_1, h_1]$ is active, span $[l_2, h_2]$ is not. When we write the constraint in conjunctive normal form, we obtain four binary constraints that employ schemata (3.18)–(3.21).

(3.18) $\quad \neg S_{l_1,h_1,\textsc{nucleus}} \vee \neg S_{l_2,h_2,\textsc{nucleus}}$

(3.19) $\quad \neg S_{l_1,h_1,\textsc{nucleus}} \vee \neg S_{l_2,h_2,\textsc{satellite}}$

(3.20) $\quad \neg S_{l_1,h_1,\textsc{satellite}} \vee \neg S_{l_2,h_2,\textsc{nucleus}}$

(3.21) $\quad \neg S_{l_1,h_1,\textsc{satellite}} \vee \neg S_{l_2,h_2,\textsc{satellite}}$

In a text of N units, for every span $[l, h]$ there are $(h-l-1)(\textsc{n}-h+l)$ spans that overlap span $[l, h]$. Therefore, the total number of overlapping spans is $\sum_{2 \leq h < \textsc{n}} \sum_{1 \leq l < h} (h - l - 1)(\textsc{n} - h + l) = \textsc{n}(\textsc{n} - 1)(\textsc{n}^2 + 5\textsc{n} - 2)/12$. It follows that the total number of binary constraints employing each of the schemata (3.18)–(3.21) is $\textsc{n}(\textsc{n} - 1)(\textsc{n}^2 + 5\textsc{n} - 2)/12$.

• **A text span with status "none" has the type and promotion "none" as well.** For every span $[l, h]$, this can be expressed as shown in (3.22) below.

$$(3.22) \qquad (\neg S_{l,h,\text{NUCLEUS}} \wedge \neg S_{l,h,\text{SATELLITE}}) \rightarrow (T_{l,h,\text{NONE}} \wedge \neg P_{l,h,l} \wedge \neg P_{l,h,l+1} \wedge \ldots \wedge \neg P_{l,h,h})$$

When we write constraint (3.22) in conjunctive normal form, we obtain one ternary constraint that employs schema (3.23) and $h-l+1$ ternary constraints that employ schema (3.24), where $i = l, \ldots, h$.

$$(3.23) \qquad S_{l,h,\text{NUCLEUS}} \vee S_{l,h,\text{SATELLITE}} \vee T_{l,h,\text{NONE}}$$

$$(3.24) \qquad S_{l,h,\text{NUCLEUS}} \vee S_{l,h,\text{SATELLITE}} \vee \neg P_{l,h,i}$$

It follows that the total number of constraints that employ the schema (3.23) is $\text{N}(\text{N}-1)/2$, and the total number of constraints that employ schema (3.24) is $\sum_{2 \le h \le \text{N}} \sum_{1 \le l < h} (h - l + 1) = \text{N}(\text{N}-1)(\text{N}+4)/6$.

• **A text span with non-"none" status has neither type "none" nor promotion "none".** For every span $[l, h]$, this can be expressed as shown in (3.25) below.

$$(3.25) \qquad (S_{l,h,\text{NUCLEUS}} \vee S_{l,h,\text{SATELLITE}}) \rightarrow (\neg T_{l,h,\text{NONE}} \wedge (P_{l,h,l} \vee P_{l,h,l+1} \vee \ldots \vee P_{l,h,h}))$$

When we write this constraint in conjunctive normal form, we obtain four constraints, each employing one of the schemata (3.26)– (3.29).

$$(3.26) \qquad \neg S_{l,h,\text{NUCLEUS}} \vee \neg T_{l,h,\text{NONE}}$$

$$(3.27) \qquad \neg S_{l,h,\text{SATELLITE}} \vee \neg T_{l,h,\text{NONE}}$$

$$(3.28) \qquad \neg S_{l,h,\text{NUCLEUS}} \vee P_{l,h,l} \vee \ldots \vee P_{l,h,h}$$

$$(3.29) \qquad \neg S_{l,h,\text{SATELLITE}} \vee P_{l,h,l} \vee \ldots \vee P_{l,h,h}$$

It follows that the total number of constraints that employ each of the schemata (3.26)–(3.29) is $\text{N}(\text{N}-1)/2$.

• **The text structure has a root.** In conjunctive normal form, this is expressed by four constraints. They express that the status of the root is either NUCLEUS or SATELLITE (3.30); that the type of the root is not NONE (3.31); that the promotion set of the root has cardinality at least one (3.32); and that there exist two immediate subspans of the root that play an active role in the representation (3.33).

(3.30) $$S_{1,\mathrm{N},\mathrm{NUCLEUS}} \vee S_{1,\mathrm{N},\mathrm{SATELLITE}}$$

(3.31) $$\neg T_{1,\mathrm{N},\mathrm{NONE}}$$

(3.32) $$P_{1,\mathrm{N},1} \vee P_{1,\mathrm{N},2} \vee \ldots \vee P_{1,\mathrm{N},\mathrm{N}}$$

(3.33) $$A_{1,\mathrm{N},1} \vee A_{1,\mathrm{N},2} \vee \ldots \vee A_{1,\mathrm{N},\mathrm{N}-1}$$

- **The text structure obeys the strong compositionality criterion.** We provide a propositional encoding of the strong compositionality criterion by considering, for each textual span $[l, h]$ that can play an active role in the final text structure, three cases in turn:

Case 1. The relation that gives the type of span $[l, h]$ is mononuclear and the satellite comes before the nucleus.

Case 2. The relation that gives the type of span $[l, h]$ is mononuclear and the nucleus comes before the satellite.

Case 3. The relation that gives the type of span $[l, h]$ is multinuclear.

Case 1. Assume first that the relation that gives the type of span $[l, h]$ is mononuclear and the satellite comes before the nucleus. In other words, assume that there exist two subspans $[l, b]$ and $[b+1, h]$ such that a mononuclear relation holds between a satellite i that belongs to span $[l, b]$ and a nucleus j that belongs to span $[b+1, h]$. In such a case, the strong compositionality criterion can be expressed as a conjunction of two formulas. The first conjunct (3.34) specifies that if a relation NAME holds between a satellite $i \in [l, b]$ and a nucleus $j \in [b+1, h]$, then the whole span $[l, b]$ has status SATELLITE, and the whole span $[b+1, h]$ has status NUCLEUS.

(3.34) $$(T_{l,h,\mathrm{NAME},i,j} \wedge \neg T_{l,b,\mathrm{NONE}} \wedge \neg T_{b+1,h,\mathrm{NONE}}) \rightarrow (S_{l,b,\mathrm{SATELLITE}} \wedge S_{b+1,h,\mathrm{NUCLEUS}})$$

When we write formula (3.34) in conjunctive normal form, we obtain for each b such that $l \leq b < h$, two formulas: one that employs schema (3.35); and one that employs schema (3.36).

(3.35) $$\neg T_{l,h,\mathrm{NAME},i,j} \vee T_{l,b,\mathrm{NONE}} \vee T_{b+1,h,\mathrm{NONE}} \vee S_{l,b,\mathrm{SATELLITE}}$$

(3.36) $$\neg T_{l,h,\mathrm{NAME},i,j} \vee T_{l,b,\mathrm{NONE}} \vee T_{b+1,h,\mathrm{NONE}} \vee S_{b+1,h,\mathrm{NUCLEUS}}$$

The second conjunct (3.37) specifies that if a relation NAME holds between a satellite $i \in [l, b]$ and a nucleus $j \in [b+1, h]$, then i is a promotion unit for span $[l, b]$; j is a promotion unit for span $[b+1, h]$; the promotion set of span $[l, h]$ is equivalent to the promotion set of span $[b+1, h]$; and, moreover, none of the units in the satellite $[l, b]$ is a promotion unit for the whole span $[l,h]$.

$$(3.37)\quad (T_{l,h,\text{NAME},i,j} \wedge S_{l,b,\text{SATELLITE}} \wedge S_{b+1,h,\text{NUCLEUS}}) \rightarrow \\ [P_{l,b,i} \wedge P_{b+1,h,j} \wedge (\forall x \in [b+1,h])(P_{b+1,h,x} \equiv P_{l,h,x}) \wedge (\forall x \in [l,b])(\neg P(l,h,x))]$$

When we write formula (3.37) in conjunctive normal form, we obtain for each b such that $l \leq b < h$ one formula that employs schema (3.38); one formula that employs schema (3.39); $h-b$ formulas that employ schema (3.40) for $x = b+1,\ldots,h$; $h-b$ formulas that employ schema (3.41) for $x = b+1,\ldots,h$; and $b-l+1$ formulas that employ schema (3.42) for $x = l,\ldots,b$.

$$(3.38)\quad \neg T_{l,h,\text{NAME},i,j} \vee \neg S_{l,b,\text{SATELLITE}} \vee \neg S_{b+1,h,\text{NUCLEUS}} \vee P_{l,b,i}$$

$$(3.39)\quad \neg T_{l,h,\text{NAME},i,j} \vee \neg S_{l,b,\text{SATELLITE}} \vee \neg S_{b+1,h,\text{NUCLEUS}} \vee P_{b+1,h,j}$$

$$(3.40)\quad \neg T_{l,h,\text{NAME},i,j} \vee \neg S_{l,b,\text{SATELLITE}} \vee \neg S_{b+1,h,\text{NUCLEUS}} \vee \neg P_{b+1,h,x} \vee P_{l,h,x}$$

$$(3.41)\quad \neg T_{l,h,\text{NAME},i,j} \vee \neg S_{l,b,\text{SATELLITE}} \vee \neg S_{b+1,h,\text{NUCLEUS}} \vee P_{b+1,h,x} \vee \neg P_{l,h,x}$$

$$(3.42)\quad \neg T_{l,h,\text{NAME},i,j} \vee \neg S_{l,b,\text{SATELLITE}} \vee \neg S_{b+1,h,\text{NUCLEUS}} \vee \neg P_{l,b,x} \vee \neg P_{l,h,x}$$

For each span $[l,h]$ there are $h-l$ ways to choose the splitting point $b \in [l,h]$. If we assume that the relations that have the satellite before the nucleus, the relations that have the nucleus before the satellite, and the relations that are multinuclear are equally distributed, it follows that the total number of formulas that employ schema (3.35) is $k/3\binom{h-l+1}{2}(h-l)$. For the whole structure, the number of constraints that employ each of the schemata (3.35)–(3.39) is at most

$$(3.43)\quad \sum_{2 \leq h \leq \text{N}} \sum_{1 \leq l < h} k/3\binom{h-l+1}{2}(h-l) = \\ k\text{N}(\text{N}-1)(3\text{N}-1)(\text{N}+1)(\text{N}+2)/360.$$

As we mentioned above, for each span $[l,h]$ there are $h-l$ ways to choose the splitting point $b \in [l,h]$. When $b = l$, there are $h-l$ units x that can be salient in the nucleus span; when $b = l+1$, there are $h-l-1$ units that can be salient in the nucleus span; and so on, when $b = h-1$, there is only one unit that can be salient in the nucleus span. It follows that for a span $[l,h]$, the number of constraints that employ schema (3.40) is given by $k/3\sum_{1 \leq b < h-l} b(h-l-b+1)(h-l)$. Hence, the number of constraints that employ schema (3.40) for the whole text is at most

$$(3.44)\quad k\sum_{2 \leq h \leq \text{N}} \sum_{1 \leq l < h} \sum_{1 \leq b < h-l} b(h-l-b+1)(h-l) = \\ k\text{N}(\text{N}-1)(\text{N}-2)(\text{N}+1)(2\text{N}^2+13\text{N}+3)/1080.$$

The number of constraints that employ schema (3.41) is the same. Reasoning similarly, we can determine that the number of constraints that employ schema (3.42) is at most

$$k/3 \sum_{2\leq h\leq \text{N}} \sum_{1\leq l<h} \sum_{1\leq b<h-l} b^2(h-l) = \\ k\text{N}(\text{N}-2)(4\text{N}+3)(\text{N}+1)(\text{N}-1)/1080. \tag{3.45}$$

Constraints (3.35)–(3.42) account for the cases in which a simple rhetorical relation holds between a satellite i that belongs to a span $[l,b]$ and a nucleus j that belongs to the adjacent span $[b{+}1,h]$. In the case there is an extended rhetorical relation that holds between the two spans, the constraints that pertain to the strong compositionality criterion are captured by two formulas. The first formula, $T_{l,h,\text{NAME},b,b} \rightarrow (S_{l,b,\text{SATELLITE}} \wedge S_{b+1,h,\text{NUCLEUS}})$, specifies that if an extended relation $rhet\_rel\_ext(\text{NAME}, l, h, b, b)$ holds between spans $[l,b]$ and $[b+1,h]$, then the status of the first span is SATELLITE, and the status of the second span is NUCLEUS. This formula yields at most $\sum_{2\leq h\leq \text{N}} \sum_{1\leq l<h}(h-l) = \text{N}(\text{N}-1)(4\text{N}+1)/6$ applications of schemata (3.46) and (3.47) respectively.

$$\neg T_{l,h,\text{NAME},b,b} \vee S_{l,b,\text{SATELLITE}} \tag{3.46}$$

$$\neg T_{l,h,\text{NAME},b,b} \vee S_{b+1,h,\text{NUCLEUS}} \tag{3.47}$$

In addition, the strong compositionality criterion requires the applications of schemata (3.48)–(3.50), which are a shorter expression of schemata (3.40)–(3.42). The number of constraints that characterize the applications of schemata (3.48)–(3.50) is the same as in the case of schemata (3.40)–(3.42).

$$\neg T_{l,h,\text{NAME},b,b} \vee \neg P_{b+1,h,x} \vee P_{l,h,x} \tag{3.48}$$

$$\neg T_{l,h,\text{NAME},b,b} \vee P_{b+1,h,x} \vee \neg P_{l,h,x} \tag{3.49}$$

$$\neg T_{l,h,\text{NAME},b,b} \vee \neg P_{l,b,x} \vee \neg P_{l,h,x} \tag{3.50}$$

Case 2. The constraints that characterize the cases in which a simple or extended rhetorical relation holds between a satellite that comes after the nucleus are analogous in form and number with the constraints that I described above in (3.35)–(3.50). For the purpose of completeness, I only enumerate them here. In schemata (3.51)–(3.62) I assume that unit j belongs to span $[b+1,h]$, and unit i belongs to span $[l,b]$.

$$\neg T_{l,h,\text{NAME},j,i} \vee T_{l,b,\text{NONE}} \vee T_{b+1,h,\text{NONE}} \vee S_{l,b,\text{NUCLEUS}} \tag{3.51}$$

$$\neg T_{l,h,\text{NAME},j,i} \vee T_{l,b,\text{NONE}} \vee T_{b+1,h,\text{NONE}} \vee S_{b+1,h,\text{SATELLITE}} \tag{3.52}$$

$$\neg T_{l,h,\text{NAME},j,i} \vee \neg S_{l,b,\text{NUCLEUS}} \vee \neg S_{b+1,h,\text{SATELLITE}} \vee P_{l,b,i} \tag{3.53}$$

$$\neg T_{l,h,\text{NAME},j,i} \lor \neg S_{l,b,\text{NUCLEUS}} \lor \neg S_{b+1,h,\text{SATELLITE}} \lor P_{b+1,h,j} \tag{3.54}$$

$$\neg T_{l,h,\text{NAME},j,i} \lor \neg S_{l,b,\text{NUCLEUS}} \lor \neg S_{b+1,h,\text{SATELLITE}} \lor \neg P_{l,b,x} \lor P_{l,h,x} \tag{3.55}$$

$$\neg T_{l,h,\text{NAME},j,i} \lor \neg S_{l,b,\text{NUCLEUS}} \lor \neg S_{b+1,h,\text{SATELLITE}} \lor P_{l,b,x} \lor \neg P_{l,h,x} \tag{3.56}$$

$$\neg T_{l,h,\text{NAME},j,i} \lor \neg S_{l,b,\text{NUCLEUS}} \lor \neg S_{b+1,h,\text{SATELLITE}} \lor \neg P_{b+1,h,x} \lor \neg P_{l,h,x} \tag{3.57}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor S_{l,b,\text{NUCLEUS}} \tag{3.58}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor S_{b+1,h,\text{SATELLITE}} \tag{3.59}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor \neg P_{l,b,x} \lor P_{l,h,x} \tag{3.60}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor P_{l,b,x} \lor \neg P_{l,h,x} \tag{3.61}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor \neg P_{b+1,h,x} \lor \neg P_{l,h,x} \tag{3.62}$$

Case 3. The constraints that characterize the cases in which a simple or extended multinuclear rhetorical relation holds across spans $[l, b]$ and $[b+1, h]$ are similar to the constraints that I described above in (3.35)–(3.50). For the purpose of completeness, I enumerate them here as well. In schemata (3.63)–(3.76) I assume that unit i belongs to span $[l, b]$, and unit j belongs to span $[b+1, h]$.

$$\neg T_{l,h,\text{NAME},i,j} \lor T_{l,b,\text{NONE}} \lor T_{b+1,h,\text{NONE}} \lor S_{l,b,\text{NUCLEUS}} \tag{3.63}$$

$$\neg T_{l,h,\text{NAME},i,j} \lor T_{l,b,\text{NONE}} \lor T_{b+1,h,\text{NONE}} \lor S_{b+1,h,\text{NUCLEUS}} \tag{3.64}$$

$$\neg T_{l,h,\text{NAME},i,j} \lor \neg S_{l,b,\text{NUCLEUS}} \lor \neg S_{b+1,h,\text{NUCLEUS}} \lor P_{l,b,i} \tag{3.65}$$

$$\neg T_{l,h,\text{NAME},i,j} \lor \neg S_{l,b,\text{NUCLEUS}} \lor \neg S_{b+1,h,\text{NUCLEUS}} \lor P_{b+1,h,j} \tag{3.66}$$

$$\neg T_{l,h,\text{NAME},i,j} \lor \neg S_{l,b,\text{NUCLEUS}} \lor \neg S_{b+1,h,\text{NUCLEUS}} \lor \neg P_{b+1,h,x} \lor P_{l,h,x} \tag{3.67}$$

$$\neg T_{l,h,\text{NAME},i,j} \lor \neg S_{l,b,\text{NUCLEUS}} \lor \neg S_{b+1,h,\text{NUCLEUS}} \lor P_{b+1,h,x} \lor \neg P_{l,h,x} \tag{3.68}$$

$$\neg T_{l,h,\text{NAME},i,j} \lor \neg S_{l,b,\text{NUCLEUS}} \lor \neg S_{b+1,h,\text{NUCLEUS}} \lor P_{l,b,x} \lor \neg P_{l,h,x} \tag{3.69}$$

$$\neg T_{l,h,\text{NAME},i,j} \lor \neg S_{l,b,\text{NUCLEUS}} \lor \neg S_{b+1,h,\text{NUCLEUS}} \lor \neg P_{l,b,x} \lor P_{l,h,x} \tag{3.70}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor S_{l,b,\text{NUCLEUS}} \tag{3.71}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor S_{b+1,h,\text{NUCLEUS}} \tag{3.72}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor \neg P_{b+1,h,x} \lor P_{l,h,x} \tag{3.73}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor P_{b+1,h,x} \lor \neg P_{l,h,x} \tag{3.74}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor P_{l,b,x} \lor \neg P_{l,h,x} \tag{3.75}$$

$$\neg T_{l,h,\text{NAME},b,b} \lor \neg P_{l,b,x} \lor P_{l,h,x} \tag{3.76}$$

The constraints described in (3.35)–(3.76) explain mostly how a text structure grows bottom-up, i.e., they explain the way the promotion sets are computed. In order to specify completely the strong compositionality criterion, we also need to explain how a discourse structure grows top-down. We do this by specifying the constraints on the active-span

Input: A sequence of textual units $U = u_1, u_2, \ldots, u_{\text{N}}$ and a set RR of simple and extended rhetorical relations that hold between units and spans in U.
Output: One valid text structure of U.

1. Create propositional variables as given in section 3.3.2 and assign them unique natural numbers.
2. Derive the set of conjunctive-normal-form constraints discussed in sections 3.3.3 and 3.3.4. Use as variables names the natural numbers that correspond to them.
3. Find a model of the logical theory derived in step 2.
4. Reconstruct the text structure that corresponds to that model.

Figure 3.6: A propositional logic, satisfiability algorithm for deriving text structures

variables.

If two adjacent spans $[l, sp]$ and $[sp+1, h]$ play an active role in the final representation and are the immediate subspans of span $[l, h]$, then their type is not NONE. The formalization of this constraint, $A_{l,h,sp} \rightarrow (\neg T_{l,sp,\text{NONE}} \vee \neg T_{sp+1,h,\text{NONE}} \vee \neg T_{l,h,\text{NONE}})$, yields three conjunctive normal form schemata, which are shown below.

(3.77) $$\neg A_{l,h,sp} \vee \neg T_{l,sp,\text{NONE}}$$

(3.78) $$\neg A_{l,h,sp} \vee \neg T_{sp+1,h,\text{NONE}}$$

(3.79) $$\neg A_{l,h,sp} \vee \neg T_{l,h,\text{NONE}}$$

Assume again that a mononuclear relation holds in the final structure between two units i, j, such that $i < j$. In such a case, there must exist a splitting point $b \in [i, j-1]$ such that both spans $[l, b]$ and $[b+1, h]$ play an active role in the final representation. The expression of this fact, $T_{l,h,\text{NONE},i,j} \rightarrow (A_{l,h,i} \vee A_{l,h,i+1} \vee \ldots \vee A_{l,h,j-1})$, yields one constraint for each span $[l, h]$, which has the schema shown in (3.80). The number of constraints having this form is $\sum_{2 \leq h \leq \text{N}} \sum_{1 \leq l < h} k\binom{h-l+1}{2} = k\text{N}(\text{N}-1)(\text{N}-1)(\text{N}+1)(\text{N}+2)/24$.

(3.80) $$\neg T_{l,h,\text{NAME},i,j} \vee A_{l,h,i} \vee A_{l,h,i+1} \vee \ldots \vee A_{l,h,j-1}$$

The status, promotion, active-span, and type constraints described in this section and the constraint schemata (3.5)–(3.80) provide a propositional, conjunctive-normal-form encoding of the valid text structures. If we assume that $k = 3\text{N}/\binom{\text{N}}{2} = 3/[2(\text{N}-1)]$ is an adequate approximation of the largest number of rhetorical relations that hold among the units of a text of N units and we sum up the number of constraints described in (3.5)–(3.80), we obtain a figure in the $O(\text{N}^5)$ range. Hence, the size of the propositional encoding of the problem of text structure derivation with respect to a text of N elementary units consists of at most $O(\text{N}^3)$ variables and at most $O(\text{N}^5)$ conjunctive-normal-form constraints.

3.3.5 Algorithm, implementation, and empirical results

Algorithm

The automatic derivation of the variables and the conjunctive-normal-form constraints of the propositional encoding of the valid text structures that pertain to a text can follow the same steps that we took in their presentation (see figure 3.6). Given an input similar to that shown in (3.4), we can determine all the variables and constraints of the valid structures that correspond to (3.4) through a trivial iterative process that considers all the possible spans and pairs of spans that can be built on units $1, 2, \ldots, \text{N}$, and all the rhetorical relations that are relevant to these spans. Because most off-the-shelf software packages that find models of logical theories represented in conjunctive normal form assume that the input is given as a sequence of disjunctions in which the non-negated variables are represented using positive integers and negated variables using negative integers, the propositional algorithm maps the names of the variables that it uses into natural numbers (see step 1 in figure 3.6). The algorithm then generates all the constraints discussed in sections 3.3.3 and 3.3.4 (step 2 in figure 3.6) and applies one of the existing software packages in order to determine a model of the logical theory that describes the problem given as input (step 3 in figure 3.6). When such a model is found, the mapping between the names of the variables to the actual structure is trivial.

Implementation and empirical results

I have written a C++ program that implements the propositional, satisfiability algorithm. The program automatically generates the variables and conjunctive normal formulas that correspond to the propositional encoding of the constraints that characterize the valid text structures of the text subsumed by the linear sequence of units given as input. Once the conjunctive normal formulas are generated, we can apply any technique for finding a model that satisfies them. I used off-the-shelf software packages to investigate empirically the computational properties of both exhaustive procedures, such as that developed by Davis and Putnam [1960], and greedy methods, such as GSAT [Selman *et al.*, 1992] and WALKSAT [Selman *et al.*, 1994].

The Davis–Putnam (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

Text	Number of variables	Number of clauses	Derivation time in seconds
A.1	45	160	<1
A.2	83	360	<1
A.3	151	818	<1
A.4	300	1856	1
A.5	306	2544	1
A.6	2298	47698	7
A.7	3865	95984	50
A.8	4558	146290	50

Table 3.2: The sizes of the propositional encodings and the amounts of time required to derive them.

Text	DP	GSAT				WALKSAT			
	Time (sec.)	Success	Time (sec.)	Max flips	Max tries	Success	Time (sec.)	Max flips	Max tries
A.1	<1	yes	<1	5000	1	yes	<1	113	1
A.2	<1	yes	<1	10000	8	yes	<1	320	1
A.3	<1	yes	17	50000	18	yes	<1	326	1
A.4	<1	no	229	50000	250	yes	<1	14711	1
A.5	<1	no	342	50000	250	yes	<1	7409	1
A.6	4	no	1821	100000	250	no	396	50000	250
A.7	137	no	2243	100000	250	no	1099	100000	250
A.8	9021	no	2262	100000	250	no	1430	100000	250

Table 3.3: Performance of the propositional logic, satisfiability-based implementations

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 3.2 shows the sizes of propositional encodings in conjunctive normal form that correspond to the texts in appendix A and the amounts of time that were required by our implementation to derive them. The data in table 3.2 suggest that as texts get larger, both the sizes of the corresponding propositional encodings and the amounts of time required to derive them can quickly exceed reasonable limits.

Table 3.3 summarizes the performance of DP, GSAT, and WALKSAT implementations in finding satisfying truth assignments for the propositional encodings of the texts given in appendix A. The second column in table 3.3 shows the amount of time required to find a satisfying truth assignment by an implementation of Davis–Putnam procedure that was downloaded from *http://www.cirl.uoregon.edu/crawford/* [Crawford and Auton, 1996]. Table 3.3 also shows whether implementations of GSAT and WALKSAT procedures [Selman *et al.*, 1992, Selman *et al.*, 1994], which were downloaded from *ftp://ftp.research.att.com/dist/ai/*, were successful in finding a satisfying truth assignment. Where a satisfying truth assignment was found, table 3.3 specifies, in the "Max tries" column, the "try" during which the procedures succeeded. Where a satisfying truth assignment was not found, table 3.3 specifies the maximum number of "tries" and "flips" that were used in attempting to find a solution. In both cases, table 3.3 shows the amount of time spent for the whole experiment.

The results in table 3.3 are interesting from two perspectives. On one hand, from a linguistic perspective, the propositional encoding shows a significant improvement over the constraint-satisfaction encoding discussed in section 3.2: 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 $O(\mathrm{N}^5)$, 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 were randomly generated. Empirical studies showed that, for such problems, the GSAT algorithm significantly outperforms the Davis–Putnam procedure. However, as table 3.3 shows, for the propositional encoding of the problem of text structure derivation it seems that it is the reverse that 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 4000 variables and more than 140000 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 thesis.

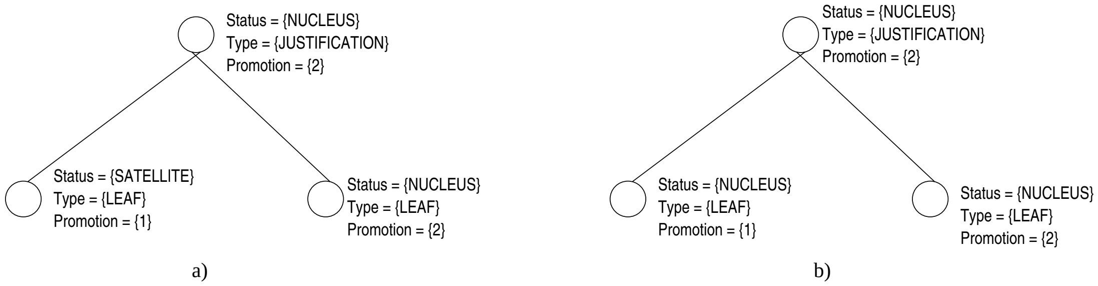

Figure 3.7: Examples of valid and invalid text structures

3.4 Deriving text structures — a proof-theoretic approach

3.4.1 Deriving text structures — a theorem proving perspective

The algorithms discussed in sections 3.2 and 3.3 derive valid text structures using model-theoretic techniques. In this section, I take a proof-theoretic stand and present a set of rules of inference (rewriting rules) that can be used to derive valid text structures starting from a given sequence of N textual units and from the set of rhetorical relations that hold among these units. The rewriting rules hence emphasize *how* valid text structures can be derived and not *what* valid text structures are.

In presenting the proof-theoretic account, I consider a universe U that consists of the set of natural numbers from 1 to N, the set of constants NUCLEUS, SATELLITE, LEAF, NULL, and the names of all rhetorical relations in a taxonomy of choice. 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 set of natural numbers from 1 to N; and $left$ and $right$ can be NULL or recursively defined objects of type $tree$. Sets of rhetorical relations such as that given in (3.4) are considered legal objects as well. We assume that the language defined over the universe U supports the traditional function symbols $+$ and $-$ and operations that are typical to sets.

The objects having the form $tree(status, type, promotion, left, right)$ can provide a functional representation of valid text structures. Assume, for example, that a rhetorical relation $rhet\_rel(\text{JUSTIFICATION}, 1, 2)$ holds among the units of a text with two elementary units. Then, the valid tree structure shown in figure 3.7.a can be represented using an object of type $tree$ as shown in (3.81). 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 3.7.b can be also represented as an object of type $tree$, as shown in (3.82), 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.

$$(3.81) \quad \begin{array}{l} tree(\text{NUCLEUS}, \text{JUSTIFICATION}, \{2\}, \\ \quad tree(\text{SATELLITE}, \text{LEAF}, \{1\}, \text{NULL}, \text{NULL}), \\ \quad tree(\text{NUCLEUS}, \text{LEAF}, \{2\}, \text{NULL}, \text{NULL})) \end{array}$$

$$(3.82) \quad \begin{array}{l} tree(\text{NUCLEUS}, \text{JUSTIFICATION}, \{2\}, \\ \quad tree(\text{NUCLEUS}, \text{LEAF}, \{1\}, \text{NULL}, \text{NULL}), \\ \quad tree(\text{NUCLEUS}, \text{LEAF}, \{2\}, \text{NULL}, \text{NULL})) \end{array}$$

Definition 3.1 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 we describe here in conjunction with universe U accepts only five predicate symbols:

- Predicate $unit(i)$ is true for each $i \leq \text{N}$ whenever the text under scrutiny can be broken into N elementary textual units. For simplicitly, we assume that these units are labelled from 1 to N. For example, for text (3.3), $unit(1)$ to $unit(4)$ are true, but $unit(5)$ 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 (3.3), the predicate $hold(RR)$ is true if RR contains the list of rhetorical relations shown in (3.4).

- Predicate $S(l, h, tree(\ldots), R_{lh})$ is true when a valid text structure that corresponds to the argument $tree(\ldots)$ 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 hold among the units in the text that have not been used in the construction of the valid structure that corresponds to the object $tree(\ldots)$. For example, given text (3.3) and the set of elementary and extended rhetorical relations that hold among its units (3.4), the predicate in (3.83) is true. In contrast, the predicate in (3.84) is false because the term $tree$ does not correspond to a valid text structure — the relation JUSTIFICATION is mononuclear.

$$(3.83) \qquad S(1,2,tree(\text{NUCLEUS}, \text{JUSTIFICATION}, \{2\},\\ tree(\text{SATELLITE}, \text{LEAF}, \{1\}, \text{NULL}, \text{NULL}),\\ tree(\text{NUCLEUS}, \text{LEAF}, \{2\}, \text{NULL}, \text{NULL})),\\ RR \setminus \{rhet\_rel(\text{JUSTIFICATION}, 1, 2)\})$$

$$(3.84) \qquad S(1,2,tree(\text{NUCLEUS}, \text{JUSTIFICATION}, \{2\},\\ tree(\text{NUCLEUS}, \text{LEAF}, \{1\}, \text{NULL}, \text{NULL}),\\ tree(\text{NUCLEUS}, \text{LEAF}, \{2\}, \text{NULL}, \text{NULL})),\\ RR \setminus \{rhet\_rel(\text{JUSTIFICATION}, 1, 2)\})$$

We 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 we use 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 we use RST, *paratactic*(CONTRAST) and *paratactic*(SEQUENCE) are both true.

We take instantiations of schemata (3.85) and (3.86) with respect to the taxonomy of relations that is used as axioms of a logical system that characterizes how text structures can be derived.

$$(3.85) \qquad hypotactic(relation\_name)$$

$$(3.86) \qquad paratactic(relation\_name)$$

Given a sequence of N textual units and a set RR of rhetorical relations that hold among these units, we take (3.87) as axiom as well.

$$(3.87) \qquad hold(RR)$$

We also take $unit(1), unit(2), \ldots, unit(\text{N})$, i.e., the applications of schema (3.88) for $1 \leq i \leq \text{N}$, as axioms in our system.

$$(3.88) \qquad unit(i)$$

We 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, we assume that for every textual unit i in the initial sequence of textual units $1, \ldots, \text{N}$ there exists a textual span S that can be associated with a valid text structure that has either status NUCLEUS or SATELLITE, type LEAF, and promotion set $\{i\}$; 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 (3.89) and N axioms having the form (3.90).

$$[unit(i) \wedge hold(RR)] \rightarrow S(i, i, tree(\text{NUCLEUS}, \text{LEAF}, \{i\}, \text{NULL}, \text{NULL}), RR) \quad (3.89)$$

$$[unit(i) \wedge hold(RR)] \rightarrow S(i, i, tree(\text{SATELLITE}, \text{LEAF}, \{i\}, \text{NULL}, \text{NULL}), RR) \quad (3.90)$$

The intuition 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 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.

Besides the axioms shown above, we consider now a set of axioms that explain how adjacent spans can be assembled into larger spans. These axioms provide a procedural account of the strong compositionality criterion. 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 $[b+1, h]$ ($rhet\_rel(name, s, n) \in rr_1 \cap rr_2$), 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 rule (3.91)) or SATELLITE (see rule (3.92)), 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 $rr_1 \cap rr_2 \setminus \{rhet\_rel(name, s, n)\}$.

(3.91)
$$\begin{aligned}
&[S(l, b, tree_1(\textsc{satellite}, type_1, p_1, left_1, right_1), rr_1) \wedge \\
&S(b+1, h, tree_2(\textsc{nucleus}, type_2, p_2, left_2, right_2), rr_2) \wedge \\
&rhet_rel(name, s, n) \in rr_1 \cap rr_2 \wedge s \in p_1 \wedge n \in p_2 \wedge hypotactic(name)] \rightarrow \\
&\qquad S(l, h, tree(\textsc{nucleus}, name, p_2, tree_1(\ldots), tree_2(\ldots)), \\
&\qquad\qquad rr_1 \cap rr_2 \setminus \{rhet_rel(name, s, n)\})
\end{aligned}$$

(3.92)
$$\begin{aligned}
&[S(l, b, tree_1(\textsc{satellite}, type_1, p_1, left_1, right_1), rr_1) \wedge \\
&S(b+1, h, tree_2(\textsc{nucleus}, type_2, p_2, left_2, right_2), rr_2) \wedge \\
&rhet_rel(name, s, n) \in rr_1 \cap rr_2 \wedge s \in p_1 \wedge n \in p_2 \wedge hypotactic(name)] \rightarrow \\
&\qquad S(l, h, tree(\textsc{satellite}, name, p_2, tree_1(\ldots), tree_2(\ldots)), \\
&\qquad\qquad rr_1 \cap rr_2 \setminus \{rhet_rel(name, s, n)\})
\end{aligned}$$

Similarly, we define rules of inference for the cases in which an extended rhetorical relation holds across spans $[l, b]$ and $[b+1, h]$ (3.93)–(3.94); for the cases in which the nucleus goes before the satellite (3.95)–(3.98); and for the cases in which the relation under scrutiny is paratactic (3.99)–(3.102).

(3.93)
$$\begin{aligned}
&[S(l, b, tree_1(\textsc{satellite}, type_1, p_1, left_1, right_1), rr_1) \wedge \\
&S(b+1, h, tree_2(\textsc{nucleus}, type_2, p_2, left_2, right_2), rr_2) \wedge \\
&rhet_rel_ext(name, l, b, b+1, h) \in rr_1 \cap rr_2 \wedge hypotactic(name)] \rightarrow \\
&\qquad S(l, h, tree(\textsc{nucleus}, name, p_2, tree_1(\ldots), tree_2(\ldots)), \\
&\qquad\qquad rr_1 \cap rr_2 \setminus \{rhet_rel(name, l, b, b+1, h)\})
\end{aligned}$$

(3.94)
$$\begin{aligned}
&[S(l, b, tree_1(\textsc{satellite}, type_1, p_1, left_1, right_1), rr_1) \wedge \\
&S(b+1, h, tree_2(\textsc{nucleus}, type_2, p_2, left_2, right_2), rr_2) \wedge \\
&rhet_rel_ext(name, l, b, b+1, h) \in rr_1 \cap rr_2 \wedge hypotactic(name)] \rightarrow \\
&\qquad S(l, h, tree(\textsc{satellite}, name, p_2, tree_1(\ldots), tree_2(\ldots)), \\
&\qquad\qquad rr_1 \cap rr_2 \setminus \{rhet_rel(name, l, b, b+1, h)\})
\end{aligned}$$

$$
\begin{aligned}
(3.95)\qquad & [S(l,b,tree_1(\textsc{nucleus},type_1,p_1,\mathit{left}_1,right_1),rr_1)\wedge\\
& S(b+1,h,tree_2(\textsc{satellite},type_2,p_2,\mathit{left}_2,right_2),rr_2)\wedge\\
& rhet\_rel(name,s,n)\in rr_1\cap rr_2\wedge s\in p_2\wedge n\in p_1\wedge hypotactic(name)]\rightarrow\\
& \quad S(l,h,tree(\textsc{nucleus},name,p_1,tree_1(\ldots),tree_2(\ldots)),\\
& \qquad rr_1\cap rr_2\setminus\{rhet\_rel(name,s,n)\})
\end{aligned}
$$

$$
\begin{aligned}
(3.96)\qquad & [S(l,b,tree_1(\textsc{nucleus},type_1,p_1,\mathit{left}_1,right_1),rr_1)\wedge\\
& S(b+1,h,tree_2(\textsc{satellite},type_2,p_2,\mathit{left}_2,right_2),rr_2)\wedge\\
& rhet\_rel(name,s,n)\in rr_1\cap rr_2\wedge s\in p_2\wedge n\in p_1\wedge hypotactic(name)]\rightarrow\\
& \quad S(l,h,tree(\textsc{satellite},name,p_1,tree_1(\ldots),tree_2(\ldots)),\\
& \qquad rr_1\cap rr_2\setminus\{rhet\_rel(name,s,n)\})
\end{aligned}
$$

$$
\begin{aligned}
(3.97)\qquad & [S(l,b,tree_1(\textsc{nucleus},type_1,p_1,\mathit{left}_1,right_1),rr_1)\wedge\\
& S(b+1,h,tree_2(\textsc{satellite},type_2,p_2,\mathit{left}_2,right_2),rr_2)\wedge\\
& rhet\_rel\_ext(name,b+1,h,l,b)\in rr_1\cap rr_2\wedge hypotactic(name)]\rightarrow\\
& \quad S(l,h,tree(\textsc{nucleus},name,p_1,tree_1(\ldots),tree_2(\ldots)),\\
& \qquad rr_1\cap rr_2\setminus\{rhet\_rel(name,b+1,h,l,b)\})
\end{aligned}
$$

$$
\begin{aligned}
(3.98)\qquad & [S(l,b,tree_1(\textsc{nucleus},type_1,p_1,\mathit{left}_1,right_1),rr_1)\wedge\\
& S(b+1,h,tree_2(\textsc{satellite},type_2,p_2,\mathit{left}_2,right_2),rr_2)\wedge\\
& rhet\_rel\_ext(name,b+1,h,l,b)\in rr_1\cap rr_2\wedge hypotactic(name)]\rightarrow\\
& \quad S(l,h,tree(\textsc{satellite},name,p_1,tree_1(\ldots),tree_2(\ldots)),\\
& \qquad rr_1\cap rr_2\setminus\{rhet\_rel(name,b+1,h,l,b)\})
\end{aligned}
$$

$$
\begin{aligned}
(3.99)\qquad & [S(l,b,tree_1(\textsc{nucleus},type_1,p_1,\mathit{left}_1,right_1),rr_1)\wedge\\
& S(b+1,h,tree_2(\textsc{nucleus},type_2,p_2,\mathit{left}_2,right_2),rr_2)\wedge\\
& rhet\_rel(name,n_1,n_2)\in rr_1\cap rr_2\wedge n_1\in p_1\wedge n_2\in p_2\wedge paratactic(name)]\rightarrow\\
& \quad S(l,h,tree(\textsc{nucleus},name,p_1\cup p_2,tree_1(\ldots),tree_2(\ldots)),\\
& \qquad rr_1\cap rr_2\setminus\{rhet\_rel(name,n_1,n_2)\})
\end{aligned}
$$

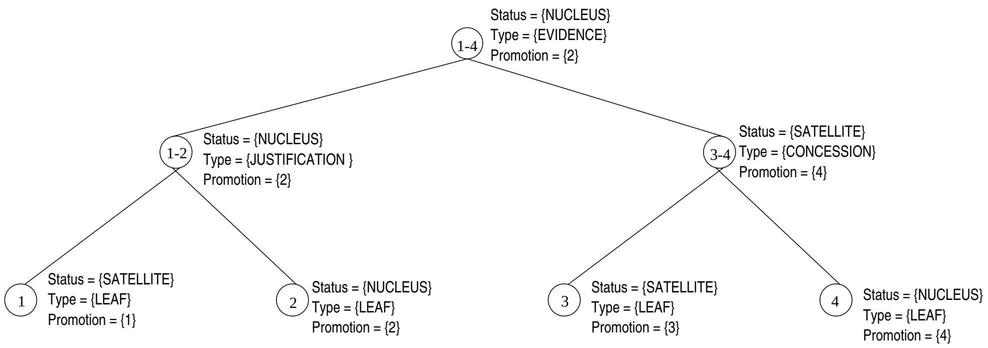

Figure 3.8: One of the valid text structures that corresponds to text (3.3).

$$
\begin{aligned}
(3.100)\quad & [S(l, b, tree_1(\textsc{nucleus}, type_1, p_1, left_1, right_1), rr_1) \wedge \\
& S(b+1, h, tree_2(\textsc{nucleus}, type_2, p_2, left_2, right_2), rr_2) \wedge \\
& rhet\_rel(name, n_1, n_2) \in rr_1 \cap rr_2 \wedge n_1 \in p_1 \wedge n_2 \in p_2 \wedge paratactic(name)] \rightarrow \\
& \qquad S(l, h, tree(\textsc{satellite}, name, p_1 \cup p_2, tree_1(\ldots), tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus \{rhet\_rel(name, n_1, n_2)\})
\end{aligned}
$$

$$
\begin{aligned}
(3.101)\quad & [S(l, b, tree_1(\textsc{nucleus}, type_1, p_1, left_1, right_1), rr_1) \wedge \\
& S(b+1, h, tree_2(\textsc{nucleus}, type_2, p_2, left_2, right_2), rr_2) \wedge \\
& rhet\_rel\_ext(name, l, b, b+1, h) \in rr_1 \cap rr_2 \wedge paratactic(name)] \rightarrow \\
& \qquad S(l, h, tree(\textsc{nucleus}, name, p_1 \cup p_2, tree_1(\ldots), tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus \{rhet\_rel(name, l, b, b+1, h)\})
\end{aligned}
$$

$$
\begin{aligned}
(3.102)\quad & [S(l, b, tree_1(\textsc{nucleus}, type_1, p_1, left_1, right_1), rr_1) \wedge \\
& S(b+1, h, tree_2(\textsc{nucleus}, type_2, p_2, left_2, right_2), rr_2) \wedge \\
& rhet\_rel\_ext(name, l, b, b+1, h) \in rr_1 \cap rr_2 \wedge paratactic(name)] \rightarrow \\
& \qquad S(l, h, tree(\textsc{satellite}, name, p_1 \cup p_2, tree_1(\ldots), tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus \{rhet\_rel(name, l, b, b+1, h)\})
\end{aligned}
$$

Axioms (3.85)–(3.102) provide a *proof-theoretic account* of the problem of text structure derivation.

1. $hold(RR)$ — Axiom (3.87)
2. $unit(1)$ — Axiom (3.88)
3. $unit(2)$ — Axiom (3.88)
4. $unit(3)$ — Axiom (3.88)
5. $unit(4)$ — Axiom (3.88)
6. $S(1, 1, tree(\text{SATELLITE}, \text{LEAF}, \{1\}, \text{NULL}, \text{NULL}), RR)$ — 1, 2, Axiom (3.90), MP
7. $S(2, 2, tree(\text{NUCLEUS}, \text{LEAF}, \{2\}, \text{NULL}, \text{NULL}), RR)$ — 1, 3, Axiom (3.89), MP
8. $S(1, 2, tree(\text{NUCLEUS}, \text{JUSTIFICATION}_1, \{2\},$
   $tree(\text{SATELLITE}, \text{LEAF}, \{1\}, \text{NULL}, \text{NULL}),$
   $tree(\text{NUCLEUS}, \text{LEAF}, \{2\}, \text{NULL}, \text{NULL})),$
   $RR_1)$ — 6, 7, Axiom (3.91), MP
9. $S(3, 3, tree(\text{SATELLITE}, \text{LEAF}, \{3\}, \text{NULL}, \text{NULL}), RR)$ — 1, 4, Axiom (3.90) , MP
10. $S(4, 4, tree(\text{NUCLEUS}, \text{LEAF}, \{4\}, \text{NULL}, \text{NULL}), RR)$ — 1, 5, Axiom (3.89) , MP
11. $S(3, 4, tree(\text{SATELLITE}, \text{CONCESSION}, \{4\},$
   $tree(\text{SATELLITE}, \text{LEAF}, \{3\}, \text{NULL}, \text{NULL}),$
   $tree(\text{NUCLEUS}, \text{LEAF}, \{4\}, \text{NULL}, \text{NULL})),$
   $RR_2)$ — 9, 10, Axiom (3.92), MP
12. $S(1, 4, tree(\text{NUCLEUS}, \text{JUSTIFICATION}_2, \{2\},$
   $tree(\text{NUCLEUS}, \text{JUSTIFICATION}, \{2\},$
   $tree(\text{SATELLITE}, \text{LEAF}, \{1\}, \text{NULL}, \text{NULL}),$
   $tree(\text{NUCLEUS}, \text{LEAF}, \{2\}, \text{NULL}, \text{NULL})),$
   $tree(\text{SATELLITE}, \text{CONCESSION}, \{4\},$
   $tree(\text{SATELLITE}, \text{LEAF}, \{3\}, \text{NULL}, \text{NULL}),$
   $tree(\text{NUCLEUS}, \text{LEAF}, \{4\}, \text{NULL}, \text{NULL}))),$
   $RR_3)$ — 8, 11, Axiom (3.95), MP

Figure 3.9: A derivation of the theorem that corresponds to the valid text structure shown in 3.8.

3.4.2 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 (3.3) and assume that the rhetorical relations RR in (3.4) hold among the units in the text. In figure 3.9, we sketch the derivation of the theorem that corresponds to the valid text structure that is shown in figure 3.8. The sets of rhetorical relations RR_1, RR_2, and RR_3 that will be used in the derivation are shown in (3.103), (3.104), and (3.105), respectively.

$$(3.103) \qquad RR_1 = \begin{cases} rhet\_rel(\textsc{justification}_2, 4, 2) \\ rhet\_rel(\textsc{evidence}, 3, 2) \\ rhet\_rel(\textsc{concession}, 3, 4) \\ rhet\_rel(\textsc{restatement}, 4, 1) \\ rhet\_rel\_ext(\textsc{justification}_3, 1, 1, 2, 4) \end{cases}$$

$$(3.104) \qquad RR_2 = \begin{cases} rhet\_rel(\textsc{justification}_1, 1, 2) \\ rhet\_rel(\textsc{justification}_2, 4, 2) \\ rhet\_rel(\textsc{evidence}, 3, 2) \\ rhet\_rel(\textsc{restatement}, 4, 1) \\ rhet\_rel\_ext(\textsc{justification}_3, 1, 1, 2, 4) \end{cases}$$

$$(3.105) \qquad RR_3 = \begin{cases} rhet\_rel(\textsc{evidence}, 3, 2) \\ rhet\_rel(\textsc{restatement}, 4, 1) \\ rhet\_rel\_ext(\textsc{justification}_3, 1, 1, 2, 4) \end{cases}$$

The derivation starts with one instantiation of axiom (3.87) and four instantiations of axiom (3.88). Using the axioms in lines 1 and 2, axiom (3.90), and the Modus Ponens rule, we derive the theorem in line 6. Using the axioms in lines 1 and 3, axiom (3.89), and Modus Ponens, we derive the theorem in line 7. Both theorems correspond to valid text structures that can be built on top of elementary units. Using the theorems in lines 6 and 7, axiom (3.91), and Modus Ponens, we derive the theorem in line 8. It corresponds to a valid text structure that can be build across span $[1, 2]$. Since this structure uses rhetorical relation $rhet\_rel(\textsc{justification}_1, 1, 2)$, the set RR_1 of rhetorical relations that can be used to expand further the text structure will no longer contain this relation. Similarly, we derive the theorem in line 11, which corresponds to a valid text structure that spans across units 3 and 4. Using the theorems derived in line 8 and 11, axiom (3.95), and Modus Ponens gives us a theorem that corresponds to a valid structure for the whole text, the structure shown in figure 3.8.

3.4.3 The proof-theoretic account of valid text structures is sound and complete

Given the formalization of text structures in chapter 2 and the set of axioms introduced in this section, it is natural to ask what the relationship between the two is. Theorem 3.1 spells out the nature of this relationship.

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

Proof. Since axioms (3.85)–(3.102) are essentially Horn clauses, for the purpose of this proof, I will treat them in the same way Prolog does. More precisely, instead on focusing on their fixed-point semantics, I will treat axioms (3.85)–(3.102) 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 (3.85)–(3.102) is consistent with the constraints described in chapter 2.

In order to prove the theorem, we first make the observation that the objects of type *tree* that are accepted by the logical language described in this section obey, by definition, most of the constraints that pertain to a valid text structure. Each of the objects of type *tree* essentially encodes a binary text structure whose nodes are characterized by a status, a type, and a promotion set. 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 among which rhetorical relations RR hold, $unit(1), \ldots, unit(n)$ 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 (3.89)–(3.102). These axioms fall into two categories. Axioms (3.89) and (3.90) 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 (3.91)–(3.102) are nothing but a one-to-one translation of the strong compositionality criterion 2.2. Therefore, the theorems that these axioms generate correspond always to valid text structures. □

Proof of completeness. The proof follows immediately from lemma 3.1. Given any text T, the algorithm shown in figure 3.10 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 the proof-theoretic
account of valid text structures.

1. apply axiom schema (3.87)
2. **for** $i := 1$ **to** N
3. apply axiom schema (3.88)
4. apply axiom schemata (3.89)–(3.90)
5. **for** $size\_of\_span := 1$ **to** $\text{N} - 1$
6. **for** $l := 1$ **to** $\text{N} - size\_of\_span$
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 $r \in RR_1 \cap RR_2$
12. apply all possible axioms (3.91)–(3.102)

Figure 3.10: An algorithm that derives all the theorems that characterize a text T with respect to the proof-theoretic account of valid text structures.

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 3.10 by means of the proof-theoretic account correspond to <u>all</u> valid structures that can be built for any span $[l, h]$ of T, where $1 \leq l \leq h \leq \text{N}$.*

Proof. The algorithm in figure 3.10 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 i of the text are described by structures that correspond either to term $tree(\text{SATELLITE}, \text{LEAF}, \{i\}, \text{NULL}, \text{NULL})$ or to term $tree(\text{NUCLEUS}, \text{LEAF}, \{i\}, \text{NULL}, \text{NULL})$. Lines 2–4 of the algorithm in figure 3.10 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 3.10 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 chapter 2, 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 (3.91)–(3.102) 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 (3.106).

$$
\begin{aligned}
(3.106)\qquad & S(l, h, tree(\text{SATELLITE}, \text{NAME}, \text{P}_2, \\
&\qquad tree(\text{SATELLITE}, \text{NAME}_1, \text{P}_1, \mathit{left}_1, right_1), \\
&\qquad tree(\text{NUCLEUS}, \text{NAME}_2, \text{P}_2, \mathit{left}_2, right_2)), \\
&\quad RR_{lh})
\end{aligned}
$$

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 $tree(\text{SATELLITE}, \text{NAME}_1, \text{P}_1, \mathit{left}_1, right_1)$ and $tree(\text{NUCLEUS}, \text{NAME}_2, \text{P}_2, \mathit{left}_2, right_2)$ respectively. According to the induction hypothesis, this means that the theorems given in (3.107) and (3.108) hold for some $rr_1, rr_2 \subseteq RR$.

$$(3.107)\qquad S(l, b, tree(\text{SATELLITE}, \text{NAME}_1, \text{P}_1, \mathit{left}_1, right_1), rr_1)$$

$$(3.108)\qquad S(b+1, h, tree(\text{NUCLEUS}, \text{NAME}_2, \text{P}_2, \mathit{left}_2, right_2), rr_2)$$

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. 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 (3.91), we only need to prove that $rhet\_rel(\text{NAME}, \text{S}, \text{N}) \in rr_1 \cap rr_2$.

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. Since $rhet\_rel(\text{NAME},\text{S},\text{N})$ 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, $rhet\_rel(\text{NAME,S,N})$ must be in the set $rr_1 \cap rr_2$.

All the antecedents that pertain to axiom (3.91) are true. Therefore, one can use axiom (3.91) and Modus Ponens to derive theorem (3.106), which contradicts our initial hypothesis that vs cannot be derived. The proof of the other cases is similar. □

3.4.4 Implementation and empirical results

There are many ways in which one can implement a set of rewriting rules of the kind described in this section. 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. Or one can write a grammar having rules such as those shown in (3.109), where each grammar rule is associated with a set of semantic constraints in the style of Montague [1973].

(3.109)
$$\begin{array}{l} S(sem) \rightarrow i \quad \{sem = \{tree(\text{NUCLEUS}, \text{LEAF}, \{i\}, \text{NULL}, \text{NULL}), RR\}\} \\ S(sem) \rightarrow i \quad \{sem = \{tree(\text{SATELLITE}, \text{LEAF}, \{i\}, \text{NULL}, \text{NULL}), RR\}\} \\ S(sem) \rightarrow S(sem_1)\ \ S(sem_2) \quad \{sem = f(sem_1, sem_2)\} \end{array}$$

The grammar-based approach assumes that the input is a sequence of textual units $1, 2, \ldots, \text{N}$. Each nonterminal S in the grammar has associated a semantics 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 (3.91)–(3.102). For example, the semantic constraint associated with rule (3.91) is that shown in (3.110) below.

(3.110)
$$\frac{\begin{array}{l} [sem_1 = \{tree_1(\text{SATELLITE}, type_1, p_1, left_1, right_1), rr_1\} \wedge \\ sem_2 = \{tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2\} \wedge \\ rhet\_rel(name, s, n) \in rr_1 \cap rr_2 \wedge s \in p_1 \wedge n \in p_2 \wedge hypotactic(name)] \end{array}}{\begin{array}{c} sem = \{tree(\text{NUCLEUS}, name, p_2, tree_1(\ldots), tree_2(\ldots)), \\ rr_1 \cap rr_2 \setminus \{rhet\_rel(name, s, n)\} \end{array}}$$

Taking the grammar-based approach, I modified the bottom-up parser described by Norvig [1992, p. 665] so that it takes as input a sequence of elementary textual units and the set of rhetorical relations that hold among these units, and builds a semantic representation that subsumes all the valid text structures that correspond to the text. The parser applies a memoization procedure[1] in order to avoid computing the same structure

[1]A memoization procedure consists in creating dinamically a database of function input/output pairs; whenever a memoized function is called, the database is checked in order to avoid computing the same

Text	Time in seconds	Number of valid structures
A.1	0.02	3
A.2	0.03	5
A.3	0.16	40
A.4	0.10	8
A.5	0.14	20
A.6	19.20	816
A.7	45.48	2584
A.8	13227.00	24055

Table 3.4: The performance of the bottom-up parser and the total number of valid trees that correspond to the texts given in appendix A.

twice, being therefore equivalent to a chart parser. Table 3.4 shows the time required by the bottom-up parser to derive *all* the valid text structures that correspond to the texts in appendix A. It is obvious that the proof-theoretic paradigm for deriving valid text structures has much better computational properties than the model-theoretic paradigms discussed in sections 3.2 and 3.3. However, the empirical data also suggests that in the cases in which the number of valid trees is very large, the performance of the algorithm degrades. Therefore, if we are to apply this algorithm on larger instances, we would need to find ways to compute only some of the valid structures.

3.5 Deriving text structures — compiling grammars in Chomsky normal form

3.5.1 From text structures to Chomsky normal-form grammars

In general, finding solutions of constraint-satisfaction problems and finding models of theories of propositional formulas are NP-complete problems [Garey and Johnson, 1979, Mackworth, 1977]. And parsing phrase structure trees in the presence of functional constraints can be exponential in the worst case [Maxwell and Kaplan, 1993, Barton *et al.*, 1985]. Therefore, deriving the valid text structures of a text using the algorithms described in sections 3.2–3.4 can be exponential in the worst case because these algorithms do not fully exploit the characteristics of the problem that we are trying to solve. In this section, we show that we can compile in polynomial time the problem of text structure derivation 2.2 into a grammar in Chomsky normal form and we prove that the size of the grammar is polynomial in the length of the input. Since one can recognize whether a string of length N

function more than once.

belongs to the language defined by a Chomsky normal-form grammar in polynomial time too, $O(\mathrm{N}^3)$, it follows that one can derive the valid text structures of a text in polynomial time.

Two crucial observations allow us to compile the problem of text structure derivation into a Chomsky normal-form grammar.

- The first observation is that a valid text structure can be recovered from an "almost-valid" text structure, i.e., a structure that associates only one unit with each promotion set. As we showed in section 3.2, one can map an "almost-valid" structure into a valid one in polynomial time. Hence, for the purpose of this section, we assume that the promotion sets of each span have cardinality one.

- The second observation is that the number of possible combinations of the values associated with the status, type, and promotion set of each node of a valid text structure is finite. Hence, given a span $[l, h]$, there exists only a finite number of symbols $S\langle l, h, status, type, promotion\_set\rangle$ that encode the variables that characterize completely each node of a valid text structure. Since the *status* of a valid span ranges over the set $\{\text{NUCLEUS}, \text{SATELLITE}\}$, the *type* over a set of $k_{[l,h]} \leq |RR|^2$ relations that are relevant to that span, and the promotion set over the elements of the set $\{\{l\}, \{l+1\}, \ldots, \{h\}\}$, it follows that there are at most $2k_{[l,h]}(h - l + 1)$ distinct symbols $S\langle l, h, status, type, promotion\_set\rangle$ that can characterize completely a span $[l, h]$ that plays an active role in a text structure.

Let us assume that we are given a sequence of textual units $U = 1, 2, \ldots, \mathrm{N}$ and a set RR that encodes all the relations that hold among these units. For example, text (3.3) is characterized by sequence $1, 2, 3, 4$ and by rhetorical relations (3.4). We present now an algorithm that starting from U and RR constructs a grammar in Chomsky normal form that can be used to derive all and only the valid text structures of U.

The compiling algorithm in figure 3.11 derives a set of rules P that fall into two categories. The rules compiled in lines 1–3 have the form $S\langle i, i, \ldots\rangle \rightarrow i$ and $S \rightarrow i$ — they are used to recognize terminal symbols $1, 2, \ldots, \mathrm{N}$. The rules compiled in lines 4–36 have the form $S\langle l, h, \ldots\rangle \rightarrow S\langle l, b, \ldots\rangle\ S\langle b+1, h, \ldots\rangle$ and $S \rightarrow S\langle l, b, \ldots\rangle\ S\langle b+1, h, \ldots\rangle$, where $l \leq b \leq h$ — they correspond to joining adjacent spans into larger spans. Hence, the compiling algorithm derives a set of production rules P that corresponds to a grammar $G = (S, T, N, P)$ in Chomsky normal form. The starting symbol of the grammar is S, the set of terminal symbols T is the set $\{1, 2, \ldots, \mathrm{N}\}$, and the set of nonterminal symbols N is given by the union of $\{S\}$ and all the symbols having the form $S\langle x, y, \ldots\rangle$ that occur in P.

[2] The symbol $|RR|$ denotes the cardinality of the initial set of relations that hold among the units of the text.

Input: A sequence $U = 1, 2, \ldots, \text{N}$ of elementary textual units and
A set RR of rhetorical relations that hold among these units.
Output: A grammar in Chomsky normal form that can be used to derive all and only the parse trees that correspond to the valid text structures of U.

1. **for** $i := 1$ **to** N
2. add rules $S \rightarrow i$, $S\langle i, i, \text{NUCLEUS}, \text{LEAF}, \{i\}\rangle \rightarrow i$, and $S\langle i, i, \text{SATELLITE}, \text{LEAF}, \{i\}\rangle \rightarrow i$
3. **endfor**
4. **for** $size\_of\_span := 1$ **to** $\text{N} - 1$
5. **for** $l := 1$ **to** $n - size\_of\_span$
6. $h = l + size\_of\_span$
7. **for** $b := l$ **to** $h - 1$
8. **for** $x := l$ **to** b
9. **for** $y := b + 1$ **to** h
10. **for each** $name_1$ for which a rule has $S\langle l, b, \text{SATELLITE}, name_1, \{x\}\rangle$ as its head
11. **for each** $name_2$ for which a rule has $S\langle b+1, h, \text{NUCLEUS}, name_2, \{y\}\rangle$ as its head
12. **for each** hypotactic relation $name$ such that $rhet\_rel(name, x, y) \in RR$ or $rhet\_rel(name, l, b, b+1, h) \in RR$
13. add rule $S \rightarrow S\langle l, b, \text{SATELLITE}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{NUCLEUS}, name_2, \{y\}\rangle$
14. add rule $S\langle l, h, \text{SATELLITE}, name, \{y\}\rangle \rightarrow S\langle l, b, \text{SATELLITE}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{NUCLEUS}, name_2, \{y\}\rangle$
15. add rule $S\langle l, h, \text{NUCLEUS}, name, \{y\}\rangle \rightarrow S\langle l, b, \text{SATELLITE}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{NUCLEUS}, name_2, \{y\}\rangle$
16. **endfor**
17. **endfor**
18. **endfor**
19. **for each** $name_1$ for which a rule has $S\langle l, b, \text{NUCLEUS}, name_1, \{x\}\rangle$ as its head
20. **for each** $name_2$ for which a rule has $S\langle b+1, h, \text{SATELLITE}, name_2, \{y\}\rangle$ as its head
21. **for each** hypotactic relation $name$ such that $rhet\_rel(name, y, x) \in RR$ or $rhet\_rel(name, b+1, h, l, b) \in RR$
22. add rule $S \rightarrow S\langle l, b, \text{NUCLEUS}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{SATELLITE}, name_2, \{y\}\rangle$
23. add rule $S\langle l, h, \text{SATELLITE}, name, \{x\}\rangle \rightarrow S\langle l, b, \text{NUCLEUS}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{SATELLITE}, name_2, \{y\}\rangle$
24. add rule $S\langle l, h, \text{NUCLEUS}, name, \{x\}\rangle \rightarrow S\langle l, b, \text{NUCLEUS}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{SATELLITE}, name_2, \{y\}\rangle$
25. **endfor**
26. **endfor**
27. **endfor**
28. **for each** $name_1$ for which a rule has $S\langle l, b, \text{NUCLEUS}, name_1, \{x\}\rangle$ as its head
29. **for each** $name_2$ for which a rule has $S\langle b+1, h, \text{NUCLEUS}, name_2, \{y\}\rangle$ as its head
30. **for each** paratactic relation $name$ such that $rhet\_rel(name, x, y) \in RR$ or $rhet\_rel(name, l, b, b+1, h) \in RR$
31. add rule $S \rightarrow S\langle l, b, \text{NUCLEUS}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{NUCLEUS}, name_2, \{y\}\rangle$
32. add rule $S\langle l, h, \text{SATELLITE}, name, \{x\}\rangle \rightarrow S\langle l, b, \text{NUCLEUS}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{NUCLEUS}, name_2, \{y\}\rangle$
33. add rule $S\langle l, h, \text{SATELLITE}, name, \{y\}\rangle \rightarrow S\langle l, b, \text{NUCLEUS}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{NUCLEUS}, name_2, \{y\}\rangle$
34. add rule $S\langle l, h, \text{NUCLEUS}, name, \{x\}\rangle \rightarrow S\langle l, b, \text{NUCLEUS}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{NUCLEUS}, name_2, \{y\}\rangle$
35. add rule $S\langle l, h, \text{NUCLEUS}, name, \{y\}\rangle \rightarrow S\langle l, b, \text{NUCLEUS}, name_1, \{x\}\rangle \; S\langle b+1, h, \text{NUCLEUS}, name_2, \{y\}\rangle$
36. **end all for loops**

Figure 3.11: A compiling algorithm that converts the problem of text structure derivation (2.2) into a Chomsky normal-form grammar.

$$
\begin{array}{lll}
S \rightarrow 1 & S\langle 1,1,\text{NUCLEUS},\text{LEAF},\{1\}\rangle \rightarrow 1 & S\langle 1,1,\text{SATELLITE},\text{LEAF},\{1\}\rangle \rightarrow 1 \\
S \rightarrow 2 & S\langle 2,2,\text{NUCLEUS},\text{LEAF},\{2\}\rangle \rightarrow 2 & S\langle 2,2,\text{SATELLITE},\text{LEAF},\{2\}\rangle \rightarrow 2 \\
S \rightarrow 3 & S\langle 3,3,\text{NUCLEUS},\text{LEAF},\{3\}\rangle \rightarrow 3 & S\langle 3,3,\text{SATELLITE},\text{LEAF},\{3\}\rangle \rightarrow 3 \\
S \rightarrow 4 & S\langle 4,4,\text{NUCLEUS},\text{LEAF},\{4\}\rangle \rightarrow 4 & S\langle 4,4,\text{SATELLITE},\text{LEAF},\{4\}\rangle \rightarrow 4
\end{array}
$$

$$
\begin{array}{ll}
S & \rightarrow S\langle 1,1,\text{SATELLITE},\text{LEAF},\{1\}\rangle \\
& \quad S\langle 2,2,\text{NUCLEUS},\text{LEAF},\{2\}\rangle \\
S\langle 1,2,\text{NUCLEUS},\text{JUSTIFICATION}_1,\{2\}\rangle & \rightarrow S\langle 1,1,\text{SATELLITE},\text{LEAF},\{1\}\rangle \\
& \quad S\langle 2,2,\text{NUCLEUS},\text{LEAF},\{2\}\rangle \\
S\langle 1,2,\text{SATELLITE},\text{JUSTIFICATION}_1,\{2\}\rangle & \rightarrow S\langle 1,1,\text{SATELLITE},\text{LEAF},\{1\}\rangle \\
& \quad S\langle 2,2,\text{NUCLEUS},\text{LEAF},\{2\}\rangle \\
S & \rightarrow S\langle 2,2,\text{NUCLEUS},\text{LEAF},\{2\}\rangle \\
& \quad S\langle 3,3,\text{SATELLITE},\text{LEAF},\{3\}\rangle \\
S\langle 2,3,\text{NUCLEUS},\text{EVIDENCE},\{2\}\rangle & \rightarrow S\langle 2,2,\text{NUCLEUS},\text{LEAF},\{2\}\rangle \\
& \quad S\langle 3,3,\text{SATELLITE},\text{LEAF},\{3\}\rangle \\
S\langle 2,3,\text{SATELLITE},\text{EVIDENCE},\{2\}\rangle & \rightarrow S\langle 2,2,\text{NUCLEUS},\text{LEAF},\{2\}\rangle \\
& \quad S\langle 3,3,\text{SATELLITE},\text{LEAF},\{3\}\rangle \\
S & \rightarrow S\langle 3,3,\text{SATELLITE},\text{LEAF},\{3\}\rangle \\
& \quad S\langle 4,4,\text{NUCLEUS},\text{LEAF},\{4\}\rangle \\
S\langle 3,4,\text{NUCLEUS},\text{CONCESSION},\{3\}\rangle & \rightarrow S\langle 3,3,\text{NUCLEUS},\text{LEAF},\{3\}\rangle \\
& \quad S\langle 4,4,\text{SATELLITE},\text{LEAF},\{4\}\rangle \\
S\langle 3,4,\text{SATELLITE},\text{CONCESSION},\{3\}\rangle & \rightarrow S\langle 3,3,\text{NUCLEUS},\text{LEAF},\{3\}\rangle \\
& \quad S\langle 4,4,\text{SATELLITE},\text{LEAF},\{4\}\rangle \\
S & \rightarrow S\langle 1,2,\text{NUCLEUS},\text{JUSTIFICATION}_1,\{2\}\rangle \\
& \quad S\langle 3,3,\text{SATELLITE},\text{LEAF},\{3\}\rangle \\
S\langle 1,3,\text{NUCLEUS},\text{EVIDENCE},\{2\}\rangle & \rightarrow S\langle 1,2,\text{NUCLEUS},\text{JUSTIFICATION}_1,\{2\}\rangle \\
& \quad S\langle 3,3,\text{SATELLITE},\text{LEAF},\{3\}\rangle \\
S\langle 1,3,\text{SATELLITE},\text{EVIDENCE},\{2\}\rangle & \rightarrow S\langle 1,2,\text{NUCLEUS},\text{JUSTIFICATION}_1,\{2\}\rangle \\
& \quad S\langle 3,3,\text{SATELLITE},\text{LEAF},\{3\}\rangle \\
S & \rightarrow S\langle 1,1,\text{SATELLITE},\text{LEAF},\{1\}\rangle \\
& \quad S\langle 2,3,\text{NUCLEUS},\text{EVIDENCE},\{2\}\rangle \\
S\langle 1,3,\text{NUCLEUS},\text{JUSTIFICATION}_1,\{2\}\rangle & \rightarrow S\langle 1,1,\text{SATELLITE},\text{LEAF},\{1\}\rangle \\
& \quad S\langle 2,3,\text{NUCLEUS},\text{EVIDENCE},\{2\}\rangle \\
S\langle 1,3,\text{SATELLITE},\text{JUSTIFICATION}_1,\{2\}\rangle & \rightarrow S\langle 1,1,\text{SATELLITE},\text{LEAF},\{1\}\rangle \\
& \quad S\langle 2,3,\text{NUCLEUS},\text{EVIDENCE},\{2\}\rangle \\
S & \rightarrow S\langle 2,3,\text{NUCLEUS},\text{EVIDENCE},\{2\}\rangle \\
& \quad S\langle 4,4,\text{SATELLITE},\text{LEAF},\{4\}\rangle \\
S\langle 2,4,\text{NUCLEUS},\text{JUSTIFICATION}_2,\{2\}\rangle & \rightarrow S\langle 2,3,\text{NUCLEUS},\text{EVIDENCE},\{2\}\rangle \\
& \quad S\langle 4,4,\text{SATELLITE},\text{LEAF},\{4\}\rangle \\
S\langle 2,4,\text{SATELLITE},\text{JUSTIFICATION}_2,\{2\}\rangle & \rightarrow S\langle 2,3,\text{NUCLEUS},\text{EVIDENCE},\{2\}\rangle \\
& \quad S\langle 4,4,\text{SATELLITE},\text{LEAF},\{4\}\rangle \\
S & \rightarrow S\langle 2,2,\text{NUCLEUS},\text{LEAF},\{2\}\rangle \\
& \quad S\langle 3,4,\text{SATELLITE},\text{CONCESSION},\{3\}\rangle \\
S\langle 2,4,\text{NUCLEUS},\text{EVIDENCE},\{2\}\rangle & \rightarrow S\langle 2,2,\text{NUCLEUS},\text{LEAF},\{2\}\rangle \\
& \quad S\langle 3,4,\text{SATELLITE},\text{CONCESSION},\{3\}\rangle \\
S\langle 2,4,\text{NUCLEUS},\text{EVIDENCE},\{2\}\rangle & \rightarrow S\langle 2,2,\text{NUCLEUS},\text{LEAF},\{2\}\rangle \\
& \quad S\langle 3,4,\text{SATELLITE},\text{CONCESSION},\{3\}\rangle
\end{array}
$$

Figure 3.12: The Chomsky normal-form grammar that is derived by the compiling algorithm for text (3.3) (see figure 3.13 for the rest of the grammar).

$$
\begin{array}{ll}
S & \rightarrow S\langle 1,3,\textsc{nucleus},\textsc{justification}_1,\{2\}\rangle \\
 & S\langle 4,4,\textsc{satellite},\textsc{leaf},\{4\}\rangle \\
S\langle 1,4,\textsc{nucleus},\textsc{justification}_2,\{2\}\rangle & \rightarrow S\langle 1,3,\textsc{nucleus},\textsc{justification}_1,\{2\}\rangle \\
 & S\langle 4,4,\textsc{satellite},\textsc{leaf},\{4\}\rangle \\
S\langle 1,4,\textsc{satellite},\textsc{justification}_2,\{2\}\rangle & \rightarrow S\langle 1,3,\textsc{nucleus},\textsc{justification}_1,\{2\}\rangle \\
 & S\langle 4,4,\textsc{satellite},\textsc{leaf},\{4\}\rangle \\
S & \rightarrow S\langle 1,3,\textsc{nucleus},\textsc{evidence},\{2\}\rangle \\
 & S\langle 4,4,\textsc{satellite},\textsc{leaf},\{4\}\rangle \\
S\langle 1,4,\textsc{nucleus},\textsc{justification}_2,\{2\}\rangle & \rightarrow S\langle 1,3,\textsc{nucleus},\textsc{evidence},\{2\}\rangle \\
 & S\langle 4,4,\textsc{satellite},\textsc{leaf},\{4\}\rangle \\
S\langle 1,4,\textsc{satellite},\textsc{justification}_2,\{2\}\rangle & \rightarrow S\langle 1,3,\textsc{nucleus},\textsc{evidence},\{2\}\rangle \\
 & S\langle 4,4,\textsc{satellite},\textsc{leaf},\{4\}\rangle \\
S & \rightarrow S\langle 1,2,\textsc{nucleus},\textsc{justification}_1,\{2\}\rangle \\
 & S\langle 3,4,\textsc{satellite},\textsc{concession},\{3\}\rangle \\
S\langle 1,4,\textsc{nucleus},\textsc{evidence},\{2\}\rangle & \rightarrow S\langle 1,2,\textsc{nucleus},\textsc{justification}_1,\{2\}\rangle \\
 & S\langle 3,4,\textsc{satellite},\textsc{concession},\{3\}\rangle \\
S\langle 1,4,\textsc{satellite},\textsc{evidence},\{2\}\rangle & \rightarrow S\langle 1,2,\textsc{nucleus},\textsc{justification}_1,\{2\}\rangle \\
 & S\langle 3,4,\textsc{satellite},\textsc{concession},\{3\}\rangle \\
S & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{evidence},\{2\}\rangle \\
S\langle 1,4,\textsc{nucleus},\textsc{justification}_3,\{2\}\rangle & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{evidence},\{2\}\rangle \\
S\langle 1,4,\textsc{satellite},\textsc{justification}_3,\{2\}\rangle & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{evidence},\{2\}\rangle \\
S & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{justification}_2,\{2\}\rangle \\
S\langle 1,4,\textsc{nucleus},\textsc{justification}_3,\{2\}\rangle & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{justification}_2,\{2\}\rangle \\
S\langle 1,4,\textsc{satellite},\textsc{justification}_3,\{2\}\rangle & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{justification}_2,\{2\}\rangle \\
S & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{evidence},\{2\}\rangle \\
S\langle 1,4,\textsc{nucleus},\textsc{justification}_1,\{2\}\rangle & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{evidence},\{2\}\rangle \\
S\langle 1,4,\textsc{satellite},\textsc{justification}_1,\{2\}\rangle & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{evidence},\{2\}\rangle \\
S & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{justification}_2,\{2\}\rangle \\
S\langle 1,4,\textsc{nucleus},\textsc{justification}_1,\{2\}\rangle & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{justification}_2,\{2\}\rangle \\
S\langle 1,4,\textsc{satellite},\textsc{justification}_1,\{2\}\rangle & \rightarrow S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\}\rangle \\
 & S\langle 2,4,\textsc{nucleus},\textsc{justification}_2,\{2\}\rangle
\end{array}
$$

Figure 3.13: The Chomsky normal-form grammar that is derived by the compiling algorithm for text (3.3) (see figure 3.12 for the rest of the grammar).

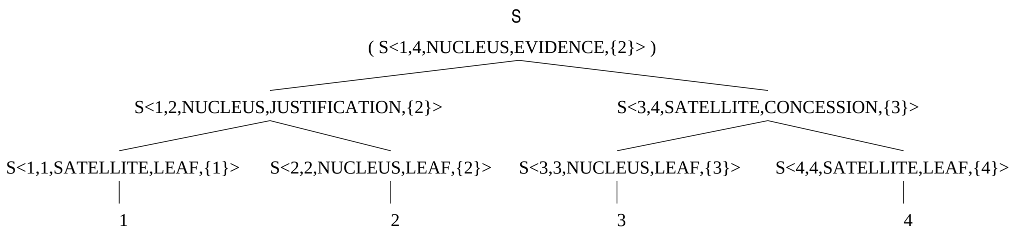

Figure 3.14: A Chomsky normal-form derivation that is isomorphic to a valid tree structure that corresponds to text (3.3).

For example, if we consider text (3.3) and its corresponding set of relations (3.4), the rules in figures 3.12 and 3.13 are the complete set of rules of a grammar in Chomsky normal form that are derived by the compiling algorithm in figure 3.11. These rules can be used to parse the input $1, 2, 3, 4$ and obtain derivations such as that shown in figure 3.14. By inspecting the derivation in 3.14, it is easy to notice that there exists a clear isomorphism between the parse tree derived using the grammar rules and the corresponding valid text structure, the structure shown in figure 3.8. In order to enable the reader visualize this isomorphism, I have represented the root of the parse tree in figure 3.14 using both the starting symbol S and the nonterminal $S\langle 1, 4, \textsc{nucleus}, \textsc{evidence}, \{2\}\rangle$, which would be obtained when a bottom-up parsing algorithm is applied.

3.5.2 Soundness and completeness results concerning the grammars generated by the compiling algorithm

In designing the compiling algorithm in figure 3.11, I have chosen to use nonterminal names that reflect all the variables that are essential for the axiomatization of valid text structures: the status, type, and promotion set of each node. Given the set of rules that the algorithm produces, we can notice that terminal symbols can be derived using only simple rules and that nonterminal symbols can be derived using only binary rules. Hence, any derivation of any input string will produce a binary parse tree. The question that still needs to be answered concerns the relationship between the parse trees that would result from the application of the grammar rules on a given text and the valid structures of that text. Theorem 3.2, which is given below, discusses the nature of the relationship.

Theorem 3.2. *Consider a sequence of textual units* $1, 2, \ldots, \textsc{n}$ *and a set* RR *that encodes all the relations that hold among these units. The compiling algorithm in figure 3.11 generates a Chomsky normal-form grammar that can be used to derive all and only the parse trees that are isomorphic with the valid structures of text* $1, 2, \ldots, \textsc{n}$.

The claim that the grammars generated by the compiling algorithm derive only parse trees that are isomorphic to valid text structures concerns the soundness of the grammar rules.

The claim that the grammars derive all parse trees that are isomorphic to valid text structures concerns the completeness of the rules.

Proof of soundness. The compiling algorithm generates all the grammar rules that correspond to building spans of size 1, 2, 3, and so on, up to N. It does so by considering for each span $[l,h]$ all the possible ways in which the span can be broken into two adjacent subspans and all the possible relations from the initial set RR that hold across the two subspans. For each relation r that holds across the adjacent subspans $[l,b]$ and $[b+1,h]$, it generates all the grammar rules that enforce the strong compositionality criterion: that is, the algorithm considers all pairs of nonterminals that characterize spans $[l,b]$ and $[b+1,h]$ and generates rules for each such pair. Consider such a rule for the case in which the relation r is hypotactic. Take, for example, rule (3.111), which is also shown in line 14 in figure 3.11).

$$S\langle l,h,\text{SATELLITE},name,\{y\}\rangle \rightarrow S\langle l,b,\text{SATELLITE},name_1,\{x\}\rangle \\ S\langle b+1,h,\text{NUCLEUS},name_2,\{y\}\rangle \tag{3.111}$$

A simple inspection of this rule, and all the other rules generated by the algorithm, shows that it enforces the compositionality criterion with respect to the statuses and promotion sets of the subspans. Since these rules are the only rules that will be used for recognizing a string $1,2,\ldots,\text{N}$, it follows that the resulting derivation will obey the strong compositionality criterion as well. However, since the rules are applied one by one, the only problem that might occur is that we might obtain a parse tree that uses the same rhetorical relation twice. We show now that this is impossible.

Each grammar rule associated with a span $[l,h]$ is built using two previously generated nonterminals that correspond to two adjacent subspans $[l,b]$ and $[b+1,h]$. Assume that $name$ is a relation that holds across the two spans, and assume that $name_1$ and $name_2$ are the names of the relations that are associated with the first and second nonterminals of the rule, as shown in (3.111). If $S\langle l,b,\text{SATELLITE},name_1,\{x\}\rangle$ is a valid nonterminal, then relation $name_1$ holds between two units found within the span $[l,b]$. If $S\langle l,h,\text{SATELLITE},name,\{y\}\rangle$ is a valid nonterminal, then relation $name$ holds between a unit of span $[l,b]$ and a unit of span $[b+1,h]$. It follows that $name$ and $name_1$ cannot be the same. Similarly, we can show that $name$ and $name_2$ cannot be the same. Since these observations hold for any span $[l,h]$, it follows that no relation is used twice in any parse of the whole input string $1,2,\ldots,\text{N}$. □

Proof of completeness. The proof of completeness is isomorphic to that of theorem 3.1. □

Text	Time in seconds
A.1	<0.01
A.2	<0.01
A.3	0.01
A.4	0.01
A.5	0.01
A.6	0.69
A.7	2.01
A.8	5.21

Table 3.5: The performance of the algorithm that compiles the fundamental problem of text processing into a grammar in Chomsky normal form.

3.5.3 An estimation of the size of the grammar

Assume that we are given a text with N elementary units and that k relations hold on average between any two elementary units. An upper bound of the number of rules that are generated by the compiling algorithm corresponds to the case in which all relations are paratactic (lines 28–35 of the algorithm). Given a span $[a, b]$ and a unit $u \in \{\{a\}, \{a+1\}, \dots, \{b\}\}$, there are at most k relations that promote unit u as a salient unit and, hence, at most k nonterminal symbols of the form $S\langle a, b, \text{NUCLEUS}, type, \{u\}\rangle$. It follows that lines 31–35 are executed at most $|RR|k^2$ times, where $|RR|$ represents the cardinality of the initial set of rhetorical relations. Hence, the algorithm in figure 3.11 generates at most

$$3\text{N} + \sum_{1 \le s < \text{N}} \sum_{1 \le l \le \text{N}-s} \sum_{l \le b < l+s} \sum_{l \le x \le b} \sum_{b+1 \le y \le l+s} 5k^2|RR| =$$
$$3\text{N} + 1/120k^2|RR|\text{N}(\text{N}-1)(\text{N}+1)(\text{N}+2)(\text{N}+3)$$

grammar rules, where s stands for *size_of_span*. If we use the same upper bounds for k and $|RR|$ as in section 3.4, we obtain that the algorithm generates at most $O(\text{N}^6)$ grammar rules in $O(\text{N}^6)$ steps. Once the grammar is generated, one can use it to derive the text structures of the text in $O(\text{N}^3)$, using the Cocke-Kasami-Younger algorithm [Younger, 1967]. Therefore, it follows that given a text T of N elementary textual units and the set RR of rhetorical relations that hold among these units, one can derive the valid text structures of text T in polynomial time $O(\text{N}^6)$.

3.5.4 Implementation and empirical results

I implemented the compiling algorithm shown in figure 3.11 in Lisp. Besides deriving the grammar rules, the implementation also stores in a chart each nonterminal symbol of the grammar, in the style of the Cocke-Kasami-Younger algorithm [Younger, 1967]. Hence, the implementation produces not only a grammar in Chomsky normal form but also the chart that the Cocke-Kasami-Younger algorithm would produce using that grammar. Hence, one can use the compiling algorithm to simultaneously generate a grammar and produce its corresponding chart. In other words, the implementation of the compiling algorithm follows closely the Cocke-Kasami-Younger approach — it stores in polynomial space a possibly exponential number of valid text structures.

Table 3.5 shows the amounts of time required by the Lisp implementation for deriving the compact chart from which any valid text structure can be extracted. Since valid structures can be extracted from this chart in polynomial time, it is obvious that the compiling algorithm significantly outperforms all the other approaches.

3.6 Related work

3.6.1 General discussion

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], 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[3] of an existing discourse tree with an incoming elementary discourse tree, the

[3]The right frontier is the set of nodes of the tree structure that are found on a path from the root to 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] 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 what follows, instead of arguing in favour of a grammar formalism or a particular set of discourse rules, I prefer to address two problems that I consider to be independent of the type of grammar or rules that all these approaches use. The first problem pertains to the assumption that discourse units can be adjoined only to nodes that belong to the right frontier of the existing discourse structure: I shall show that the notion of "right frontier" is weaker than the notion of discourse compositionality that I introduced in chapter 2. The second problem pertains to the inherently nonmonotonic nature of incremental tree derivation.

3.6.2 The notion of "right frontier" is weaker than compositionality criterion 2.1

All grammar-based approaches to discourse assume that only the right frontier of a discourse tree can accommodate a new unit. Consider, however, the naturally occurring text (3.112), which is given below.[4]

(3.112) [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]]

Assume that we are using an incremental approach and wish to derive the discourse structure of text (3.112) and assume that we have already processed the first two units of the text (see figure 3.15.a) and are about to process the third unit (see figure 3.15.b–d). We follow Cristea and Webber's notation [1997] and assume that the processing of the third unit of text (3.112) gives rise to the auxiliary tree shown in figure 3.15.c. The node labelled

right-most leaf.

[4]Text (3.112) is a fragment of text (2.1), which is discussed in section 2.2.1 and chapter 6 and for which a discourse structure was built by two independent analysts.

with an asterisk in tree 3.15.c is a "foot" node, which can be adjoined to a node that belongs to the right frontier. According to incremental approaches to discourse derivation, adjoining corresponds to identifying a discourse relation between the new material, in this case unit 3, and material on the right frontier of the discourse structure built so far. If we take this requirement literally, we can adjoin tree 3.15.c either at the node labelled 2 or at the root of the tree 3.15.b. Obviously, since the third unit in the text is related to the first unit through a JOINT relation, we cannot adjoin tree 3.15.c at the node labelled 2. But, we cannot adjoin tree 3.15.c at the root of tree 3.15.b either, because the third unit is in a JOINT relation with the first unit and not with the ELABORATION relation that holds between the first two units. And we cannot adjoin tree 3.15.c at the node labelled 1 in tree 3.15.b because, although units 3 and 1 are related through a JOINT relation, unit 1 is not a node of the right frontier of tree 3.15.b. The only way we can make the notion of right frontier work is by associating with the root of tree 3.15.b some information that will enable tree 3.15.c to be adjoined to it. But this information is unit 1 and associating unit 1 with the root of tree 3.15.b corresponds to applying compositionality criterion 2.1.

In other words, if we obey only the right frontier principle but do not promote unit 1 to the set of salient units of the root of tree 3.15.c, we can never determine the discourse structure of text (3.112): the parsing process would fail when unit 3 would need to be inserted in the partial tree 3.15.c.

One can easily imagine texts in which salient units that are embedded more deeply in the structure to the left of the right frontier are eventually elaborated or contrasted. For example, in order to adjoin unit 4 (see figure 3.15.e–g) to the tree that corresponds to the processing of units 1 to 3, we need the root of the tree in figure 3.15.e be characterized by both units 1 and 3 because the BACKGROUND relation holds between both these units and unit 4. Unless the salient information, in this case the information corresponding to units 1 and 3, is propagated upwards during the tree construction, the application of the right-frontier principle is impossible. Because of this, I consider the notion of "right frontier" to be weaker than compositionality criterion 2.1. In fact, the treatment of anaphora proposed by van den Berg [1996] and the treatment of adjunction proposed by Gardent [1997] are nothing but a semantic expression of compositionality criterion 2.1. Van den Berg associates with the nodes of a discourse tree feature structures that store the discourse referents. Whenever a new node is added to a partial discourse tree, the mother node inherits the discourse referents of the children. These referents can be subsequently used for anaphora resolution. And Gardent distinguishes between feature structures that are relevant to the mother nodes and feature structures that are relevant to the daughter nodes [Gardent, 1997] and provides mechanisms through which adjunction operations affect not only the feature structures of daughter nodes but also the feature structures of mother nodes.

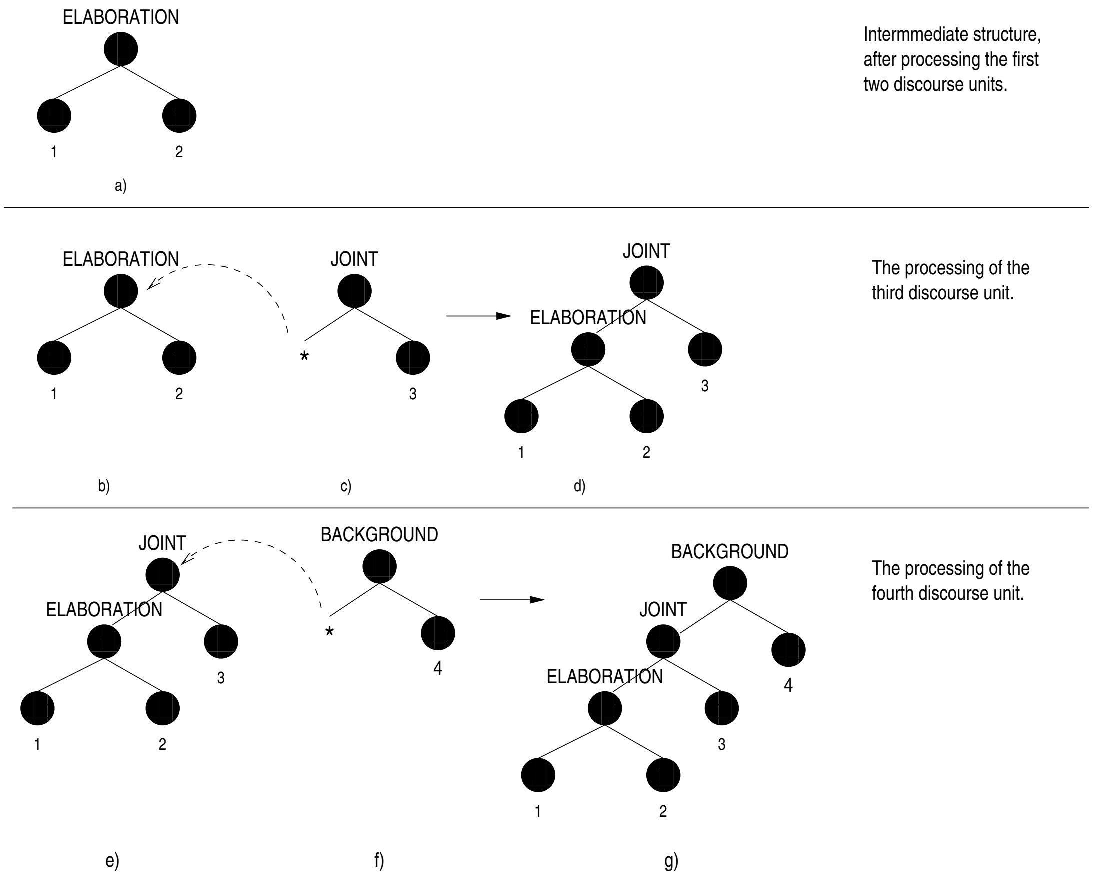

Figure 3.15: The incremental derivation of the discourse structure of text (3.112).

3.6.3 The incremental derivation of discourse structures is nonmonotonic

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 all current grammar-based approaches may be inefficient from a computational perspective. Consider example (3.113), which is reproduced from [Cristea and Webber, 1997].

(3.113) [Because John is such a generous man[1]][— whenever he is asked for money,[2]] [he will give whatever he has, for example[3]][— he deserves the "Citizen of the Year" award.[4]]

As Cristea and Webber note, the fact that unit 2 provides together with unit 3 an example for 1, rather than satisfying the expectation raised by "Because", becomes apparent only when unit 3 is processed — more specifically, when the discourse marker "for example" is considered. Obviously, in order to accommodate the finding that units 2 and 3 are an

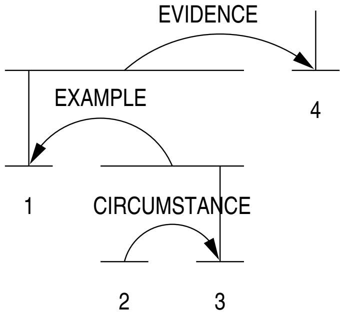

Figure 3.16: The valid text structure of text (3.113).

example for the idea presented in the first unit, we have to undo the adjoining of node 2. Therefore, the incremental processing of discourse cannot be monotonic. In order to deal with the nonmonotonicity of incremental discourse derivation, we either have to 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.

The paradigm that I propose in this thesis is to determine first all possible rhetorical relations that hold among the units of a text, and determine only afterwards the discourse structure of that text. For example, for text (3.113), we will first determine that the relations given in (3.114) hold among the elementary units of the text, and then apply any of the algorithms discussed in this chapter to derive the valid discourse structure shown in figure 3.16.

$$(3.114) \quad \begin{cases} rhet\_rel(\textsc{evidence}, 1, 4) \\ rhet\_rel(\textsc{circumstance}, 2, 3) \\ rhet\_rel(\textsc{example}, 3, 1) \end{cases}$$

The non-incremental paradigm that I presented in this chapter is efficient but admittedly, it 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 re-interpretations that are observed in naturally occurring conversations [Hirst *et al.*, 1993, McRoy, 1993].[5]

Studying the ways in which the algorithms presented in this chapter can be modified in order to derive valid text structures incrementally is, however, outside the scope of this thesis.

[5] I thank Graeme Hirst for bringing up this hypothesis.

3.7 Summary

In this chapter, I have investigated both theoretically and empirically the computational properties of four paradigms that can be used to derive valid text structures. I showed how the problem of text structure derivation 2.2 can be mapped into a constraint-satisfaction problem and I showed that the direct formulation of the strong compositionality criterion has a negative effect on the performance of the CSP-based approach.

I then showed how the problem of text structure derivation can be mapped into a propositional logic encoding in conjunctive normal form that is polynomial in size with respect to the number of units in the input text. Surprisingly, the empirical experiments that attempted to determine satisfying truth assignments for propositional encodings of eight discourse problems showed that the Davis-Putnam exhaustive procedure [1960] outperformed the stochastic procedures GSAT and WALKSAT [Selman *et al.*, 1992, Selman *et al.*, 1994].

I presented a set of axioms and inference rules that can be used to derive valid text structures through proof-theoretic techniques. The implementation of this approach significantly outperformed the approaches that attempted to derive valid structure on the basis of model-theoretic techniques.

I also gave an algorithm that compiles in polynomial time the problem of text structure derivation into a grammar in Chomsky normal form whose size is polynomial in the number of elementary units of the input text. Using this approach proved to be the most efficient method for deriving the valid structures of texts.

Chapter 4

A corpus analysis of cue phrases

4.1 Towards determining the discourse structure of unrestricted texts

Given the formalization of text structures and the algorithms that derive them that we have seen in chapters 2 and 3, in order to automatically build the valid text structures of an arbitrary text, we need only to determine the elementary units of that text and the rhetorical relations that hold among the units. An accurate determination of the elementary units of a text and of the relations that hold among them is beyond the current state of the art in natural language processing. However, empirical and computational research suggests that we can find and exploit approximate solutions to both of these problems by capitalizing on the occurrence of certain lexicogrammatical constructs.

In this chapter, I first discuss the lexicogrammatical constructs that can be used to determine the elementary units of a text and to hypothesize rhetorical relations among them. These constructs include grammatical morphemes, tense and aspect, certain lexical and syntactic structures, certain patterns of pronominalization and anaphoric usages, cohesive devices, and cue phrases. In section 4.3, I argue that a shallow analysis of text that relies primarily on knowledge about the way cue phrases like *because, however,* and *in addition* are used can indicate the underlying structure of text. The rest of the chapter discusses an exploratory corpus study of cue phrases. The study is meant to provide empirical grounding for a set of algorithms that bridge the gap between the problem of deriving valid text structures for unrestricted texts and the theoretical problem of text structure derivation that was discussed in chapter 3.

4.2 From linguistic constructs to discourse structures

Grammatical morphemes

The role of grammatical morphemes in structuring discourse relies on extending the role that they have in signalling the syntactic structures that are licensed by a generative approach to grammar [Chomsky, 1965]. 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 direct 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].

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 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 verb *meet* in example (4.1)); but the use of simple past verbs in relative clauses usually corresponds to background material (see the use of verb *engineer* in example (4.2)) [Decker, 1985].

(4.1) 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]

(4.2) "The ice has been broken," proclaimed President Belisario Betancur of Colombia, who *engineered* the meeting. [Decker, 1985, p. 317]

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 formalization of discourse relations [1993], the event of pushing associated with the second sentence in example (4.3) is normally assumed to have produced the event of falling associated with the first sentence, if the pushing event occurred before the falling event.

(4.3) Max fell. John pushed him. [Lascarides and Asher, 1993]

Hence, a causal relation is normally assumed to hold between the sentences in (4.3).

Syntactic constructs

Traditionally, *cleft* constructions have been considered to enable a reader select which element of a sentence is in focus. According to Quirk et.al. [1985, p. 89], a cleft sentence is divided into two parts: an initial focal element, and a "background" structure which follows the initial element and which resembles a relative clause. For example, "Julie" is the focal element and "who buys her vegetables in the market" is the background structure in the cleft sentence shown in (4.4), below.

(4.4) It is Julie who buys her vegetables in the market.

Prince [1978] and Delin and Oberlander [1992] have observed that cleft constructions could also 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 (4.5) 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].

(4.5) 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.

Pronominalization and anaphoric usages

Sidner [1981], Grosz and Sidner [1986], Sumita [1992], and Grosz, Joshi, and Weinstein [1995] have speculated that certain patterns of pronominalization and anaphoric usages correlate with the structure of discourse. Vonk's experimental 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 signalling of thematic shifts.

For example, in the sequence of sentences given in (4.6), 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.

(4.6)
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.
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 [1997a, 1997b] 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 [1997], 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.

Cohesive devices

The automatic detection of overspecified anaphoric expressions is still a computational challenge. However, Hearst [1994, 1997] has shown that even simple forms of lexical cohesion that are computationally tractable, such as word co-occurrences, can be used to detect topic shifts in expository texts. Much more sophisticated studies of the correlation between lexical cohesion and discourse structure are given by Morris and Hirst [Morris, 1988, Morris and Hirst, 1991], Hoey [1991], and Langleben [1983]. For example, Morris and Hirst showed that there exists a correlation between lexical chains, i.e., sequences of words related via lexical cohesion that span topical units of texts, and the structure of discourse. The lexical chains can be derived using knowledge from thesauri, such as *Roget's Thesaurus*, as used by Morris [1988] and Morris and Hirst [1991], or from lexical knowledge bases, such as

Wordnet, as used by St-Onge [1995] and Hirst and St-Onge [1997].

Cue phrases or connectives

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. Besides their structural role, connectives have been shown to have highly elaborate pragmatic functions, such as signalling shifts in the subjective perspective [Segal *et al.*, 1991, Segal and Duchan, 1997], presupposing various states of beliefs [Wing and Scholnick, 1981], and licensing inferences through mechanisms that are similar to those of scalar implicatures [Grice, 1975, Fillenbaum, 1977, Anscombre and Ducrot, 1983, Hirschberg, 1991, Oberlander and Knott, 1996]. For example, in the text shown in (4.7), which was produced by a five-year-old boy, the connectives are used to explain the thinking process of a little lion, the main character of the story.

(4.7) Once upon a time there was a little lion and he lived alone *because* his mother and father was dead. And one day he went hunting. And he saw two lions. And they were his mother and father. *So* he took his blanket to their den. *Because* it was bigger. [Segal and Duchan, 1997, p. 98]

More precisely, the *So* and the second *because* are used to build a complex *subjective* argument that explains why the lion moved in with his parents (because their space was bigger than his) and what the move entailed (taking his blanket to their den).

Psycholinguistic research also suggests that some connectives not only enable readers to process text faster, but also to recall better the related information [Deaton and Gernsbacher, 1997, Gernsbacher, 1997]. In three experiments, Deaton and Gernsbacher have shown that two-clause sentences that describe moderately causal events were read faster when the clauses were conjoined by *because* (Susan called the doctor for help *because* the baby cried in his playpen) than when they were conjoined by *and*, *then*, or *after*. In addition, when the clauses were conjoined by *because*, subjects recalled the second clauses more frequently when prompted with the first clause.

The facet of connectives that I explore in this thesis 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 pertaining to an inferential process that is structural in nature.

4.3 Arguments for a shallow approach to discourse processing

As I argued in the previous section, the problem of determining with high accuracy the elementary textual units and the rhetorical relations that hold among elementary and non-elementary units is not yet solvable. However, we saw that a significant set of lexicogrammatical constructs can be used to provide approximate solutions for it. In the rest of this thesis, I investigate how far we can get in building valid structures for unrestricted texts by focusing our attention only on discourse connectives and lexicogrammatical constructs that can be detected by means of a *shallow analysis* of natural language texts. The intuition behind this choice relies on the following facts.

- Psycholinguistic and other empirical research has shown that discourse markers are consistently used by human subjects both as cohesive ties between adjacent clauses and as "macroconnectors" between larger textual units. For example, in Halliday and Hasan's view [1976], connectives are linguistic devices that provide textual cohesion over successive sentences. Thus, their view is more local than global. The local function of connectives has been also proved to be essential for understanding the intentions of the participants in dialogues [Schiffrin, 1987]; increasing a reader's recall of information pertaining to related clauses and sentences; and contributing to the information represented in the text [Segal *et al.*, 1991].

  Empirical studies of narratives, stories, and naturally occurring conversations have shown that connectives have a global role as well. 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].

- 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 of readers to interpret the unmarked textual units as continuations of the topics of the units that precede them [Segal *et al.*, 1991].

- Discourse markers are used in a manner that is consistent with the semantics and pragmatics of the discourse segments that they relate. In other words, I assume that the texts that we process are well-formed from a discourse perspective, much as researchers in sentence parsing assume that they are well-formed from a syntactic perspective. As a consequence, I assume that one can bootstrap the full syntactic, semantic, and pragmatic analysis of the clauses that make up a text and still end up with a reliable discourse structure for that text. In fact, in many cases, a deep semantic analysis will not help, because rhetorical relations cannot be inferred only on the basis of the semantics and pragmatics of the considered textual units; rather, a connective is required in order to trigger that inference [Segal and Duchan, 1997]. Consider, for example, the following two utterances, which are taken from [Paley, 1981, p. 4]:

  (4.8) There was a little boy with no mother and no father. *But* he had seven brothers and seven sisters.

  As Segal and Duchan aptly point out [1997, p. 117], had there been an *And* in place of the *But*, one would interpret the second sentence as an assertion of this family situation. It is the occurrence of *But* that instructs the reader to contrast the situation of being an orphan with that of having many siblings.

Given the above discussion, the immediate objection that one can raise is that discourse markers are three-ways ambiguous. In some cases, their use is only sentential, i.e., they make a semantic contribution to the interpretation of a clause. And even in the cases where markers have a discourse usage, they are ambiguous with respect to the rhetorical relations that they mark and the sizes of the textual spans that they connect. I address now each of these objections in turn.

Sentential and discourse usages of cue phrases

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

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

that these results can be improved if one uses machine learning techniques [Litman, 1994, Litman, 1996] or genetic algorithms [Siegel and McKeown, 1994]. I have taken Hirschberg and Litman's research one step further and designed a comprehensive corpus analysis of cue phrases that enabled me to design algorithms that improved their results and coverage. The method, procedure, and results of the corpus analysis are discussed in this chapter. The algorithm that determines elementary unit boundaries and identifies discourse usages of cue phrases will be discussed in chapter 5.

Discourse markers are ambiguous with respect to the rhetorical relations that they mark and the sizes of the units that they connect

When I began this research, no empirical data supported the extent to which this ambiguity characterizes natural language texts. To better understand this problem, the corpus analysis that is to be described in this chapter was designed so as to also provide information about the types of rhetorical relations, rhetorical statuses (nucleus or satellite), and sizes of textual spans that each marker can indicate. I expected from the beginning that it would be impossible to predict exactly the types of relations and the sizes of the spans that a given cue marks. However, given that the structure that we are trying to build is highly constrained, such a prediction proved to be unnecessary; the overall constraints on the structure of discourse that I enumerated in chapter 2 cancel out most of the configurations of elementary constraints that do not yield valid discourse trees.

Consider, for example, the following text:

(4.9) [*Although* discourse markers are ambiguous,[1]] [one can use them to build discourse trees for unrestricted texts:[2]] [this will lead to many new applications in natural language processing.[3]]

For the sake of argument, assume that we are able to break text (4.9) into textual units as labelled above and that we are interested now in finding rhetorical relations between these units. Assume now that we can infer that *Although* marks a CONCESSIVE relation between satellite 1 and nucleus 2, and the colon, an ELABORATION between satellite 3 and nucleus either 1 or 2. A representation of text (4.9) is then the set of relations given in (4.10), where $\oplus$ denotes exclusive disjunction:

$$(4.10)\quad \begin{cases} rhet\_rel(\textsc{concession}, 1, 2) \\ rhet\_rel(\textsc{elaboration}, 3, 1) \oplus rhet\_rel(\textsc{elaboration}, 3, 2) \end{cases}$$

Despite the ambiguity of the relations, the overall rhetorical structure constraints will associate only one discourse tree with text (4.9), namely the tree given in figure 4.1. Any discourse tree configuration that uses relation $rhet\_rel(\textsc{elaboration}, 3, 1)$ will be ruled

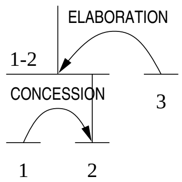

Figure 4.1: The discourse tree of text (1).

out because unit 1 is not an important unit for span $[1, 2]$ and, as discussed in chapter 2, a rhetorical relation that holds between two spans of a valid text structure must also hold between their most important units: the important unit of span $[1, 2]$ is unit 2, i.e., the nucleus of the relation $rhet\_rel(\text{CONCESSION}, 1, 2)$.

4.4 A corpus analysis of cue phrases

4.4.1 Motivation

The discussion in section 4.3 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 chapter, I describe it in detail and provide some general results.

4.4.2 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 10 text fragments for each occurrence of a cue phrase that was found at the beginning of a paragraph or sentence, and 20 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 17 text fragments per cue phrase, having few texts for the cue phrases that do not occur very often in the corpus and up to 60 for cue phrases such as *and*, which I considered to be highly ambiguous. Overall, I randomly selected more than 7600 texts. Appendix B 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.

The reader is warned that the given number of occurrences of each cue phrase in the Brown Corpus is only a rough estimate. For example, according to the table shown in appendix B, there are 950 occurrences of the cue phrase *even* in the Brown corpus: 150 of them at the beginning of a sentence and 800 in the middle or at the end. However, this number includes also occurrences of *even after*, *even before*, *even if*, *even so*, *even then*, *even though*, and *even when*, which are assigned separate entries in the table. Hence, because the program that randomly extracted text samples was not written so as to avoid extracting text fragments that contained the cue phrase *even though*, for example, when looking for the phrase *even*, the list in appendix B exhibits a certain degree of redundancy. To avoid analyzing the same text fragment more than once, the fragments that were automatically assigned to a simple cue phrase, such as *even*, but were actually characterized by a complex cue phrase, such as *even after*, that had been assigned a separate entry in the initial list, were ignored during the analysis.

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 with the LaTeX macro for emphasizing text, {\em }. The cue phrase occurrence was located approximately 200 words from the beginning of the text fragment. Text (4.11) is an example fragment with the cue phrase *accordingly*.

(4.11) 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.

And text (4.12) is an example fragment with the cue phrase *Although*.

(4.12) 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.

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 4.4.3 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".

Usually, a discourse marker signals one rhetorical relation. However, in some cases, the occurrence of a simple or multiple marker, such as *and although*, which is obtained by concatenating a set of simple markers, can signal more than one rhetorical relation. The set of rhetorical relations that are signalled by such markers may relate different textual units, have different rhetorical statuses, etc. 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 markers that "link forward" and the preceding text. For example, in text (4.12), 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 text (4.12), 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

Algorithm	Field names
The clause-boundary and discourse-marker identification algorithm (section 5.3.3)	"Marker", "Usage", "Position", "Right boundary", "Break action"
The discourse-marker-based algorithm for hypothesizing rhetorical relations (section 5.4.2)	"Marker", "Usage", "Where to link", "Rhetorical relation", "Statuses", "Types of textual units", "Clause distance", "Sentence distance", "Distance to salient unit"
The bottom-up, text planning algorithms (section 7.4.3)	"Usage", "Where to link", "Rhetorical relation", "Statuses", "Types of textual units", "Clause distance", "Sentence distance", "Distance to salient unit"

Table 4.1: The fields from the corpus that were used in developing the algorithms discussed in the rest of the thesis.

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.

Discussion

The information in the fields associated with each text fragment and cue phrase constitutes the empirical foundation of five algorithms: an algorithm that identifies elementary unit boundaries and discourse usages of cue phrases; an algorithm that hypothesizes rhetorical relations that hold among textual units; and three algorithms that construct text plans in a bottom-up fashion. Table 4.1 enumerates explicitly the fields that were used in developing each of these algorithms.

4.4.3 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.

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 text (4.11).

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 non-elementary units are each enclosed within square brackets as well. For example, the field "Example" that corresponds to text (4.11) will contain the information shown in (4.13), because the cue phrase under scrutiny, *accordingly*, marks a VOLITIONAL-CAUSE relation between the units "[However, the Federal Court held that] [it was now a State matter.]" and "[It *accordingly* refused to assume jurisdiction]".

(4.13) 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.

The field "Example" that corresponds to text (4.12) is shown in (4.14) below. (The sentence containing the cue phrase *Although* is an ELABORATION on the question "[How well do faculty members govern themselves?]"; hence, all the textual units in between are enclosed within square brackets.)

(4.14) 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.

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 "They sometimes pay more attention to their rights than to their own internal problems of government." is a simple clause, I decided to break it into two elementary textual units because the cue phrase "than" can provide grounds for inferring that a COMPARISON is made between the attention that faculties pay to their rights and the attention that they pay to their own internal problems of government.

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 cue phrase signalled a rhetorical relation between two spans of text of significant sizes, I assigned those spans an elementary unit status, although in some cases they were not fully fleshed clauses. In the rest of the thesis I use the term *clause-like unit* in order to refer to such elementary units.

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 text (4.11), the "Marker" environment will contain the string "⊔accordingly⊔" because no punctuation marks or cue phrases surround the cue phrase under scrutiny.[2] 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 fragment (4.14) is "#⊔Although⊔".

Usage

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

- SENTENTIAL (S), when the cue phrase has no function in structuring the discourse. For example, in text 4.15, *above all* is used purely sententially: *above* is a preposition and *all* is a quantifier.

  (4.15) 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.

- DISCOURSE (D), when the cue phrase signals a discourse relation between two textual units. For example, in text 4.16, *Although* signals a concession relation between two clauses of the same sentence; the clauses are enclosed within square brackets.

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

- 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

[2] The symbol ⊔ denotes a blank character.

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 stated in discourse; rather, it is the role of the cue phrase to help the reader infer them.[3] For example, in text (4.17), *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 (4.17) and the speaker's knowledge and that *again* provides the means through which the hearer can infer that knowledge.

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

In text (4.18), *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.

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

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 text (4.13) 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 texts (4.14) and (4.16) 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 text (4.13) is rhetorically

[3] 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.

related to a textual unit that goes before it (B). In contrast, the clause that contains the discourse marker *Although* in text (4.16) is rhetorically related to the clause that comes immediately after it (A).

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 text (4.13) is MC;C because it relates the multiclause-like unit "[However the Federal Court held that] [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 (4.16) 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 examples (4.13) and (4.16) 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 example (4.13) has value 0. However, the field for example (4.16) 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

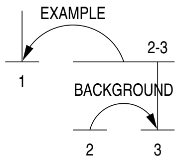

Figure 4.2: The discourse tree of text (4.19).

cases, this is not so. Consider, for example, the text given in (4.19) below, with respect to the cue phrase *for example.*

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

A rhetorical analysis of text (4.19) is shown in figure 4.2. 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 (4.19) and marker *for example*, its value is 0.

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 example (4.13) is M. However, the content of the field "Position$_1$" for example (4.16) 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 text (4.13) is S;N because the multiclause-like units "[However the Federal Court held that] [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 (4.16) 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. Appendix C contains the list of rhetorical relations that were used in the corpus analysis. 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 examples (4.13) and (4.16) are VOLITIONAL-CAUSE and CONCESSION, respectively.

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 5.3.3, in conjunction with the clause-like unit boundary and marker-identification algorithm.

4.4.4 Method and results

Once the database had been created, I analyzed each record in it and updated its fields according to the requirements described in section 4.4.3. Tables 4.2 and 4.3 show the

Field	Content
Example	(4.13)
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

Table 4.2: A corpus analysis of the segmentation and integration function of the cue phrase *accordingly* from text (4.13).

---

information that I associated with the fields when I analyzed the text fragments shown in (4.13) and (4.14) 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.[4]

The taxonomy of relations that I used to label the 1197 discourse usages in the corpus contained 54 relations. The table shown in appendix C lists their names and the number of instances in which each rhetorical relation was used. As one can note, the number of relations is much larger than 24, which is the size of the taxonomy 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 is exploratory, I considered it appropriate to introduce relations that would better capture the meaning of these relationships. The rhetorical relation names listed in appendix C were chosen so as to reflect the intended semantics of the relations.

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

- its recognition in text;

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

Field	Content
Example	(4.14)
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$_2$	6
Sentence distance$_2$	4
Distance to salient unit$_2$	-1
Position$_2$	B
Statuses$_2$	N;S
Rhetorical relation$_2$	ELABORATION
Break action	COMMA

Table 4.3: A corpus analysis of the segmentation and integration function of the cue phrase *Although* from text (4.14).

- 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 chapter 5, where I establish the connection between the corpus analysis and the algorithms that derive text structures for unrestricted texts.

In the context of natural language generation (chapter 7), I show how the corpus can be used to compute the strengths of the preferences of rhetorical relations to realize their satellites and nuclei in a certain order and to cluster their satellites and nuclei into larger textual spans.

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 thesis. For the moment, I prefer to make no claims with respect to the size and nature of an appropriate taxonomy of rhetorical relations.

4.4.5 Discussion

The main advantage of the empirical work described here consists in the empirical grounding that it provides for a set of algorithms that derive text structures of unrestricted texts in the context of discourse analysis and build valid text plans in the context of natural language generation. 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. In chapters 5 and 7, I discuss in detail the relationship between the corpus analysis and these algorithms.

The most important consequence of the fact that I was the only analyst of 2100 of the 7600 of the text fragments in the corpus concerns the evaluation procedures that I chose to use. In order to avoid evaluating the algorithms that I developed against my own subjective standard, I used the corpus analysis only for algorithm development. The testing of the algorithms was done against data that did not occur in the corpus and that was analyzed independently by a relatively large number of judges.

As I have already mentioned, I am aware of no previous empirical study that has investigated the relationship between cue phrases, rhetorical relations, and discourse units to the extent that was aimed at here. Because of this, I assumed from the beginning that my corpus analysis would have, primarily, an exploratory nature.

Ideally, the corpus analysis would be performed by more than one analyst. Unfortunately, time and cost constraints are factors that cannot be neglected when such a corpus study is designed. The magnitude of a corpus study that can provide data that is both reliable and statistically significant is beyond the scope of a PhD thesis. However, the size of the corpus is not the only problem that an analyst has to face. During my corpus analysis, I noticed a set of other problems that I consider worthy of being brought to the reader's attention. These problems stem from the lack of objective definitions for the notions of elementary textual unit, nuclearity, and rhetorical relation. Below, I discuss each of these problems in turn.

Problems with identifying the elementary units of text

My initial intent was to take clauses as the elementary units of discourse. Consider, however, the text shown in (4.20), below.

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

If I had taken my initial intent literally, I would have not broken sentence (4.20) into two units, because "light leakage from one ultraviolet source to another" does not contain a verb, and therefore, is not a clause. However, the marker *Because of* clearly signals a causal relation between the textual spans “light leakage from one ultraviolet source to another,” and “the lights are switched by a commutator-like assembly rotated by a synchronous motor”. Uncovering this relation can be only beneficial from a text understanding perspective. However, how far should one go in this attempt of using phrases rather than clauses as the elementary units of discourse?

As I have already discussed in section 4.4.3, in the texts that I analyzed, I did not use an objective definition of elementary textual unit. Rather, I relied on a more intuitive one: whenever I found that a cue phrase signalled a rhetorical relation between two spans of text of significant sizes, I assigned those spans an elementary unit status, although, in some cases, they were not fully fleshed clauses.

Problems with identifying the rhetorical status of the textual units involved in a discourse relation

As we have seen, nuclearity plays a major role in the formalization of text structures that I proposed in chapter 2. One of the main assumptions that this formalization relies upon is that the rhetorical statuses of the units involved in a rhetorical relation can be determined unambiguously. However, in few cases in the corpus, although the rhetorical relation that held between two units was easy to label, it was ambiguous as to which unit was the nucleus and which was the satellite. Consider the following example:

(4.21) [It is not enough for man to be an ontological esse.] [He needs existential completion,] [he needs, that is, to move in the direction of completion.] [And the direction of that movement is determined by his perception of the truth about himself.] [He must, *consequently*, exist as a self-perceived substantive, developing agent,] [or he does not exist as man.] [Thus, it is no mystical intuition,] [but an analyzable conception to say that man and his tradition can “fall out of existence”.] [This happens at the moment man loses the perception of moral substance in himself,] [of a nature that, in Maritain’s words, is perceived as a “locus of intelligible necessities”.] [An existentialist is a man who perceives himself only as “esse”,] [as existence without substance.]

The cue phrase *consequently* clearly marks a causal relation between units “And the direction of that movement is determined by his perception of the truth about himself” and “He must, *consequently*, exist as a self-perceived substantive, developing agent”. It is, however, not obvious which unit should be assigned the status of nucleus and which that of satellite. In fact, in general, causal relations are difficult to assign a nuclear status: in some cases, the context provides enough evidence with respect to whether the writer intended to assign a more important role to the cause or to the result. In some cases, however, it seems that the nuclearity assignment can go either way.

More precisely, it is not that the taxonomy of relations does not distinguish between causal relations in which the cause is the nucleus and causal relations in which the result is the nucleus, but rather that we lack an objective definition that would allow us to determine which of these relations to use.

Problems with identifying the rhetorical relations that holds between two textual units

During the corpus analysis, it was sometimes difficult to determine one rhetorical relation that would most adequately characterize the relation between two units. Consider, as an example the text shown in (4.22) below.

(4.22) [Certain badly disillusioned market critics are often apt to feel that there is something somehow unfair, dirty, or even thoroughly criminal about this interplay of competitive forces.] [*But after all*, can anyone imagine a market wherein the reverse of these things were true?] [Try to imagine a market in which only a minority of traders would lose,] [and the majority would make consistent profits.] [How much and how many profits could a majority take out of the losses of a few?]

What is the rhetorical relation that best describes the relationship between the first two sentences? To a certain degree is a CONTRAST between the features of a real market and the features of an imaginary one. But at the same time, an INTERPRETATION relation can be considered to hold between the two sentences as well. Which one should we choose? And if we choose both, how do we objectively assign a strength or preference to one of the relations? In my analysis, I chose to label a relation between two textual spans with *all* the names of the rhetorical relations whose definitions seemed to apply.

4.5 Related work

As I discussed at the beginning of this chapter, in order to automatically determine the valid text structures of an arbitrary text, we need only to determine the elementary units of that text and the rhetorical relations that hold among them. The corpus analysis that I have presented in the previous section, which aims at providing solutions for both of these problems, owes much to inspiration from recent developments in empirical discourse analysis. Particularly relevant is the work that pertains to segmenting discourse, distinguishing between discourse and sentential usages of cue phrases, and determining the correlation between cue phrases and discourse structure.

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; and in terms of intentional- and informational-based accounts that reflect the functional role of segments in text [Moser and Moore, 1997]. 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, 1997b, Passonneau and Litman, 1997a] 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, 1997] 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, 1997a, Moser and Moore, 1997, 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, 1997a]. 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%. In a more recent proposal, Yaari [1997] suggested that by using hierarchical agglomerative clustering algorithms one can identify topical segments in expository texts. Yaari's algorithm looks promising, but has not yet been evaluated extensively.

The corpus study discussed in this chapter was designed so as to enable the development of an algorithmic approach to identifying the elementary units of discourse. Because the notions of intention and topic yield discourse segments that are too coarse for our purpose, we could not use the algorithms described in this section.

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 co-occurrence data may provide useful information for cue phrase disambiguation. They also showed that part-of-speech 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 analysis presented in this chapter has benefited extensively from the lessons learnt from Hirschberg and Litman's study. As will become apparent in section 5.3.3, the orthographic environment and the neighboring cues play an important role in determining whether a given cue phrase has a discourse function in a text. The corpus analysis discussed in this chapter is also meant to fill a coverage gap in Hirschberg, Litman, Siegel, and McKeown's work: the corpus that they relied upon had only 953 occurrences of 34 cue phrases, which were uttered by one speaker during a speech of 75 minutes that contained approximately 12,500 words.

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. [1995] 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.* A broad empirical investigation of cue phrases was also 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, 1997]. They collected a set of 17 student-tutor interactions encompassing 144 question-answer exchanges that had 854 clauses. For each interaction in the corpus, the analysts determined the elementary and non-elementary discourse constituents and the discourse relations that hold between them. The analysts also labelled 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 labelled 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 of 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 corpus, it had a very 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.

4.6 Summary

In this chapter, 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.

The most important part of the chapter is dedicated to the presentation of an exploratory corpus study of the discourse function of cue phrases. Besides the materials and methods that I used in the corpus analysis of 450 cue phrases, I also provided some general results and discussed the need to use objective definitions of elementary textual unit, nuclear status, and rhetorical relation. At the end of the chapter, I compared the empirical work described here with previous empirical work in discourse segmentation, cue phrase disambiguation, and the discourse function of cue phrases.

Chapter 5

The rhetorical parsing of unrestricted natural language texts

5.1 Preamble

5.1.1 Pros and cons for an underspecified hierarchical representation of text

In devising a *rhetorical parsing algorithm*, i.e., an algorithm that finds the valid discourse structures of an unrestricted text, we have two choices: we can assume a text to be a "flat" sequence of elementary textual units (which, for simplicity, can be assimilated with the sequence of clauses that corresponds to that text); or we can assume a text to have a predefined, underspecified hierarchical structure whose elements are clauses, sentences, paragraphs, information blocks, sections, chapters, etc. More precisely, we can assume that the paragraphs and sections of a text are meaningful from a discourse processing perspective as much as clauses and sentences are, i.e., the paragraph and section breaks correlate with the structure of discourse. Each approach has advantages and disadvantages.

From a linguistic perspective, the advantage of taking a text to be a flat sequence of textual units is that it puts no constraints on the places where the boundaries between large textual spans can occur. If we are able to determine the rhetorical relations between textual units accurately, then the text structures that we will eventually build will be accurate as well. The disadvantage of such an approach is primarily computational. A real text may have hundreds or even thousands of elementary units. If we build a tree over such a large number of units, it is very likely that the time required by the tree-derivation process will be significant. Because my intent is to devise an algorithm that can be used in practice, on real texts, and because the rhetorical indicators that I rely upon are not very accurate, I assume that texts have a predefined, underspecified, hierarchical structure.

Consider, for example, a text that has three paragraphs with a total of 11 sentences.

The text is represented schematically in (5.1): each of the first two paragraphs has four sentences, while the third paragraph has three sentences.

(5.1) [. . . . . ^{1}] [. . . . . ^{2}] [. .
. . . . ^{3}] [. . . . . ^{4}]
[. . . . . ^{5}] [. . . . . ^{6}] [. . . . . .
. . . . ^{7}] [. . . . . ^{8}]
[. ^{9}] [. . . . . ^{10}] [. . . . . .
. . . . ^{11}]

If we assume that text (5.1) is a flat sequence of elementary units, in this case a sequence of sentences, the rhetorical parsing of text (5.1) consists in building a discourse tree over a sequence of 11 textual units. However, if we assume that paragraphs are legitimate high-level units that correlate with the structure of discourse, the rhetorical parsing of text (5.1) can be divided into three stages:

1. Find the discourse trees of each of the three paragraphs.

2. Find the discourse trees of a sequence that has only three units, corresponding to the three paragraphs of text (5.1).

3. Replace the leaves of the discourse structure that was built in step 2 with the trees that were built for each paragraph, thus obtaining a discourse tree for the whole text.

Hence, from a computational perspective, instead of deriving the discourse structure of a sequence of 11 units, we derive the discourse structure of two sequences of four units and two sequences of three units, which is a much faster process.

Although such an approach is computationally attractive, it may pose some problems in those cases in which the paragraph breaks do not match closely the thematic and intentional breaks. For example, text (5.1) may be very well characterized by a topic that ranges across sentences 1 to 5 and a topic that ranges across sentences 6 to 11. If the two topics are in contrast, an adequate discourse tree will have two major subspans: one across units 1 to 5, and another one across units 6 to 11. Obviously, an algorithm that assumes that the structure of paragraphs correlates with the structure of discourse will inappropriately build a discourse tree that has a span between units 1 and 4, a span between units 5 and 8, and a span between units 9 and 11.

Deciding whether paragraph breaks correlate well enough with the structure of discourse is not straightforward; in fact, psycholinguistic and empirical research provide contradictory evidence. For example, the psychological 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 my own empirical investigation of the relationship between text structures and text summaries (see chapter 6) suggests that paragraph breaks can help readers determine what textual units are most important in a text.

In contrast, the 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. One of the explanations of this finding is that the criteria that are used by readers in segmenting text do not fit exactly those that have been used by authors when writing them. An extreme position is taken by Longacre [1979], who mentions that paragraph breaks are often introduced only for esthetic reasons. And an experiment described by Stark [1988] seems to confirm this; reinstating paragraph breaks by students led to poor results: only nine of the paragraph breaks used by the author of a text with 17 paragraph breaks were identified as such by more that 50% of the subjects.

One way to circumvent this problem is by considering, still, that texts have a hierarchical, underspecified structure and that the larger textual units are given not by paragraphs but by "information blocks" [Heurley, 1997]. An information block is a set of sentences and paragraphs that are semantically related and that are built around a unique topic; the boundaries of an information block are independent of any orthographic marking in the surface structure of the text. Research in computational linguistics and information retrieval has shown that information blocks can be determined through a semantically-based process, which assumes that such blocks "talk about" the same thing. Word co-occurrences [Hearst, 1994, Hearst, 1997, Salton and Allan, 1995, Salton *et al.*, 1995, Richmond *et al.*, 1997, Yaari, 1997] and simple or complex chains of semantic relations, such as synonymy, hyponymy, meronymy, etc. [Morris, 1988, Morris and Hirst, 1991, Hoey, 1991, Hirst and St-Onge, 1997, Green, 1997], provide the means for determining the boundaries of these blocks.

Although appealing, the use of information blocks as legitimate, high-level textual units is hampered by the fact that word co-occurrences and even elaborate forms of semantic relatedness do not provide strong-enough means for correctly determining textual boundaries that correlate well enough with the structure of discourse [Hearst, 1994, Hearst, 1997, Morris, 1988, Morris and Hirst, 1991]. In addition, the relationship between the semantically based, cohesive devices and the rhetorical relations that they license is still insufficiently known to be applicable in determining the rhetorical relations that hold between information blocks. Even if we can determine that two information blocks are semantically related, it is still difficult to infer the nature of the rhetorical relation that would appropriately characterize this relationship [Green, 1997].

```
Input: A text T.
Output: The valid text structures of T.

1.  I. Determine the set D of all cue phrases (potential discourse markers) in T.
2.  II. Use information derived from the corpus analysis in order to determine
3.      recursively all the sections, paragraphs, sentences, and clause-like units of the
4.      text and the set D_d ∈ D of cue phrases that have a discourse function.
5.  III. For each of the three highest levels of granularity (sentences, paragraphs,
6.       and sections)
7.          III.1 Use information derived from the corpus analysis about the
8.                discourse markers D_d in order to hypothesize rhetorical relations
9.                among the elementary units that correspond to that level.
10.         III.2 Use cohesion in order to hypothesize rhetorical relations among
11.               the units for which no hypotheses were made in step III.1.
12.         III.3 Apply one of the algorithms discussed in section 5.5 in order to
13.               determine all the valid text trees that correspond to that level.
14.         III.4 Assign a weight to each of the text trees and determine the tree
15.               with maximal weight.
16. IV. Merge the best trees that correspond to each level into a discourse tree that
17.     spans the whole text and that has clause-like units as its elementary units.
```

Figure 5.1: Outline of the rhetorical parsing algorithm

5.1.2 The rhetorical parsing algorithm — a bird's-eye view

In this chapter, I present a rhetorical parsing algorithm that derives the valid discourse structures of unrestricted texts. The algorithm is outlined in figure 5.1. It assumes that texts have a predetermined, underspecified hierarchical structure with the following main levels: clause-like units, sentences, paragraphs, and sections. 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 5.1), the rhetorical parser uses information derived from the corpus analysis in chapter 4 for determining 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 5.1). 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 disjunctive hypotheses.

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

III.3 Once the set of textual units and the set of rhetorical relations that hold among the units have been determined, the algorithm derives discourse trees at each of the three levels that are assumed to be in correlation with the discourse structure: sentence, paragraph, and section levels (see lines 12-13).

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 5.6 (see lines 14–15).

In the final step, the algorithm assembles the trees built at each level of granularity, thus obtaining a discourse tree that spans over the whole text (lines 16–17 in figure 5.1).

In the rest of the chapter, 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 chapter 4, 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 text (1.1), which was taken from *Scientific American*, November 1996 and which is reproduced for convenience in (5.2) below.

(5.2) 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.

5.2 Determining the potential discourse markers of a text

5.2.1 From the corpus analysis to the potential discourse markers of a text

The corpus analysis discussed in chapter 4 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. For example, if the cue phrase *Besides* occurs at the beginning of a sentence and is not followed by a comma, as in text (5.3), it usually signals a rhetorical relation that holds between the clause-like unit that contains it and the clause that comes after. However, if the same cue phrase occurs at the beginning of a sentence and is immediately followed by a comma, as in text (5.4), it usually signals a rhetorical relation that holds between the sentence to which *Besides* belongs and a textual units that precedes it.

(5.3) *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.

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

I have taken each 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. As a result of this evaluation, I partitioned cue phrase occurrences into three classes:

1. In the first class are the cue phrases that played a discourse role in most of the text fragments in the corpus. For example, whenever the cue phrase *Although* was used, it marked a CONCESSION relation between two clauses of the same sentence. In addition, in most cases, the right boundary of the clause to which *Although* belonged was given by the occurrence of the first comma in that sentence.

2. In the second class are the cue phrases that played a discourse role in most of the text fragments in which they were adjacent to other cue phrases. For example, the cue phrase *and* had a discourse role whenever it occurred before another cue phrase, although it had both a sentential and discourse role when it occurred in isolation. In addition, when it occurred before another cue phrase, the left boundary of the clause-like unit to which *and* belonged was located just before its occurrence.

<table>
<tr><th>Marker</th><th>Regular expression</th></tr>
<tr><td>Although</td><td>[⊔\t\n]Although(⊔ | \t | \n)</td></tr>
<tr><td>because</td><td>[,][⊔\t\n]+because(⊔ | \t | \n)</td></tr>
<tr><td>but</td><td>[⊔\t\n]+but(⊔ | \t | \n)</td></tr>
<tr><td>for example</td><td>[,][⊔\t\n]+for[⊔\t \n]+example(⊔ |, | \t | \n)</td></tr>
<tr><td>where</td><td>,[⊔\t\n]+where(⊔ | \t | \n)</td></tr>
<tr><td>With</td><td>[⊔\t\n]With(⊔ | \t | \n)</td></tr>
<tr><td>Yet</td><td>[⊔\t\n]Yet(⊔ | \t | \n)</td></tr>
<tr><td>COMMA</td><td>,(⊔ | \t | \n)</td></tr>
<tr><td>OPEN_PAREN</td><td>[,][⊔\t\n]+(</td></tr>
<tr><td>CLOSE_PAREN</td><td>)(⊔ | \t | \n)</td></tr>
<tr><td>DASH</td><td>[,][⊔\t\n]+—(⊔ | \t | \n)</td></tr>
<tr><td>END_SENTENCE</td><td>(“.”)|(“?”)|(“!”)|(“."”)|(“?"”)|(“!"”))</td></tr>
<tr><td>BEGIN_PARAGRAPH</td><td>⊔⋆((\n\t[⊔\t]⋆)|(\n[⊔\t\n]{2,}))</td></tr>
</table>

Table 5.1: A list of regular expressions that correspond to occurrences of some of the potential discourse markers and punctuation marks.

---

3. In the third class are the cue phrases that played a sentential role in a majority of the text fragments and the cue phrases for which I was not able to infer straightforward rules that would allow a shallow algorithm to discriminate between their discourse and sentential usages. For example, *after* was a cue phrase for which I found it impossible to predict whether it had a discourse or sentential function by analyzing only the orthographic environment and the markers found in its neighborhood.

I used the cue phrases and the orthographic environments that characterized the cue phrases of the first two classes 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, as was the case with *Besides*, I created one regular expression for each function. I ignored the cue phrases in the third class because they were not appropriate for the surface-based approach that I investigated. Table 5.1 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. In fact, such a position is consistent with recent developments in the linguistics of punctuation [Nunberg, 1990, Briscoe, 1996, Pascual and Virbel, 1996, Say and Akman, 1996, Shiuan and Ann, 1996], which emphasize the importance of punctuation marks in a variety of natural language processing tasks that range from parsing to information packaging.

The regular expressions shown in table 5.1 obey the conventions used by the Unix tool *lex*. Table 5.2 describes the semantics of the symbols used in 5.1. For example, the regular

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

Table 5.2: The semantics of the symbols used in table 5.1.

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 dot, 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 which are followed by one of the following:

- 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.

5.2.2 An algorithm for determining the potential discourse markers of a text

Once the regular expressions that match potential discourse markers were derived, it was trivial to implement the first step of the rhetorical parser (line 1 in figure 5.1). A program that uses the Unix tool *lex* traverses the text given as input and determines the locations at which potential discourse markers occur. For example, when the regular expressions are matched against text (5.2), the algorithm recognizes all punctuation marks and the cue phrases shown in italics in text (5.5) below.

(5.5) *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.

5.3 Determining the elementary units of a text

5.3.1 From the corpus analysis to the elementary textual units of a text

As I discussed in chapter 4, 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 eight actions that constitutes the foundation of an algorithm to determine automatically 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. If no comma is found before the end of the sentence, a textual boundary is created at the end of the sentence.

- Action NORMAL_THEN_COMMA instructs the analyzer to insert a textual boundary immediately before the occurrence of the marker and to another textual boundary immediately after the occurrence of the first comma in the input stream. As in the case of the action COMMA, if the first comma is followed by an *and* or an *or*, the textual boundary is set after the occurrence of the next comma. If no comma is found before the end of the sentence, a textual boundary is created at the end of the sentence.

- Action END instructs the analyzer to insert a textual boundary immediately after the cue phrase.

- Action MATCH_PAREN instructs the analyzer to insert textual boundaries both before the occurrence of the open parenthesis that is normally characterized by such an action, and after the closed parenthesis that follows it.

- Action COMMA_PAREN instructs the analyzer to insert textual boundaries both before the cue phrase and after the occurrence of the next comma in the input stream.

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

The preceding three actions, MATCH_PAREN, COMMA_PAREN, and MATCH_DASH, are usually used for determining the boundaries of parenthetical units. These units, such as those shown in italics in (5.6), (5.7), (5.8), and (5.9) below, are related only to the larger units that they belong to or to the units that immediately precede them.

(5.6) 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).*

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

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

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

Because the deletion of parenthetical units does not affect the readibility of a text, in the algorithms that we present here we 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, we 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 (5.10) signals a rhetorical relation of CONCESSION between the clause to which it belongs and the previous clause. However, in text (5.11), 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.

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

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

In addition to the algorithmic information that is explicitly encoded in the field "Break action", the shallow analyzer 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:

- 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 5.3 lists the actions and the positions in the elementary units of the cue phrases and orthographic markers shown in table 5.1.

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

Table 5.3: The list of actions that correspond to the potential discourse markers and punctuation marks shown in table 5.1.

5.3.2 The section, paragraph, and sentence identification algorithm

As I discussed in section 5.1.2, the rhetorical parser assumes that texts have a predetermined, underspecified hierarchical structure with an optional title and four levels: sections, paragraphs, sentences, and clause-like units. Each section is assumed to be characterized by a title and by a collection of paragraphs — in fact, this is the format of most articles found in magazines and newspapers.

The algorithm that determines the section, paragraph and sentence boundaries is a very simple one. It uses the set of regular expressions that identify paragraph and sentence boundaries (see table 5.1) 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. For the purpose of the research described here, this algorithm was enough: it 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 5.3.3. However, I expect that future implementations of the rhetorical parser will take advantage of recent research in sentence boundary identification [Palmer and Hearst, 1997]. This research shows that on the basis of the orthographic environment and the part-of-speech tags of the words found in the neighborhood of a period, one can correctly determine sentence boundaries in 98 to 99 percent of the cases.

5.3.3 The clause-like unit and discourse-marker identification algorithm

On the basis of the information derived from the corpus (see table 5.3), I have designed an algorithm that identifies textual unit boundaries in a sentence and cue phrases that have a

```
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.

status := NIL; clauses := NIL; parentheticals := NIL;
currClauseStart := 1; currParentStart := −1;
for i from 1 to n
  if MATCH_PAREN ∈ status
    if markerTextEqual(i,“)”)
      parentheticals := parentheticals ∪ textFromTo(currParentStart,offset(i));
      status := status \ {MATCH_PAREN}; currParentStart := −1;
    continue;
  if MATCH_DASH ∈ status
    if markerTextEqual(i,“—”)
      parentheticals := parentheticals ∪ textFromTo(currParentStart,offset(i));
      status := status \ {MATCH_DASH}; currParentStart := −1;
    continue;
  if COMMA_PAREN ∈ status
    if markerTextEqual(i,“,”) ∧
        NextAdjacentMarkerIsNotAnd() ∧ NextAdjacentMarkerIsNotOr()
      parentheticals := parentheticals ∪ textFromTo(currParentStart,offset(i));
      status := status \ {COMMA_PAREN}; currParentStart := −1;
    continue;
  if COMMA ∈ status ∧ markerTextEqual(i,“,”) ∧
      NextAdjacentMarkerIsNotAnd() ∧ NextAdjacentMarkerIsNotOr()
    clauses := clauses ∪ textFromTo(currClauseStart,offset(i),parentheticals);
    currClauseStart := i; status := status \ {COMMA};
    parentheticals := NIL; currParentStart := −1;
    continue;
  if SET_AND ∈ status
    if markerAdjacent(i − 1,i) ∧ currClauseStart < i − 1
      clauses := clauses ∪ textFromTo(currClauseStart,offset(i − 1),parentheticals);
      currClauseStart := i − 1;
      setDiscourse(i − 1,yes); setDiscourse(i,yes);
      parentheticals := NIL;
    status := status \ {SET_AND};
  if SET_OR ∈ status
    if markerAdjacent(i − 1,i) ∧ currClauseStart < i − 1
      clauses := clauses ∪ textFromTo(currClauseStart,offset(i − 1),parentheticals);
      currClauseStart := i − 1;
      setDiscourse(i − 1,yes); setDiscourse(i,yes);
      parentheticals := NIL;
    status := status \ {SET_OR};
```

Figure 5.2: The clause-like unit and discourse-marker identification algorithm — see continuation in figure 5.3, on the next page.

```
for i from 1 to n
    ⋮
    switch(getActionType(i)){
        case DUAL:
            if markerAdjacent(i − 1,i)
              status := status ∪ {COMMA};
              setDiscourse(i − 1,yes); setDiscourse(i,yes);
            else
              clauses := clauses ∪ textFromTo(currClauseStart,offset(i),
                                                    parentheticals);
              currClauseStart := offset(i); parentheticals := NIL;
              setDiscourse(i,yes);
        case NORMAL:
            clauses := clauses ∪ textFromTo(currClauseStart,offset(i),
                                                  parentheticals);
            currClauseStart := offset(i); parentheticals := NIL;
            setDiscourse(i,yes);
        case COMMA:
            if markerAdjacent(i − 1,i)
              setDiscourse(i − 1,yes);
            setDiscourse(i,yes);
            status := status ∪ {COMMA};
        case NORMAL_THEN_COMMA:
            clauses := clauses ∪ textFromTo(currClauseStart,offset(i),
                                                  parentheticals);
            currClauseStart := offset(i); parentheticals := NIL;
            setDiscourse(i,yes);
            status := status ∪ {COMMA};
        case NOTHING:
            if signalsRhetoricalRelations(i)
              setDiscourse(i,yes);
        case MATCH_PAREN, COMMA_PAREN, MATCH_DASH:
            status := status ∪ {getActionType(i)};
            currParentStart = offset(i);
        case SET_AND, SET_OR:
            if status is neither MATCH_PAREN nor MATCH_DASH
              status := status ∪ {getActionType(i)};
    }
% end for
finishUpParentheticalsAndClauses();
```

Figure 5.3: The clause-like unit and discourse-marker identification algorithm — continuation from the previous page (figure 5.2).

discourse function. Figures 5.2 and 5.3 show its main steps. 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 the algorithm described in section 5.2.2. Each element in markers[n] is characterized by a feature structure with the following entries:

- the action associated with the cue phrase (see table 5.3);
- the position in the elementary unit of the cue phrase (see table 5.3);
- 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 3 and 71) and identifies the elementary textual units in the sentence on the basis of the types of the markers that it processes. The algorithm makes use of the following variables and functions:

- Variable "status" records the set of markers that have been processed earlier that may still influence the identification of clause and parenthetical unit boundaries. At the beginning, its value is set to NIL.
- Variable "parenthetical" records the set of parenthetical units that pertain to a given clause. At the beginning, its value is set to NIL.
- Variable "clauses" records all the elementary units that pertain to a given sentence and are not parenthetical. At the beginning, its value is NIL.
- Variable "currParentStart" records the offset in the sentence where the parenthetical unit under consideration begins. At the beginning, its value is set to -1, which means that no parenthetical unit is yet under consideration.
- Variable "currClauseStart" records the offset in the sentence where the elementary unit under consideration begins. At the beginning, its value is 1 — the first elementary unit of the sentence starts always at offset 1.
- Function markerTextEqual(i, s) returns *true* if the i-th cue phrase in the array markers[n] is equal with the string s. Otherwise, the function returns *false*.
- Function offset(i) returns the position relative to the beginning of the sentence where the i-th cue phrase of the array markers[n] occurs. The offset depends on the feature "Position" that characterizes the cue phrase. If its value is B, the function returns the position where the cue phrase starts. If its value is E, the function returns the position where the cue phrase ends.
- Function textFromTo(i, j) returns the textual unit between offsets i and j in sentence S.

- Function textFromTo(i, j,parentheticals) returns the textual unit between offsets i and j in sentence S. The textual unit is characterized by the parenthetical units stored in the variable "parentheticals".

- Function setDiscourse(i,yes) sets the feature *has_discourse_function* of the i-th cue phrase to "yes".

- Function getActionType(i) returns the action that characterizes the i-th cue phrase in the sentence S.

- Function signalsRhetoricalRelations(i) returns true if the i-th cue phrase can play a discourse role in the sentence (see section 5.4.2 for details).

- Function finishUpParentheticalsAndClauses() accounts for the text that might remain unassigned to a clause-like unit after processing the potential discourse markers of a sentence.

The clause-like unit identification algorithm has two main parts: lines 40–71 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 4–39 concern actions that are executed when the "status" variable is not NIL. We discuss now in turn each of these actions.

Lines 4–19 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 5.3.1, to assign parenthetical information no elementary textual unit status. Because of this, the text shown in italics in text (5.12), 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". The extra conditions in line 16 of the algorithm avoid seting parenthetical unit 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 (5.12) shows, taking the first comma as boundary of the parenthetical unit would be inappropriate.

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

Obviously, one can easily find counterexamples to this rule (and to other rules that are employed by the algorithm). For example, the clause-like unit and discourse-marker identification algorithm will produce erroneous results when it processes the sentence shown in (5.13) below.

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

Nevertheless, the evaluation results discussed in section 5.3.4 show that the algorithm produces correct results in the majority of the cases.

If the "status" variable contains the action COMMA, the occurrence of the first comma that is not adjacent to an *and* or *or* marker determines the identification of a new elementary unit (see lines 20–25 in figure 5.2). The boundaries of the new unit are given by the offset recorded in the variable "currClauseStart" and by the offset of the i-th marker. The third argument of the function "textFromTo" in line 22 shows that the parentheticals that have been created up to that point are considered subordinated to the elementary unit that is created. The creation of a clause-like unit also implies the reseting of the variables "currClauseStart", "status", "parentheticals", and "currParentStart".

Usually, the discourse role of the cue phrases *and* and *or* is ignored because the surface-form algorithm that we propose is unable to distinguish accurately enough between their discourse and sentential usages. However, lines 26–32 and 33–39 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, they had a discourse function. For example, in sentence (5.14), *and* acts as an indicator of a JOINT relation between the first two clauses of the text.

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

If a discourse marker is found that immediately follows the occurrence of an *and* (or an *or*) and if the left boundary of the elementary unit under consideration is found to the left of the *and* (or the *or*), a new elementary unit is identified whose right boundary is just before the *and* (or the *or*). In such a case the *and* (or the *or*) is considered to have a discourse function as well, so the flag *has_discourse_function* is set to "yes" (lines 30 and 37 in figure 5.2).

Lines 40–71 of the algorithm concern the cases in which the "status" variable is NIL. If the type of the marker is DUAL (see lines 41–49), 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 (5.15), but identify a textual unit boundary when processing the same marker in example (5.16). The final textual unit boundaries that are assigned by the algorithm are shown using square brackets.

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

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

Lines 50–54 of the algorithm concern the most frequent marker type. The type NORMAL determines the identification of a new clause-like unit whose boundaries are given by the variable "currClauseStart" and by the offset of the marker under scrutiny. Lines 55–59 concern 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. 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 a marker of type NOTHING is processed, the only action that is taken consists in assigning the marker a discourse usage. Lines 69–71 of the algorithm concern the treatment of markers that introduce expectations with respect to the occurrence of parenthetical units: the effect of processing such a marker consists of updating the "status" variable. The same updating 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 77 of figure 5.3 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 text (5.5), it determines that the text has ten elementary units and that seven cue phrases have a discourse function. Text (5.17) shows the elementary units within square brackets. The instances of parenthetical information are shown within curly brakets. The cue phrases that are assigned by the algorithm as having a discourse function are shown in italics.

(5.17) [*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]]

5.3.4 Evaluation of the clause-like unit and discourse-marker identification algorithm

The algorithm shown in figures 5.2 and 5.3 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 chapter 4.

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 at least 90% of the examples that were extracted from the corpus. The enforcement of this criterion reduces on one hand the recall of the discourse markers that can be detected, but on the other hand, significantly increases the precision. I chose this deliberately because, during the corpus analysis, I noticed that most 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 vast 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 ran-

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%

Table 5.4: Evaluation of the marker identification procedure.

domly 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 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 signalled 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. 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 5.4), a result that outperforms Hirschberg and Litman's [1993]. In fact, Hirschberg and Litman's algorithm and all its extensions that use machine learning techniques [Litman, 1994, Litman, 1996] or genetic algorithms [Siegel and McKeown, 1994] rely on manually encoded features. In contrast, the algorithm described here is fully automated: it takes as input unrestricted text, it uses the regular expressions described in section 5.2 in order to

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%

Table 5.5: Evaluation of the clause-like unit boundary identification procedure.

determine the potential discourse markers in the text, and then it determines those that have a discourse function. 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 percent of these occurrences.

The algorithm correctly identified 81.3% of the clause-like unit boundaries, with a precision of 90.3% (see table 5.5). 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 (5.18); as a consequence, the algorithm incorrectly inserts a textual unit boundary before it.

(5.18) [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.]

It is the purpose of future research to improve the algorithm described here and to investigate the benefits of using more sophisticated methods.

5.4 Hypothesizing rhetorical relations between textual units of various granularities

5.4.1 From discourse markers to rhetorical relations

In sections 5.2 and 5.3, 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 5.2.1) 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 were adjacent in all the instances in the corpus.

- The feature "Distance to salient unit" is given by the maximum of the values of the database field **Distance to salient unit**.

Marker	Stat- uses	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 $\vee$ 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 $\vee$ P	EXAMPLE	2	1
where	NULL	NULL	NULL	NULL		
With	N_S	B	S $\vee$ P	ELABORATION	5	-1
	S_N	A	C	BACKGROUND JUSTIFICATION	0	1
Yet	S_N	B	S $\vee$ 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		

Table 5.6: The list of features sets that are used to hypothesize rhetorical relations for the discourse markers and punctuation marks shown in table 5.1.

Table 5.6 lists the feature sets associated with the cue phrases that were initially listed in table 5.1. Table 5.6 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 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).

5.4.2 A discourse-marker-based algorithm for hypothesizing rhetorical relations

At the end of step II of the rhetorical parsing algorithm (see figure 5.1), 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 5.4.

At each level of granularity (sentence, paragraph, and section), the discourse-marker-based hypothesizing algorithm 5.4 iterates over all textual units of that level and over all discourse markers that are relevant to them (see lines 2–4 in figure 5.4). For each discourse marker, the algorithm constructs a 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 5.6, an appropriate disjunctive hypothesis is that shown in (5.19) below, where NAME is the name of the rhetorical relation that can be signalled by the marker, Maximal_distance(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 Distance_to_salient_unit(m) is the maximum distance to the salient unit found in the rightmost position.

(5.19)
$$\begin{array}{l} rhet\_rel(\textsc{name}, i, i-1) \oplus \ldots \oplus rhet\_rel(\textsc{name}, i, i-Max(m)) \oplus \\ rhet\_rel(\textsc{name}, i+1, i-1) \oplus \ldots \oplus rhet\_rel(\textsc{name}, i+1, i-Max(m)) \oplus \\ \vdots \\ rhet\_rel(\textsc{name}, i+Dist\_sal(m)+1, i-1) \oplus \ldots \oplus \\ \quad rhet\_rel(\textsc{name}, i+Dist\_sal(m)+1, i-Max(m)) \end{array}$$

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.

```
RR_d := NULL;
for i from 1 to n
  for each marker m ∈ D_d that belongs to U[i] and that
        relates units having the same type as those in U
    if Where_to_link(m) = BEFORE
      rr := NULL;
      l := i − 1;
      while (l ≥ 0 ∧ i − l ≤ Maximal_distance(m))
          r := i;
          while (r ≤ n ∧ r − i ≤ Distance_to_salient_unit(m) + 1)
              if (Statuses(m) = SATELLITE_NUCLEUS)
                rr := rr ⊕ rhet_rhel(name(d), l, r);
              else
                rr := rr ⊕ rhet_rhel(name(d), r, l);
              r := r + 1;
          l := l − 1;
    else
      rr := NULL;
      r := i + 1;
      while (r ≤ n ∧ r − i ≤ Maximal_distance(m))
          l := i;
          while (l ≥ 0 ∧ i − l ≤ Distance_to_salient_unit(m) + 1)
              if (Statuses(m) = SATELLITE_NUCLEUS)
                rr := rr ⊕ rhet_rhel(name(d), l, r);
              else
                rr := rr ⊕ rhet_rhel(name(d), r, l);
              l := l − 1;
          r := r + 1;
    endif
  RR_d := RR_d ∪ {rr};
  endfor
endfor
```

Figure 5.4: The discourse-marker-based hypothesizing algorithm

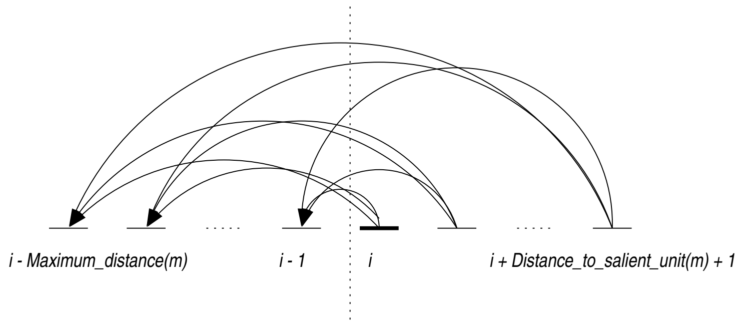

Figure 5.5: A graphical representation of the disjunctive hypothesis that is generated by the discourse-marker-based hypothesizing algorithm for a discourse marker m that belongs to unit i and that signals a rhetorical relation whose nucleus comes before the satellite.

Essentially, the disjunctive hypothesis enumerates relations of type NAME over members of the Cartesian product $\{i, i+1, \ldots, i + \text{Distance\_to\_salient\_unit}(m) + 1\} \times \{i - \text{Maximum\_distance}(m), i - \text{Maximum\_distance}(m) + 1, \ldots, i-1\}$, i.e., all the pairs of units that are separated by an imaginary line drawn between units $i-1$ and i (see figure 5.5). 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 5.4 automatically builds disjunctive hypotheses of the kind shown in (5.19) 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 5.5 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 spans that contains unit i and a unit that comes after it.

5.4.3 A word co-occurrence-based algorithm for hypothezing rhetorical relations

The rhetorical relations hypothesized by the discourse-marker-based algorithm rely entirely on occurrences of discourse markers. In the building of the valid text structures of sentences, the set of rhetorical relations that are hypothesized on the basis of discourse marker occurrences provides sufficient information. After all, the clause-like units of a sentence are determined on the basis of discourse marker occurrences as well; so every unit of a sentence is related to at least one other unit of the same sentence. 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 co-occurrences 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. 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 counting the number of words that co-occur in both textual units. If the number of word co-occurrences 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 co-occurrence-based hypothesizing algorithm are shown in figure 5.6. 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 correpond to these units. The suffixes of the remaining words are removed as well (see

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** $(l \geq 0 \wedge i - l \leq LD)$
8. $r := i + 1$;
9.   **while**$(r \leq n \wedge r - i \leq RD)$
10.   **if** (numberWordCoOccurrences(cleanedUp(l),cleanedUp(r)) $>$
11.   UnitThreshHold)
12. $rr := rr \oplus rhet_rel($ELABORATION$, r, l)$;
13.   **else**
14. $rr := rr \oplus rhet_rel($JOINT$, l, r)$;
15. $r = r + 1$;
16. $l = l - 1$;
17.   **endif**
18. $RR_c = RR_c \cup \{rr\}$;
19. **endfor**
20. $RR := RR_d \cup RR_c$;

Figure 5.6: The word co-occurrence-based hypothesizing algorithm.

function "cleanedUp" on line 9 in figure 5.6), so that words that have the same root could be considered to co-occur even in the cases in which they are used in different cases, moods, tenses, etc. If the number of co-occurrences of root words is greater than a certain threshold, an ELABORATION 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 depends on the type of the textual units that are under scrutiny and the number of units in the sequence. I have experimented with a range of different values and noticed that when the number of sentences or the number of paragraphs in a section is small, it is likely that the rhetorical relation that holds between two adjacent units is ELABORATION (this corresponds to a threshold of value -1). For longer paragraphs and sections, I consider two sentences to be related if the number of co-occurrences is larger than 1; and two paragraphs to be related if the number of co-occurrences is larger than 6.

5.4.4 Hypothesizing rhetorical relations — an example

Let us consider, again, text (5.2). Given the textual units and the discourse markers that were identified by the clause-like unit and discourse-marker identification algorithm (see text (5.17)), we now examine the relations that are hypothesized by the discourse-marker- and word co-occurrence-based hypothesizing algorithms at the sentence, paragraph, and section levels. Text (5.17) has three sentences that have more than one elementary unit. For the sentence shown in (5.20), the discourse-marker-based algorithm hypothesizes the disjunction shown in (5.21). This hypothesis is consistent with the information given in table 5.6, which shows that, in the corpus, the marker "With" consistently signalled BACKGROUND and JUSTIFICATION relations between a satellite, the unit that contained the marker, and a nucleus, the unit that followed it.

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

$$rhet\_rel(\textsc{background}, 1, 2) \oplus rhet\_rel(\textsc{justification}, 1, 2) \tag{5.21}$$

For the sentence shown in (5.22), the discourse-marker-based algorithm hypothesizes the two disjunctions shown in (5.23).

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

$$\left\{\begin{array}{l} rhet\_rel(\textsc{contrast}, 4, 5) \oplus rhet\_rel(\textsc{contrast}, 4, 6) \\ rhet\_rel(\textsc{cause}, 6, 4) \oplus rhet\_rel(\textsc{evidence}, 6, 4) \oplus \\ \quad rhet\_rel(\textsc{cause}, 6, 5) \oplus rhet\_rel(\textsc{evidence}, 6, 5) \end{array}\right. \tag{5.23}$$

This hypothesis is consistent with the information given in table 5.6 as well: *but* signals a CONTRAST between the clause-like unit that contains the marker and a unit that went before; however, it is also possible that this relation 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 $rhet\_rel(\textsc{contrast}, 4, 6)$ is hypothesized as well. The second disjunct concerns the marker *because*, which can signal either a CAUSE or an EVIDENCE relation.

For sentence (5.24), there is only one rhetorical relation that is hypothesized, that shown in (5.25).

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

$$rhet\_rel(\textsc{concession}, 7, 8) \tag{5.25}$$

Text (5.17) 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 co-occurrence-based algorithm, no stemmed words are found to co-occur in the first two sentences, but the stem *sun* is found to co-occur in the first and third sentences. Therefore, the algorithm hypothesizes the first disjunct in (5.26). 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 (5.26).

$$\left\{\begin{array}{l} rhet\_rel(\textsc{joint}_1, [1,2], 3) \oplus rhet\_rel(\textsc{elaboration}_1, [4,6], 3) \\ rhet\_rel(\textsc{elaboration}_2, [4,6], 3) \oplus rhet\_rel(\textsc{joint}_2, [1,2], 3) \end{array}\right. \tag{5.26}$$

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 co-occurrence-based algorithm are those shown in (5.27).

$$\left\{\begin{array}{l} rhet\_rel(\textsc{elaboration}, 3, [1,2]) \\ rhet\_rel(\textsc{elaboration}, [4,6], 3) \end{array}\right. \tag{5.27}$$

In contrast with the situation discussed with respect to the first paragraph of text (5.17), the second paragraph uses markers that provide enough information for linking 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 5.6, which specifies that a rhetorical relation of EXAMPLE holds between a satellite, the sentence that contains the marker, and a nucleus, the sentence that went before. However, the satellite of the relation 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 (5.28) 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 co-occurrence-based algorithm makes no further predictions.

(5.28) $$\left\{ \begin{array}{l} rhet\_rel(\text{EXAMPLE}, 9, [7,8]) \oplus rhet\_rel(\text{EXAMPLE}, 10, [7,8]) \\ rhet\_rel(\text{ANTITHESIS}, 9, 10) \oplus rhet\_rel(\text{ANTITHESIS}, [7,8], 10) \end{array} \right.$$

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 text (5.17), that shown in (5.29), below.

(5.29) $$rhet\_rel(\text{ELABORATION}, [7,10], [1,6])$$

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

5.5 Building valid text structures with disjunctive rhetorical relations

5.5.1 Preamble

The paradigms and algorithms that were developed in chapter 3 assumed that the input to the problem of text structure derivation was a sequence of elementary textual units and the set of simple and extended rhetorical relations that held among these units (see definition 2.2). However, as we discussed in the beginning of this chapter, the surface-form methods that the rhetorical parser employs cannot determine exactly the rhetorical relations that hold among textual units. Rather, these methods make exclusively disjunctive hypotheses. From this perspective, the problem of text structure derivation can be then reformulated as follows:

Definition 5.1. An extended formulation of the problem of text structure derivation — the disjunctive case: *Given a sequence of textual units $U = u_1, u_2, \ldots, u_n$ and a set RR of simple, extended, and disjunctive rhetorical relations that hold among these units and among textual spans that are defined over U, find all valid text structures of U.*

Disjunctive hypotheses can be immediately integrated into the algorithms that derive valid text structures by means of model-theoretic techniques because they are nothing but a set of logical constraints. However, the experiments described in chapter 3 suggest that the most efficient algorithms are those that employ proof-theoretic techniques and that compile the problem of text structure derivation into a grammar in Chomsky normal form. When the input to the problem of text structure derivation contains exclusively disjunctive hypotheses, the efficient algorithms described in chapter 3 cannot be applied directly. We discuss now how these algorithms can be modified so that they can derive valid text structures in the presence of disjunctive rhetorical relations.

5.5.2 A proof-theoretic approach to deriving valid text structures — the disjunctive case

The proof-theoretic approach that I discussed in section 3.4 needs only a few cosmetic changes in order to support disjunctive hypotheses. These changes concern the treatment of the set rr of rhetorical relations that is available to extend a given tree. Let us focus on one of the axioms that were given in section 3.4, for example, axiom (3.99), which is reproduced here for convenience, in (5.30), below.

$$\begin{aligned}(5.30)\quad & [S(l, b, tree_1(\textsc{nucleus}, type_1, p_1, \mathit{left}_1, right_1), rr_1) \wedge \\ & S(b+1, h, tree_2(\textsc{nucleus}, type_2, p_2, \mathit{left}_2, right_2), rr_2) \wedge \\ & rhet\_rel(name, n_1, n_2) \in rr_1 \cap rr_2 \wedge n_1 \in p_1 \wedge n_2 \in p_2 \wedge paratactic(name)] \rightarrow \\ & \quad S(l, h, tree(\textsc{nucleus}, name, p_1 \cup p_2, tree_1(\ldots), tree_2(\ldots)), \\ & \quad\quad rr_1 \cap rr_2 \setminus \{rhet\_rel(name, n_1, n_2)\})\end{aligned}$$

Axiom (5.30) specifies that if there exists a span from unit l to unit b that is characterized by valid text structure $tree_1(\ldots)$ and rhetorical relations rr_1 and another span from unit $b+1$ to unit h that is characterized by valid text structure $tree_2(\ldots)$ and rhetorical relations rr_2; if rhetorical relation $rhet\_rel(name, n_1, n_2)$ holds between a unit n_1 that is among the promotion units of span $[l, b]$ and a unit n_2 that is among the promotion units of span $[b+1, h]$; if $rhet\_rel(name, n_1, n_2)$ can still be used to extend both spans $[l, b]$ and $[b+1, h]$ ($rhet\_rel(name, n_1, n_2) \in rr_1 \cap rr_2$); and if that the relation is paratactic, then one can combine spans $[l, b]$ and $[b+1, h]$ into a larger span $[l, h]$ that has a structure whose status is NUCLEUS, type $name$, promotion set $p_1 \cup p_2$, and whose substructures are given by the structures of the immediate subspans. The set of rhetorical relations that can be used to further extend this structure is given by $rr_1 \cap rr_2 \setminus \{rhet\_rel(name, n_1, n_2)\}$.

In order for an axiom like (5.30) to be applicable in the case in which the set of rhetorical relations rr contains disjunctive hypotheses, we need to understand how the $\in, \cap$, and $\setminus$ set operations are affected by the disjunctions. Let us assume, for example, that we want to derive the valid structures of a text that has three units, which are labelled from 1 to 3, and that the rhetorical relations shown in (5.31) below hold among the units in the text.

$$(5.31)\quad RR = \left\{ \begin{array}{l} rhet\_rel(\textsc{contrast}, 1, 2) \oplus rhet\_rel(\textsc{contrast}, 1, 3) \\ rhet\_rel(\textsc{elaboration}, 3, 1) \end{array} \right.$$

Assume that we have already derived valid text structures for the elementary units 1 and 2 and that we want to use an axiom similar to (5.30) in order to derive a text structure for span [1,2]. Assume that we use $rhet\_rel(\textsc{contrast}, 1, 2)$ to create a span over units 1 and 2, and that we do not delete from the list of rhetorical relations that are still available to extend the span $[1, 2]$ the disjunction $rhet\_rel(\textsc{contrast}, 1, 2) \oplus rhet\_rel(\textsc{contrast}, 1, 3)$, but merely the relation that has been used. In such a case, we could still use $rhet\_rel(\textsc{contrast}, 1, 3)$ later, in order to join span $[1, 2]$ with unit 3, thus obtaining the tree in figure 5.7, which is obviously incorrect because it uses the same relation twice.

In order to apply the proof-theoretic-based approach described in section 3.4 to sets of rhetorical relations that contain disjunctive hypotheses, we need only to redefine the simple set operations $\in$ and $\setminus$ so that they can handle exclusive disjunctions. The new operations

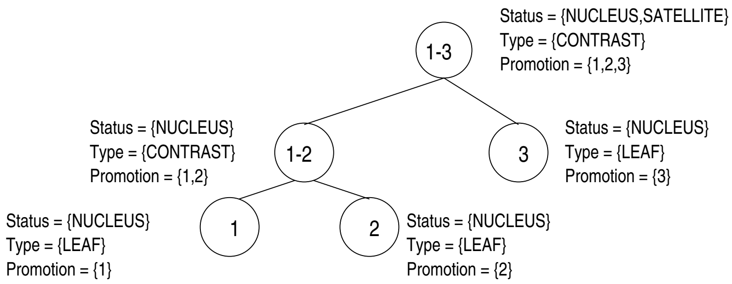

Figure 5.7: Example of invalid text structure.

are labelled by the symbols $\in_\oplus$ and $\setminus_\oplus$. In explaining their semantics, we use the sets of rhetorical relations shown in (5.32) and (5.33) below.

(5.32) $$rr_1 = \begin{cases} rhet\_rel(\textsc{contrast}, 1, 2) \oplus rhet\_rel(\textsc{contrast}, 1, 3) \\ rhet\_rel(\textsc{elaboration}, 3, 1) \end{cases}$$

(5.33) $$rr_2 = \begin{cases} rhet\_rel(\textsc{contrast}, 1, 2) \\ rhet\_rel(\textsc{elaboration}, 3, 1) \\ rhet\_rel(\textsc{concession}, 2, 3) \end{cases}$$

Definition 5.2. *The expression $rhet\_rel(name, s, n) \in_\oplus rr$ holds if and only if $rhet\_rel(name, s, n)$ occurs in set rr either as a simple or extended relation, or as one of the disjuncts of an exclusive disjunction of rhetorical relations.*

For example, the following relations hold.

$$rhet\_rel(\textsc{contrast}, 1, 2) \in_\oplus rr_1$$
$$rhet\_rel(\textsc{contrast}, 1, 2) \in_\oplus rr_2$$

Definition 5.3. *The elements that remain in a set of rhetorical relations after the operation $\setminus_\oplus$ that takes $\{rhet\_rel(name, s, n)\}$ as second argument are the simple, extended, and disjunctive rhetorical relations that are not equal to $rhet\_rel(name, s, n)$ and that do not have a disjunct equal to $rhet\_rel(name, s, n)$. In the case in which one of the disjuncts is $rhet\_rel(name, s, n)$, the whole collection of related disjuncts is eliminated from the set.*

For example, the following relations hold.

$$rr_1 \setminus_\oplus \{rhet\_rel(contrast, 1, 2)\} = \{rhet\_rel(\textsc{elaboration}, 3, 1)\}$$
$$rr_2 \setminus_\oplus \{rhet\_rel(contrast, 1, 2)\} = \{rhet\_rel(\textsc{elaboration}, 3, 1), rhet\_rel(\textsc{concession}, 2, 3)\}$$

Using the new set operators $\in_\oplus$ and $\setminus_\oplus$, we can modify axiom (5.30) as shown in (5.34) below.

(5.34)
$$
\begin{aligned}
&[S(l, b, tree_1(\text{NUCLEUS}, type_1, p_1, left_1, right_1), rr_1) \wedge \\
&S(b+1, h, tree_2(\text{NUCLEUS}, type_2, p_2, left_2, right_2), rr_2) \wedge \\
&rhet\_rel(name, n_1, n_2) \in_\oplus rr_1 \wedge rhet\_rel(name, n_1, n_2) \in_\oplus rr_2 \wedge \\
&n_1 \in p_1 \wedge n_2 \in p_2 \wedge paratactic(name)] \rightarrow \\
&\qquad S(l, h, tree(\text{NUCLEUS}, name, p_1 \cup p_2, tree_1(\ldots), tree_2(\ldots)), \\
&\qquad\qquad rr_1 \cap rr_2 \setminus_\oplus \{rhet\_rel(name, n_1, n_2)\})
\end{aligned}
$$

Axiom (5.34) treats each exclusive disjunction as a whole, thus ensuring that no rhetorical relations occur more than once in a discourse structure.

In order to apply the proof-theoretic approach described in section 3.4 to the sets of rhetorical relations that are hypothesized by the discourse-marker- and word co-occurrence-based algorithms, we need only to rewrite all the axioms (3.91)–(3.102) in the same way that we rewrote axiom (3.99). Below, I show the complete set of axioms that handle disjunctive hypotheses.

As in section 3.4, we take instantiations of axioms (5.35), (5.36), (5.37), and (5.38) as the only atomic axioms of a system that corresponds to a sequence of N textual units and a set RR of rhetorical relations that hold among these units.

(5.35) $hypotactic(relation\_name)$

(5.36) $paratactic(relation\_name)$

(5.37) $hold(RR)$

(5.38) $unit(i)$

The complete set of axioms is given below.

(5.39) $[unit(i) \wedge hold(RR)] \rightarrow S(i, i, tree(\text{NUCLEUS}, \text{LEAF}, \{i\}, \text{NULL}, \text{NULL}), RR)$

(5.40) $[unit(i) \wedge hold(RR)] \rightarrow S(i, i, tree(\text{SATELLITE}, \text{LEAF}, \{i\}, \text{NULL}, \text{NULL}), RR)$

$$
\begin{aligned}
(5.41)\qquad & [S(l,b,tree_1(\textsc{satellite},type_1,p_1,\mathit{left}_1,right_1),rr_1) \wedge \\
& S(b+1,h,tree_2(\textsc{nucleus},type_2,p_2,\mathit{left}_2,right_2),rr_2) \wedge \\
& rhet\_rel(name,s,n) \in_{\oplus} rr_1 \wedge\ rhet\_rel(name,s,n) \in_{\oplus} rr_2 \wedge \\
& s \in p_1 \wedge n \in p_2 \wedge hypotactic(name)] \rightarrow \\
& S(l,h,tree(\textsc{nucleus},name,p_2,tree_1(\ldots),tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name,s,n)\})
\end{aligned}
$$

$$
\begin{aligned}
(5.42)\qquad & [S(l,b,tree_1(\textsc{satellite},type_1,p_1,\mathit{left}_1,right_1),rr_1) \wedge \\
& S(b+1,h,tree_2(\textsc{nucleus},type_2,p_2,\mathit{left}_2,right_2),rr_2) \wedge \\
& rhet\_rel(name,s,n) \in_{\oplus} rr_1 \wedge\ rhet\_rel(name,s,n) \in_{\oplus} rr_2 \wedge \\
& s \in p_1 \wedge n \in p_2 \wedge hypotactic(name)] \rightarrow \\
& \qquad S(l,h,tree(\textsc{satellite},name,p_2,tree_1(\ldots),tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name,s,n)\})
\end{aligned}
$$

$$
\begin{aligned}
(5.43)\qquad & [S(l,b,tree_1(\textsc{satellite},type_1,p_1,\mathit{left}_1,right_1),rr_1) \wedge \\
& S(b+1,h,tree_2(\textsc{nucleus},type_2,p_2,\mathit{left}_2,right_2),rr_2) \wedge \\
& rhet\_rel\_ext(name,l,b,b+1,h) \in_{\oplus} rr_1 \wedge \\
& rhet\_rel\_ext(name,l,b,b+1,h) \in_{\oplus} rr_2 \wedge hypotactic(name)] \rightarrow \\
& \qquad S(l,h,tree(\textsc{nucleus},name,p_2,tree_1(\ldots),tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name,l,b,b+1,h)\})
\end{aligned}
$$

$$
\begin{aligned}
(5.44)\qquad & [S(l,b,tree_1(\textsc{satellite},type_1,p_1,\mathit{left}_1,right_1),rr_1) \wedge \\
& S(b+1,h,tree_2(\textsc{nucleus},type_2,p_2,\mathit{left}_2,right_2),rr_2) \wedge \\
& rhet\_rel\_ext(name,l,b,b+1,h) \in_{\oplus} rr_1 \wedge \\
& rhet\_rel\_ext(name,l,b,b+1,h) \in_{\oplus} rr_2 \wedge hypotactic(name)] \rightarrow \\
& \qquad S(l,h,tree(\textsc{satellite},name,p_2,tree_1(\ldots),tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name,l,b,b+1,h)\})
\end{aligned}
$$

$$
\begin{aligned}
(5.45)\qquad & [S(l, b, tree_1(\textsc{nucleus}, type_1, p_1, \mathit{left}_1, right_1), rr_1) \wedge \\
& S(b+1, h, tree_2(\textsc{satellite}, type_2, p_2, \mathit{left}_2, right_2), rr_2) \wedge \\
& rhet\_rel(name, s, n) \in_{\oplus} rr_1 \wedge rhet\_rel(name, s, n) \in_{\oplus} rr_2 \wedge \\
& s \in p_2 \wedge n \in p_1 \wedge hypotactic(name)] \rightarrow \\
& \qquad S(l, h, tree(\textsc{nucleus}, name, p_1, tree_1(\ldots), tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, s, n)\})
\end{aligned}
$$

$$
\begin{aligned}
(5.46)\qquad & [S(l, b, tree_1(\textsc{nucleus}, type_1, p_1, \mathit{left}_1, right_1), rr_1) \wedge \\
& S(b+1, h, tree_2(\textsc{satellite}, type_2, p_2, \mathit{left}_2, right_2), rr_2) \wedge \\
& rhet\_rel(name, s, n) \in_{\oplus} rr_1 \wedge rhet\_rel(name, s, n) \in_{\oplus} rr_2 \wedge \\
& s \in p_2 \wedge n \in p_1 \wedge hypotactic(name)] \rightarrow \\
& \qquad S(l, h, tree(\textsc{satellite}, name, p_1, tree_1(\ldots), tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, s, n)\})
\end{aligned}
$$

$$
\begin{aligned}
(5.47)\qquad & [S(l, b, tree_1(\textsc{nucleus}, type_1, p_1, \mathit{left}_1, right_1), rr_1) \wedge \\
& S(b+1, h, tree_2(\textsc{satellite}, type_2, p_2, \mathit{left}_2, right_2), rr_2) \wedge \\
& rhet\_rel\_ext(name, b+1, h, l, b) \in_{\oplus} rr_1 \wedge \\
& rhet\_rel\_ext(name, b+1, h, l, b) \in_{\oplus} rr_2 \wedge hypotactic(name)] \rightarrow \\
& \qquad S(l, h, tree(\textsc{nucleus}, name, p_1, tree_1(\ldots), tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, b+1, h, l, b)\})
\end{aligned}
$$

$$
\begin{aligned}
(5.48)\qquad & [S(l, b, tree_1(\textsc{nucleus}, type_1, p_1, \mathit{left}_1, right_1), rr_1) \wedge \\
& S(b+1, h, tree_2(\textsc{satellite}, type_2, p_2, \mathit{left}_2, right_2), rr_2) \wedge \\
& rhet\_rel\_ext(name, b+1, h, l, b) \in_{\oplus} rr_1 \wedge \\
& rhet\_rel\_ext(name, b+1, h, l, b) \in_{\oplus} rr_2 \wedge hypotactic(name)] \rightarrow \\
& \qquad S(l, h, tree(\textsc{satellite}, name, p_1, tree_1(\ldots), tree_2(\ldots)), \\
& \qquad\qquad rr_1 \cap rr_2 \setminus_{\oplus} \{rhet\_rel(name, b+1, h, l, b)\})
\end{aligned}
$$

$$(5.49)\quad \begin{aligned}&[S(l,b,tree_1(\textsc{nucleus},type_1,p_1,left_1,right_1),rr_1)\wedge\\ &S(b+1,h,tree_2(\textsc{nucleus},type_2,p_2,left_2,right_2),rr_2)\wedge\\ &rhet\_rel(name,n_1,n_2)\in_{\oplus} rr_1 \wedge rhet\_rel(name,n_1,n_2)\in_{\oplus} rr_2\wedge\\ &n_1\in p_1\wedge n_2\in p_2\wedge paratactic(name)]\rightarrow\\ &\qquad S(l,h,tree(\textsc{nucleus},name,p_1\cup p_2,tree_1(\ldots),tree_2(\ldots)),\\ &\qquad\qquad rr_1\cap rr_2\setminus_{\oplus}\{rhet\_rel(name,n_1,n_2)\})\end{aligned}$$

$$(5.50)\quad \begin{aligned}&[S(l,b,tree_1(\textsc{nucleus},type_1,p_1,left_1,right_1),rr_1)\wedge\\ &S(b+1,h,tree_2(\textsc{nucleus},type_2,p_2,left_2,right_2),rr_2)\wedge\\ &rhet\_rel(name,n_1,n_2)\in_{\oplus} rr_1 \wedge rhet\_rel(name,n_1,n_2)\in_{\oplus} rr_2\wedge\\ &n_1\in p_1\wedge n_2\in p_2\wedge paratactic(name)]\rightarrow\\ &\qquad S(l,h,tree(\textsc{satellite},name,p_1\cup p_2,tree_1(\ldots),tree_2(\ldots)),\\ &\qquad\qquad rr_1\cap rr_2\setminus_{\oplus}\{rhet\_rel(name,n_1,n_2)\})\end{aligned}$$

$$(5.51)\quad \begin{aligned}&[S(l,b,tree_1(\textsc{nucleus},type_1,p_1,left_1,right_1),rr_1)\wedge\\ &S(b+1,h,tree_2(\textsc{nucleus},type_2,p_2,left_2,right_2),rr_2)\wedge\\ &rhet\_rel\_ext(name,l,b,b+1,h)\in_{\oplus} rr_1\wedge\\ &rhet\_rel\_ext(name,l,b,b+1,h)\in_{\oplus} rr_2\wedge paratactic(name)]\rightarrow\\ &\qquad S(l,h,tree(\textsc{nucleus},name,p_1\cup p_2,tree_1(\ldots),tree_2(\ldots)),\\ &\qquad\qquad rr_1\cap rr_2\setminus_{\oplus}\{rhet\_rel(name,l,b,b+1,h)\})\end{aligned}$$

$$(5.52)\quad \begin{aligned}&[S(l,b,tree_1(\textsc{nucleus},type_1,p_1,left_1,right_1),rr_1)\wedge\\ &S(b+1,h,tree_2(\textsc{nucleus},type_2,p_2,left_2,right_2),rr_2)\wedge\\ &rhet\_rel\_ext(name,l,b,b+1,h)\in_{\oplus} rr_1\wedge\\ &rhet\_rel\_ext(name,l,b,b+1,h)\in_{\oplus} rr_2\wedge paratactic(name)]\rightarrow\\ &\qquad S(l,h,tree(\textsc{satellite},name,p_1\cup p_2,tree_1(\ldots),tree_2(\ldots)),\\ &\qquad\qquad rr_1\cap rr_2\setminus_{\oplus}\{rhet\_rel(name,l,b,b+1,h)\})\end{aligned}$$

Axioms (5.35)–(5.52) provide a *disjunctive proof-theoretic account* of the disjunctive case of the problem of text structure derivation.

Theorem 5.1 is the sibling of theorem 3.1, which was given in section 3.4. Its proof mirrors the proof of theorem 3.1.

Theorem 5.1. *Given a text T that is characterized by a set of rhetorical relations RR that may be exclusively disjunctive, the application of the disjunctive proof-theoretic account 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-theoretic account 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 disjunctive proof-theoretic account.*

Implementing the disjunctive proof-theoretic account

There are many ways in which one can implement the set of rewriting rules described in this section. My rhetorical parser implements the disjunctive proof-theoretic account as a chart-parsing algorithm. 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 labelled 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 $\text{N} + 1$ 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 (5.53), below.

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

$S \rightarrow S\ \ S$

The grammar rules that are associated with the chart might be only partially completed. We use the traditional bullet symbol $\bullet$ 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 $[0, \text{N}, S \rightarrow S\, S\, \bullet, vn_2, r_2]$.

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
$S \rightarrow \bullet\, S\, S$	STARTCOMPOUND
$S \rightarrow S \bullet S$	MIDDLECOMPOUND
$S \rightarrow S\, S\, \bullet$	ENDCOMPOUND

The chart-parsing algorithm that implements the disjunctive proof-theoretic account for deriving text structures in given in figure 5.8. 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.[1] 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 (5.39)–(5.52). 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 (5.39)–(5.52). 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.

[1]The rhetorical parser assumes that the input has at least two units.

```
Input: A sequence U = 1, 2, ..., 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.

function CHART-PARSER(N, RR)
  chart := nil;
  INITIALIZER(RR);
  for i from 1 to N
     SCANNER(i);
  return chart;

procedure ADD-EDGE(edge)
  if edge ∉ chart[EndOf(edge)]
    push edge in chart[EndOf(edge)];
    if GrammarRuleOf(edge) ∈ {ENDUNIT,ENDCOMPOUND})
      COMPLETER(edge);
    else
      PREDICTOR(edge);

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

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

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

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

Figure 5.8: A chart-parsing algorithm that implements the disjunctive proof-theoretic account of building valid text structures.

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.

5.5.3 Deriving valid text structures through compilation of grammars in Chomsky normal form — the disjunctive case

We have seen that, when the rhetorical relations that hold among textual units are precisely known, the valid structures of a text can be derived in polynomial time by compiling the problem of text structure derivation into a grammar in Chomsky normal form (see theorem 3.2). Unfortunately, the compiling algorithm shown in figure 3.11 is not applicable in the case the rhetorical relations that hold among textual units are exclusive disjunctions. In proving that the compiling algorithm generates a grammar that can be used to derive all and only the valid text structures of a text, we have shown that the rules of the grammar never generate text trees that use the same rhetorical relation twice. If the set RR of rhetorical relations that hold among the units in the text contains disjunctive hypotheses, this property no longer holds. Reconsider, for example, a text with three elementary units, and assume that the rhetorical relations in (5.54) hold among the units of the text.

$$(5.54)\quad \begin{cases} rhet\_rel(\textsc{contrast}, 1, 2) \oplus rhet\_rel(\textsc{contrast}, 1, 3) \\ rhet\_rel(\textsc{elaboration}, 3, 1) \end{cases}$$

If we use the compiling algorithm, we obtain a grammar that contains among its rules, those shown in (5.55).

$$(5.55)\quad \begin{cases} S\langle 1, 1, \textsc{nucleus}, \textsc{leaf}, \{1\}\rangle \rightarrow 1 \\ S\langle 2, 2, \textsc{nucleus}, \textsc{leaf}, \{2\}\rangle \rightarrow 2 \\ S\langle 3, 3, \textsc{nucleus}, \textsc{leaf}, \{3\}\rangle \rightarrow 3 \\ S\langle 1, 2, \textsc{nucleus}, \textsc{contrast}, \{1\}\rangle \rightarrow S\langle 1, 1, \textsc{nucleus}, \textsc{leaf}, \{1\}\rangle \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad S\langle 2, 2, \textsc{nucleus}, \textsc{leaf}, \{2\}\rangle \\ S\langle 1, 3, \textsc{nucleus}, \textsc{contrast}, \{1\}\rangle \rightarrow S\langle 1, 2, \textsc{nucleus}, \textsc{contrast}, \{1\}\rangle \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad S\langle 3, 3, \textsc{nucleus}, \textsc{leaf}, \{3\}\rangle \\ S \rightarrow S\langle 1, 2, \textsc{nucleus}, \textsc{contrast}, \{1\}\rangle \quad S\langle 3, 3, \textsc{nucleus}, \textsc{leaf}, \{3\}\rangle \end{cases}$$

If we apply the rules in (5.55) on the input $1, 2, 3$, we obtain a parse tree that corresponds to the invalid text structure in figure 5.7, which uses the rhetorical relation CONTRAST twice. This happens because the disjunctive relation in (5.54) is relevant in the sense of definition (2.8) both to spans $[1, 2]$ and $[1, 3]$. In contrast, in the case of non-disjunctive relations, when a rhetorical relation r was used to join two textual spans in a larger span $[l, h]$, it was guaranteed that relation r could not be used to join span $[l, h]$ with other adjacent spans.

It follows that if we are to use a grammar-based approach to deriving text structures, we need to provide mechanisms to prevent the use of a rhetorical relation more than once in a derivation. We do this by assigning to each nonterminal symbol of the grammar an extra index. Hence, instead of nonterminals of the form $S\langle x, y, status, type, promotion\_set\rangle$, we are going to use nonterminals of the form $S\langle x, y, status, type, promotion\_set, used\_relations\rangle$, where *used_relations* is the set of rhetorical relations that are used in a parse that has $S\langle x, y, status, type, promotion\_set, used\_relations\rangle$ as its root. The new algorithm that derives text structures by means of a grammar in Chomsky normal form relies on the same facts as the one in section 3.5. That is, it still uses the fact that valid text structures can be recovered from an "almost-valid" text structure, i.e., a structure that associates only one unit with each promotion set. And it still takes advantage of the fact that the number of nonterminal symbols of type $S\langle x, y, status, type, promotion\_set, used\_relations\rangle$ is finite. Since the *status* of a valid span ranges over a set of cardinality 2, {NUCLEUS, SATELLITE}, the *type* over a set of $k_{[x,y]} < |RR|$ relations that are relevant to span $[x, y]$, the *promotion_set* over the elements of the set $\{\{x\}, \{x+1\}, \ldots, \{y\}\}$, and the *used_relations* over $\binom{|RR|}{y-x}$ possible combinations of rhetorical relations that are members of the initial set RR of cardinality $|RR|$, it follows that there are at most $2k_{[x,y]}(y-x+1)\binom{|RR|}{y-x}$ nonterminal symbols for each span $[x, y]$ that plays an active role in the structure of a text.

Theorem 5.2. *Consider a sequence of textual units* $1, 2, \ldots, \text{N}$ *and a set* RR *that encodes all the relations that hold among these units. The relations can be simple, extended, and disjunctive. The* disjunctive compiling algorithm *in figure 5.9 generates a Chomsky normal-form grammar that can be used to derive all and only the parse trees that are isomorphic with the valid text structures of text* $1, 2, \ldots, \text{N}$.

Sketch of the proof. The proof of theorem 5.2 is similar to that of theorem 3.2. We sketch here only its main steps.

The disjunctive compiling algorithm in figure 5.9 derives all the grammar rules that correspond to building spans of size 1, 2, 3, and so on, up to N. It does so by considering for each span $[l, h]$, all the possible ways in which the span can be broken into two adjacent subspans and all the possible relations from the initial set RR that hold across the two subspans. For each relation r that holds across the adjacent subspans $[l, b]$ and $[b+1, h]$, if the relation has not been used in the derivation of the nonterminals that characterize spans $[l, b]$ and $[b+1, h]$, the algorithm generates all the grammar rules that enforce the strong compositionality criterion: that is, the algorithm considers all pairs of nonterminals that characterize spans $[l, b]$ and $[b+1, h]$ and generates rules for each such pair.

A simple inspection of the rules generated by the disjunctive compiling algorithm shows that they enforce the compositionality criterion with respect to the statuses, types, and

Input: A sequence $1, 2, \ldots, \textsc{n}$ of elementary textual units.
A set RR of rhetorical relations that hold among these units.
Output: A grammar in Chomsky normal form that can be used to derive all and only the parse trees that correspond to the valid text structures of U.

for i **from** 1 **to** $\textsc{n}$
add rule $S \rightarrow i$
add rules $S\langle i, i, \textsc{nucleus}, \textsc{leaf}, \{i\}, \emptyset\rangle \rightarrow i$ and $S\langle i, i, \textsc{satellite}, \textsc{leaf}, \{i\}, \emptyset\rangle \rightarrow i$
endfor
for *size_of_span* **from** 1 **to** $\textsc{n} - 1$
for l **from**1 to $n - size\_of\_span$
$h := l + size\_of\_span;$
for b **from** l **to** $h - 1$
for x **from** l **to** b
for y **from** $b + 1$ **to** h
for each $name_1$ for which a rule has $S\langle l, b, \textsc{satellite}, name_1, \{x\}, r_1\rangle$ as head
for each $name_2$ for which a rule has $S\langle b + 1, h, \textsc{nucleus}, name_2, \{y\}, r_2\rangle$ as head
for each hypotactic relation *name* such that
$(r = rhet\_rel(name, x, y) \in_{\oplus} RR \vee$
$r = rhet\_rel(name, l, b, b + 1, h) \in_{\oplus} RR \vee$
$r = rhet\_rel(name, x, y) \oplus \ldots \oplus rhet\_rel(name_k, x_k, y_k) \in_{\oplus} RR) \wedge r \notin_{\oplus} rr_1 \cup rr_2$
add rule $S \rightarrow S\langle l, b, \textsc{satellite}, name_1, \{x\}, r_1\rangle, S\langle b + 1, h, \textsc{nucleus}, name_2, \{y\}, r_2\rangle$
add rule $S\langle l, h, \textsc{satellite}, name, \{y\}, r_1 \cup r_2 \cup \{r\}\rangle \rightarrow$
$S\langle l, b, \textsc{satellite}, name_1, \{x\}, r_1\rangle, S\langle b + 1, h, \textsc{nucleus}, name_2, \{y\}, r_2\rangle$
add rule $S\langle l, h, \textsc{nucleus}, name, \{y\}, r_1 \cup r_2 \cup \{r\}\rangle \rightarrow$
$S\langle l, b, \textsc{satellite}, name_1, \{x\}, r_1\rangle, S\langle b + 1, h, \textsc{nucleus}, name_2, \{y\}, r_2\rangle$
endfor
endfor
endfor
for each $name_1$ for which a rule has $S\langle l, b, \textsc{nucleus}, name_1, \{x\}, r_1\rangle$ as head
for each $name_2$ for which a rule has $S\langle b + 1, h, \textsc{satellite}, name_2, \{y\}, r_2\rangle$ as head
foreach hypotactic relation *name* such that
$(r = rhet\_rel(name, y, x) \in_{\oplus} RR \vee$
$r = rhet\_rel(name, b + 1, h, l, b) \in_{\oplus} RR \vee$
$r = rhet\_rel(name, y, x) \oplus \ldots \oplus rhet\_rel(name_k, y_k, x_k) \in_{\oplus} RR) \wedge r \notin_{\oplus} rr_1 \cup rr_2$
add rule $S \rightarrow S\langle l, b, \textsc{nucleus}, name_1, \{x\}, r_1\rangle, S\langle b + 1, h, \textsc{satellite}, name_2, \{y\}, r_2\rangle$
add rule $S\langle l, h, \textsc{satellite}, name, \{x\}, r_1 \cup r_2 \cup \{r\}\rangle \rightarrow$
$S\langle l, b, \textsc{nucleus}, name_1, \{x\}, r_1\rangle, S\langle b + 1, h, \textsc{satellite}, name_2, \{y\}, r_2\rangle$
add rule $S\langle l, h, \textsc{nucleus}, name, \{x\}, r_1 \cup r_2 \cup \{r\}\rangle \rightarrow$
$S\langle l, b, \textsc{nucleus}, name_1, \{x\}, r_1\rangle, S\langle b + 1, h, \textsc{satellite}, name_2, \{y\}, r_2\rangle$
endfor
endfor
endfor

Figure 5.9: A disjunctive compiling algorithm that converts the disjunctive case of the problem of text structure derivation into a Chomsky normal-form grammar (see continuation in figure 5.10).

for y **from** $b+1$ **to** h

$\vdots$

for each $name_1, r_1$ for which a rule has $S\langle l, b, \textsc{nucleus}, name_1, \{x\}, r_1\rangle$ as head
for each $name_2, r_2$ for which a rule has $S\langle b+1, h, \textsc{nucleus}, name_2, \{y\}, r_2\rangle$ as head
for each paratactic relation $name$ such that
$(r = rhet\_rel(name, x, y) \in_{\oplus} RR \vee$
$r = rhet\_rel(name, l, b, b+1, h) \in_{\oplus} RR \vee$
$r = rhet\_rel(name, x, y) \oplus \ldots \oplus rhet\_rel(name_k, x_k, y_k) \in_{\oplus} RR) \wedge r \notin_{\oplus} rr_1 \cup rr_2$
add rule $S \rightarrow S\langle l, b, \textsc{nucleus}, name_1, \{x\}, r_1\rangle, S\langle b+1, h, \textsc{nucleus}, name_2, \{y\}, r_2\rangle$
add rule $S\langle l, h, \textsc{satellite}, name, \{x\}, r_1 \cup r_2 \cup \{r\}\rangle \rightarrow$
$S\langle l, b, \textsc{nucleus}, name_1, \{x\}, r_1\rangle, S\langle b+1, h, \textsc{nucleus}, name_2, \{y\}, r_2\rangle$
add rule $S\langle l, h, \textsc{satellite}, name, \{y\}, r_1 \cup r_2 \cup \{r\}\rangle \rightarrow$
$S\langle l, b, \textsc{nucleus}, name_1, \{x\}, r_1\rangle, S\langle b+1, h, \textsc{nucleus}, name_2, \{y\}, r_2\rangle$
add rule $S\langle l, h, \textsc{nucleus}, name, \{x\}, r_1 \cup r_2 \cup \{r\}\rangle \rightarrow$
$S\langle l, b, \textsc{nucleus}, name_1, \{x\}, r_1\rangle, S\langle b+1, h, \textsc{nucleus}, name_2, \{y\}, r_2\rangle$
add rule $S\langle l, h, \textsc{nucleus}, name, \{y\}, r_1 \cup r_2 \cup \{r\}\rangle \rightarrow$
$S\langle l, b, \textsc{nucleus}, name_1, \{x\}, r_1\rangle, S\langle b+1, h, \textsc{nucleus}, name_2, \{y\}, r_2\rangle$
end all for loops

Figure 5.10: A disjunctive compiling algorithm that converts the disjunctive case of the problem of text structure derivation into a Chomsky normal-form grammar (continuation from figure 5.9).

promotion sets of the subspans. Because, at each step, the algorithm generates only grammar rules that introduce rhetorical relations that have not been used before, no derivation will use a rhetorical relation more than once.

The algorithm generates rules that correspond to all possible ways in which two textual spans can be put together into a valid text structure. Each of these rules is valid, so by induction, it immediately follows that the parse trees on a given input correspond to valid text structures: hence, the disjunctive compiling algorithm is sound. Because the grammar enumerates rules that correspond to all the possible ways in which text spans can be joined into larger text structures, it follows that the algorithm is also complete. □

Example

Given a sequence of three textual units $1, 2, 3$ among which the rhetorical relations shown in (5.54) hold, the disjunctive compiling algorithm generates a grammar having the rules shown in figure 5.11. These rules can be used to parse the input $1, 2, 3$ and obtain derivations such as that shown in figure 5.12. The labels in the nodes of the parse tree in figure 5.12 correspond to the disjunctive rhetorical relation shown in (5.56) and to the complete set of relations that hold among the units of the text, which was given in (5.54).

(5.56) $$rhet\_rel(\textsc{contrast}, 1, 2) \oplus rhet\_rel(\textsc{contrast}, 1, 3)$$

$$
\begin{array}{lll}
S \rightarrow 1; & S\langle 1,1,\textsc{nucleus},\textsc{leaf},\{1\},\emptyset\rangle \rightarrow 1; & S\langle 1,1,\textsc{satellite},\textsc{leaf},\{1\},\emptyset\rangle \rightarrow 1 \\
S \rightarrow 2; & S\langle 2,2,\textsc{nucleus},\textsc{leaf},\{2\},\emptyset\rangle \rightarrow 2; & S\langle 2,2,\textsc{satellite},\textsc{leaf},\{2\},\emptyset\rangle \rightarrow 2 \\
S \rightarrow 3; & S\langle 3,3,\textsc{nucleus},\textsc{leaf},\{3\},\emptyset\rangle \rightarrow 3; & S\langle 3,3,\textsc{satellite},\textsc{leaf},\{3\},\emptyset\rangle \rightarrow 3
\end{array}
$$

$$
\begin{array}{l}
S \rightarrow S\langle 1,1,\textsc{nucleus},\textsc{leaf},\{1\},\emptyset\rangle\ S\langle 2,2,\textsc{nucleus},\textsc{leaf},\{2\},\emptyset\rangle \\
S\langle 1,2,\textsc{nucleus},\textsc{contrast},\{1\},\{rhet\_rel(\textsc{contrast},1,2) \oplus rhet\_rel(\textsc{contrast},1,3)\}\rangle \\
\quad \rightarrow\ S\langle 1,1,\textsc{nucleus},\textsc{leaf},\{1\},\emptyset\rangle\ S\langle 2,2,\textsc{nucleus},\textsc{leaf},\{2\},\emptyset\rangle \\
S\langle 1,2,\textsc{satellite},\textsc{contrast},\{1\},\{rhet\_rel(\textsc{contrast},1,2) \oplus rhet\_rel(\textsc{contrast},1,3)\}\rangle \\
\quad \rightarrow\ S\langle 1,1,\textsc{nucleus},\textsc{leaf},\{1\},\emptyset\rangle\ S\langle 2,2,\textsc{nucleus},\textsc{leaf},\{2\},\emptyset\rangle \\
S\langle 1,2,\textsc{nucleus},\textsc{contrast},\{2\},\{rhet\_rel(\textsc{contrast},1,2) \oplus rhet\_rel(\textsc{contrast},1,3)\}\rangle \\
\quad \rightarrow\ S\langle 1,1,\textsc{nucleus},\textsc{leaf},\{1\},\emptyset\rangle\ S\langle 2,2,\textsc{nucleus},\textsc{leaf},\{2\},\emptyset\rangle \\
S\langle 1,2,\textsc{satellite},\textsc{contrast},\{2\},\{rhet\_rel(\textsc{contrast},1,2) \oplus rhet\_rel(\textsc{contrast},1,3)\}\rangle \\
\quad \rightarrow\ S\langle 1,1,\textsc{nucleus},\textsc{leaf},\{1\},\emptyset\rangle\ S\langle 2,2,\textsc{nucleus},\textsc{leaf},\{2\},\emptyset\rangle \\
S \rightarrow S\langle 1,2,\textsc{nucleus},\textsc{contrast},\{1\},\{rhet\_rel(\textsc{contrast},1,2)\oplus \\
\qquad\qquad\qquad\qquad\qquad\qquad rhet\_rel(\textsc{contrast},1,3)\}\rangle \\
\qquad S\langle 3,3,\textsc{satellite},\textsc{leaf},\{3\},\emptyset\rangle \\
S\langle 1,3,\textsc{nucleus},\textsc{elaboration},\{1\},\{rhet\_rel(\textsc{contrast},1,2)\oplus \\
\qquad\qquad\qquad\qquad\qquad\qquad rhet\_rel(\textsc{contrast},1,3), \\
\qquad\qquad\qquad\qquad\qquad\qquad rhet\_rel(\textsc{elaboration},3,1)\}\rangle \\
\quad \rightarrow S\langle 1,2,\textsc{nucleus},\textsc{contrast},\{1\},\{rhet\_rel(\textsc{contrast},1,2)\oplus \\
\qquad\qquad\qquad\qquad\qquad\qquad rhet\_rel(\textsc{contrast},1,3)\}\rangle \\
\qquad S\langle 3,3,\textsc{satellite},\textsc{leaf},\{3\},\emptyset\rangle \\
S\langle 1,3,\textsc{satellite},\textsc{elaboration},\{1\},\{rhet\_rel(\textsc{contrast},1,2)\oplus \\
\qquad\qquad\qquad\qquad\qquad\qquad rhet\_rel(\textsc{contrast},1,3), \\
\qquad\qquad\qquad\qquad\qquad\qquad rhet\_rel(\textsc{elaboration},3,1)\}\rangle \\
\quad \rightarrow S\langle 1,2,\textsc{nucleus},\textsc{contrast},\{1\},\{rhet\_rel(\textsc{contrast},1,2)\oplus \\
\qquad\qquad\qquad\qquad\qquad\qquad rhet\_rel(\textsc{contrast},1,3)\}\rangle \\
\qquad S\langle 3,3,\textsc{satellite},\textsc{leaf},\{3\},\emptyset\rangle
\end{array}
$$

Figure 5.11: The Chomsky normal-form grammar that is derived by algorithm 5.9 for a text with three units that is characterized by rhetorical relations (5.54).

The derivation shown in figure 5.12 corresponds to the valid text structure shown in figure 5.13.

An estimation of the size of the grammar

Assume that we are given a text with N elementary units and that k relations hold on average between any two elementary units. An upper bound of the number of rules that are generated by the disjunctive compiling algorithm corresponds to the case in which all relations are paratactic (lines 39–53 in figure 5.10). Given a span $[a,b]$ and a unit $u \in \{\{a\},\{a+1\},\ldots,\{b\}\}$, there are at most k relations that could promote unit u as a salient unit and, hence, at most $k\binom{|RR|}{b-a}$ nonterminal symbols of the form $S\langle a,b,\textsc{nucleus},type,\{u\},r\rangle$, where

$S\langle 1, 3, \textsc{nucleus}, \textsc{elaboration}, \{1\}, \{(5.54)\}\rangle$

$S\langle 1, 2, \textsc{nucleus}, \textsc{contrast}, \{1\}, \{(5.56)\}\rangle$ $S\langle 3, 3, \textsc{satellite}, \textsc{leaf}, \{3\}, \emptyset\rangle$

$S\langle 1, 1, \textsc{nucleus}, \textsc{leaf}, \{1\}, \emptyset\rangle$ $S\langle 2, 2, \textsc{nucleus}, \textsc{leaf}, \{2\}, \emptyset\rangle$

Figure 5.12: A Chomsky normal-form derivation of a valid tree structure that corresponds to relations (5.54).

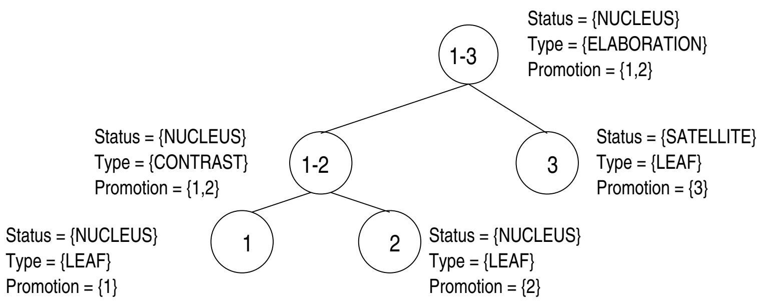

Figure 5.13: The valid text structure that corresponds to the derivation shown in figure 5.12.

$|RR|$ represents the cardinality of the initial set of rhetorical relations. It follows that lines 45–53 are executed at most $|RR|k^2\binom{|RR|}{h-l}$ times. Hence, the disjunctive compiling algorithm generates a grammar G with at most $|G|$ rules, where $|G|$ is given by the expression below.

$$|G| = 3\textsc{n} + \sum_{1\le s<\textsc{n}} \sum_{1\le l\le \textsc{n}-s} \sum_{l\le b<l+s} \sum_{l\le x\le b} \sum_{b+1\le y\le l+s} 5k^2|RR|\binom{|RR|}{h-l} \tag{5.57}$$

If we take as upper bound for $|RR|$ the value $3\textsc{n}$, this gives an exponential number of grammar rules ($O(2^{3\textsc{n}})$). Hence, in the worst case, the disjunctive compiling algorithm generates an exponential number of grammar rules. This result suggests that if the rhetorical relations that hold among the elementary units of a text are disjunctive, determining all the valid structures of a text might require exponential time.

5.5.4 Deriving valid text structures — an example

The rhetorical parsing algorithm shown in figure 5.1 employs in step III.3 the chart-parsing algorithm that implements the disjunctive proof-theoretic account, which was shown in figure 5.8. When the chart-parsing algorithm uses as input the rhetorical relations that were hypothesized by the discourse-marker- and word co-occurrence-based algorithms at the sentence, paragraph, and section levels of text 5.17, it derives the valid text structures shown in figures 5.14–5.19.

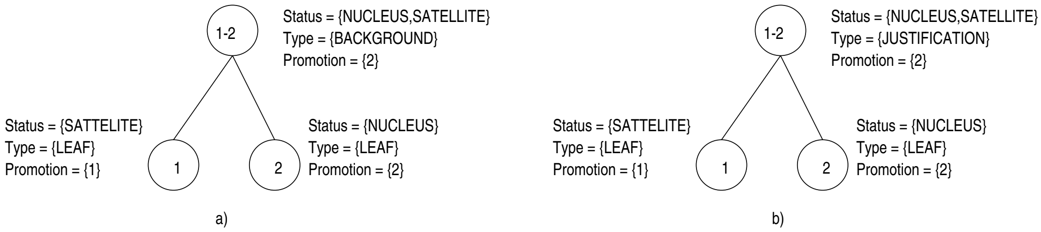

Figure 5.14: The valid text structures of sentence (5.20).

5.6 The ambiguity of discourse

5.6.1 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 (5.22) (see figure 5.15). In my experiments, I noticed, at least for English, that the "best" discourse trees are usually 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 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 (5.58), 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.

$$(5.58)\quad w(tree) = \begin{cases} 0 & \text{if } \mathit{isLeaf}(tree), \\ w(\mathit{leftOf}(tree)) + w(\mathit{rightOf}(tree)) + & \text{otherwise.} \\ \quad \mathit{depth}(\mathit{rightOf}(tree)) - \mathit{depth}(\mathit{leftOf}(tree)) & \end{cases}$$

For example, when applied to the valid text structures of sentence (5.22), the weight function

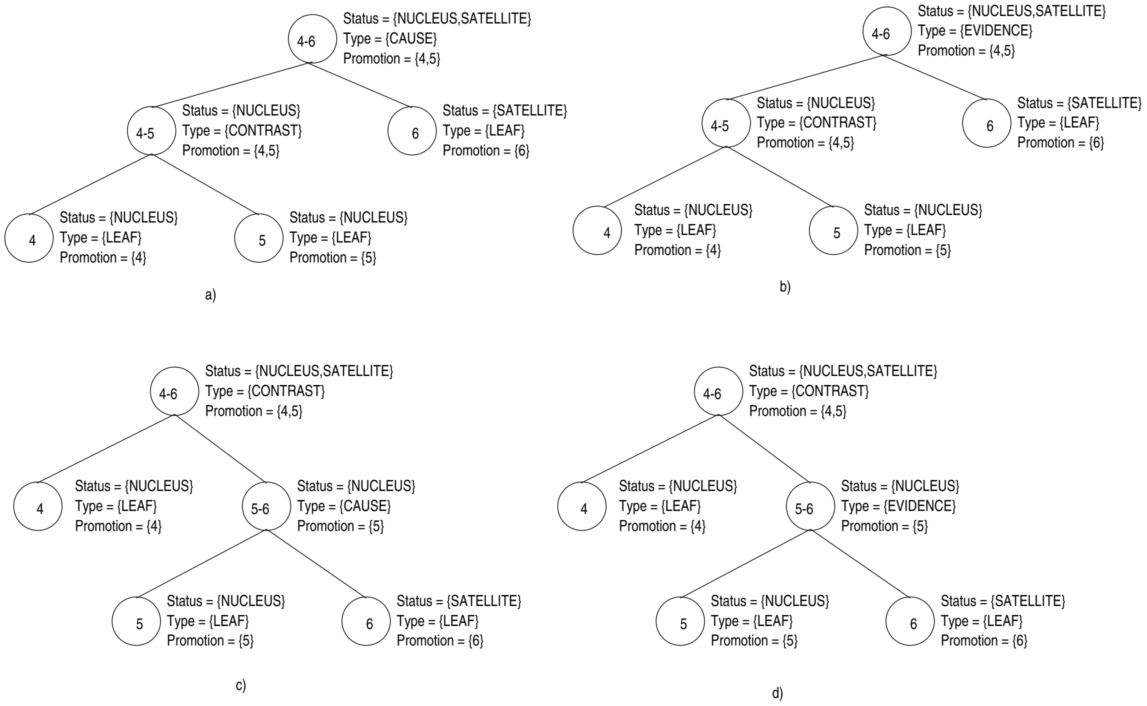

Figure 5.15: The valid text structures of sentence (5.22).

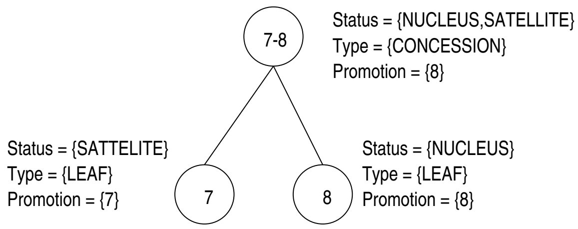

Figure 5.16: The valid text structure of sentence (5.24).

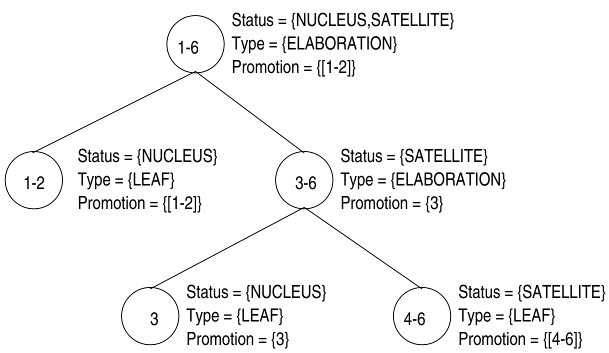

Figure 5.17: The valid text structure of the first paragraph of text (5.17) (see relations (5.27)).

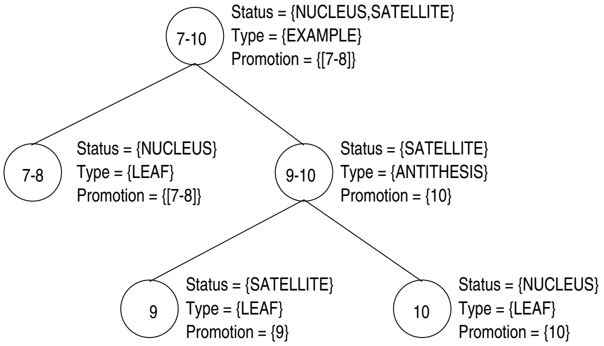

Figure 5.18: The valid text structure of the second paragraph of text (5.17) (see relations (5.28)).

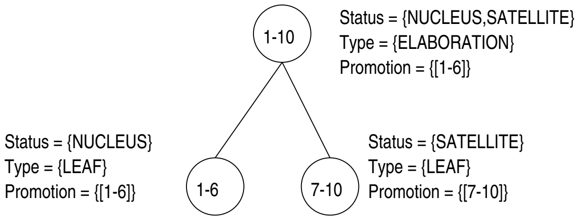

Figure 5.19: The valid text structure of text (5.17) (see relation (5.29)).

will assign the value -1 to the trees shown in figures 5.15.a and 5.15.b, and the value $+1$ to the trees shown in figures 5.15.c and 5.15.d.

5.6.2 The ambiguity of discourse — an implementation perspective

There are two ways one can disambiguate discourse. One way is to consider, during the parsing process, all of the valid 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 algorithm 5.8 is used, we can keep in the chart only the partial structures that could lead to a final structure of maximal weight.

In step III.4, the rhetorical parser shown in figure 5.1 implements the second approach. Hence, instead of keeping in the chart all the partial structures that characterize sentence (5.22), it will keep only the partial structures of maximal weight, i.e., the structures shown in figures 5.15.c and 5.15.d. In this way, the overall efficiency of the system is increased. In order to keep in the chart only the partial structures that could lead to valid structures of maximal weight, we need to modify only the procedure AddEdge in figure 5.8 so that it pushes an *edge* into the *chart* only if the *edge* corresponds to a partial structure that has a greater weight than any other partial structure that promotes the same units with respect to the span under consideration. In this case, pushing an *edge* into the *chart* is also accompanied by the deletion of the edges that span the same units, have the same promotion units, and have lower weights.

When more than one valid text structure has the same maximal weight, the rhetorical parser chooses randomly one of the structures of maximal weight at each of the three levels: sentence, paragraph, and section. For example, when the rhetorical parser selects the trees of maximal weight for text (5.17) at each of the three levels of abstraction, it selects the trees shown in figures 5.14.a, 5.15.c, 5.16, 5.17, 5.18, and 5.19. If no weight function were used, the rhetorical parser would generate eight distinct valid text structures for the whole text.

5.7 Deriving the final text structure

In the last step (lines 16–17 in figure 5.1), 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 correpond 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 text (5.2), the rhetorical parser builds the text structure shown in figure 5.20. The convention that I use is that nuclei are surrounded by solid boxes and satellites by dotted boxes; 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 by curly brackets. The leaves of the discourse structure are numbered from 1 to N, where N represents the number of elementary units in the whole text. The numbers associated with each node denote the units that are members of its promotion set.

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

5.8 Discussion and evaluation

I believe that there are two ways to evaluate the correctness of the discourse trees that an automatic process builds. One is to compare the automatically derived trees with trees that have been built manually. The other is to evaluate the impact that they have on the

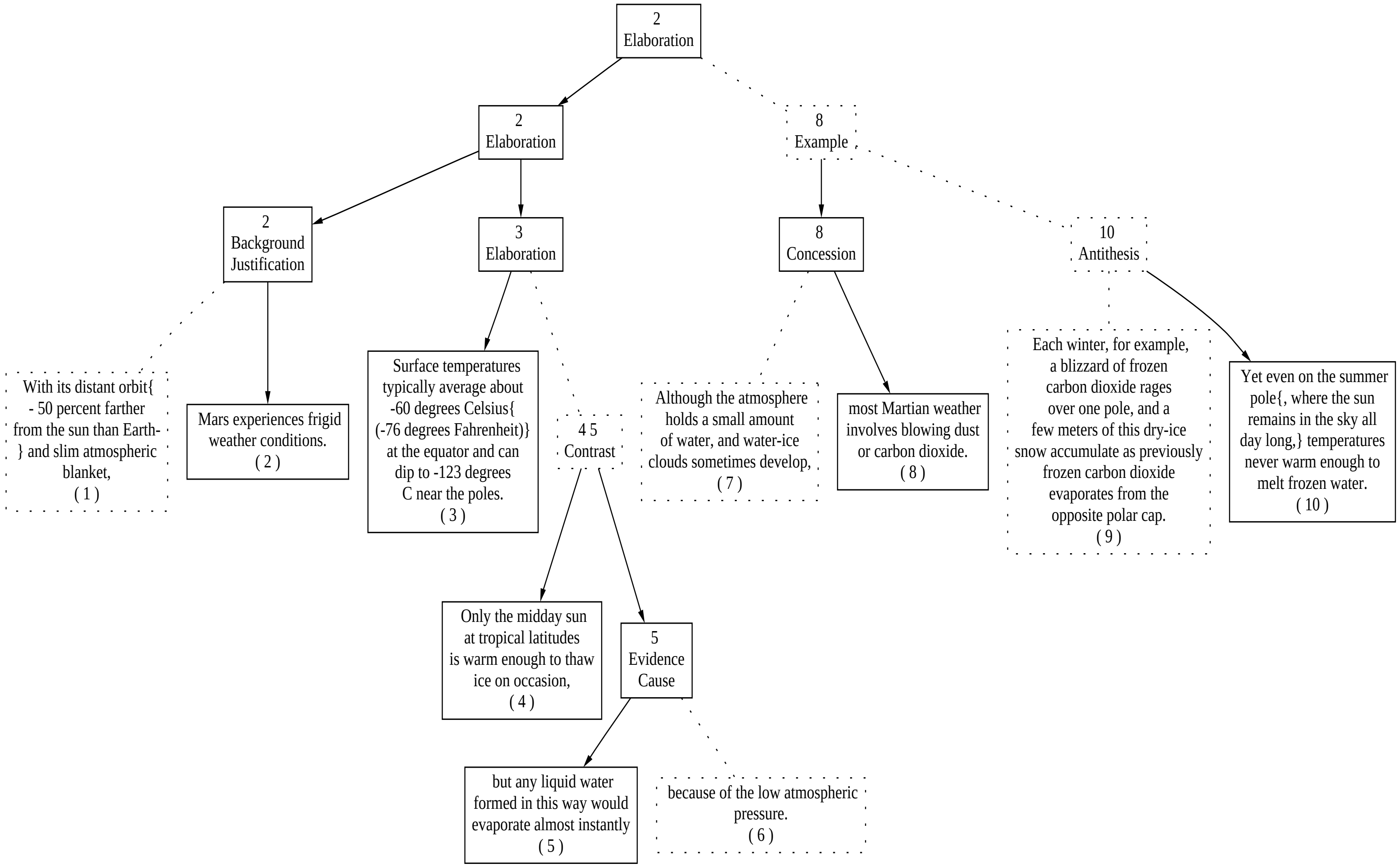

Figure 5.20: The discourse tree of maximal weight that is built by the rhetorical parsing algorithm for text (5.2).

accuracy of other natural language processing tasks, such as anaphora resolution, intention recognition, or text summarization. In this thesis, I describe evaluations that follow both these avenues.

Unfortunately, the linguistic community has not yet built a corpus of discourse trees against which rhetorical parsers can be evaluated with the effectiveness that traditional parsers are. To circumvent this problem, I asked two analysts to manually build the discourse trees for five texts that ranged from 161 to 725 words (for details, see chapter 6). Although there were some differences with respect to the names of the relations that the analysts used, the agreement with respect to the status assigned to various units (nuclei and satellites) and the overall shapes of the trees was statistically significant.

In order to measure this agreement I associated an importance score to each textual unit in a tree and computed the Spearman correlation coefficients between the importance scores derived from the discourse trees built by each analyst.[2] The correlation was very high: 0.798, $p < 0.0001$. Differences between the two analysts came mainly from their interpretations of two of the texts: the discourse trees of one analyst mirrored the paragraph structure of the texts, while the discourse trees of the other mirrored a logical organization of the text, which that analyst believed to be important.

The Spearman correlation coefficients with respect to the importance of textual units between the discourse trees built by the rhetorical parser and those built by each analyst were 0.480, $p < 0.0001$, and 0.449, $p < 0.0001$. These lower correlation values were due to the differences in the overall shape of the trees and to the fact that the granularity of the discourse trees built by the program was not as fine as that of the trees built by the analysts.

Besides directly comparing the trees built by the program with those built by analysts, I also evaluated the impact that the trees could have on the task of summarizing text. A summarization program that uses the rhetorical parsing algorithm 5.1 recalled 66% of the sentences considered important by 13 judges in the same five texts, with a precision of 68%. In contrast, a random procedure recalled, on average, only 38.4% of the sentences considered important by the judges, with a precision of 38.4%. And the Microsoft Office 97 summarizer recalled 41% of the important sentences with a precision of 39%. In chapter 6, I discuss at length the experiments from which the data presented above was derived.

The rhetorical parser presented here uses only the structural constraints that were enumerated in chapter 2. Co-relational constraints (such as those described by Sumita et

[2]The Spearman rank correlation coefficient is an alternative to the usual correlation coefficient. It 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.

al. [1992]), focus, theme, anaphoric links, and other syntactic, semantic, and pragmatic factors do not yet play a role in the rhetorical parsing algorithm, but I nevertheless expect them to reduce the number of valid discourse trees that can be associated with a text. I also expect that other robust methods for determining coherence relations between textual units, such as those described by Harabagiu and Moldovan [1995, 1996], will improve the accuracy of the routines that hypothesize the rhetorical relations that hold between adjacent units.

5.9 Related work

I am not aware of the existence of any other rhetorical parser for English. I believe that the research that comes closest to that described in this chapter is that of Sumita et al. [1992] and Kurohashi and Nagao [1994].

Sumita et al. [1992] report on a discourse analyzer for Japanese. Even if one ignores some computational "bonuses" that can be easily exploited by a Japanese discourse analyzer (such as co-reference and topic identification), there are still some key differences between Sumita's work and the one presented here. Particularly important is the fact that the theoretical foundations of Sumita et al.'s analyzer do not seem to be able to accommodate the ambiguity of discourse markers; in their system, discourse markers are considered unambiguous with respect to the relations that they signal. In contrast, my rhetorical parser uses a mathematical model in which this ambiguity is acknowledged and appropriately treated. Also, the discourse trees that the rhetorical parser builds are very constrained structures (see chapter 2): as a consequence, the rhetorical parser does not overgenerate invalid trees as Sumita et al.'s does. Furthermore, my rhetorical parser uses only surface-form methods for determining the markers and textual units and uses clause-like units as the minimal units of the discourse trees. In contrast, Sumita et al. use deep syntactic and semantic processing techniques for determining the markers and the textual units and use sentences as minimal units in the discourse structures that they build.

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

A parallel line of research has been recently 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 differrent from that exploited by the rhetorical parser: their algorithms rely primarily on cohesion and not on coherence. Because of this, the referential structures are not as constrained as the discourse structures that the rhetorical parser builds. In fact, the discourse relations between the referential segments are not even labelled. Still, I believe that studying the commonalities and differences between the referential and rhetorical segments could provide new insights into the nature of discourse.

5.10 Summary

The rhetorical parser that I have presented in this chapter takes as input unrestricted English text and generates the valid text structures of that text. The rhetorical parser relies on the following algorithms:

- A surface-form algorithm that determines the elementary units of the text and the cue phrases that have a discourse structuring function.
- An algorithm that uses information that was derived from the corpus analysis discussed in chapter 4 in order to hypothesize exclusively disjunctive rhetorical relations that hold between the textual units of a text.
- An algorithm that uses word co-occurrences in order to hypothesize exclusively disjunctive rhetorical relations that hold between the textual units of a text.
- A chart-parsing algorithm that uses sets of exclusively disjunctive rhetorical relations in order to derive the valid discourse structures of a text.

I have also presented mechanisms that deal with the ambiguity of discourse and discussed two different ways in which discourse trees can be evaluated.

Chapter 6

The summarization of natural language texts

6.1 Preamble

The rhetorical parser presented in chapter 5 not only constructs discourse structures that make explicit the rhetorical relations between different spans of text but also assigns to each node in a discourse tree the elementary units of its promotion set. These units are also shown in the PostScript representations of the discourse trees that are generated by the rhetorical parser. In this chapter 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 6.2, 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 6.4). The experiment confirms that there exists a correlation between the nuclei of a text structure and what readers perceive as being important in the corresponding text.

In section 6.5, I evaluate an implementation of the discourse-based summarization algorithm and discuss its strengths and weaknesses. I end the chapter with a review of related work.

6.2 From discourse structures to text summaries

6.2.1 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 6.1, for example, allows us to determine that, according to the formalization in chapter 2, unit 2 is the most important textual unit in text (6.1) 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 6.1 is the same as the tree in figure 5.20; and text (6.1) is the same as text (5.17).[1] They have been replicated here only for convenience.)

(6.1) [*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}]

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

[1] The only difference between texts (6.1) and (5.17) concerns the labelling of the parenthetical units. In text (6.1), they are labelled with strings having the form Pn, where n denotes the elementary unit to which the parenthetical unit is related. In text (5.17), the parenthetical units were not labelled.

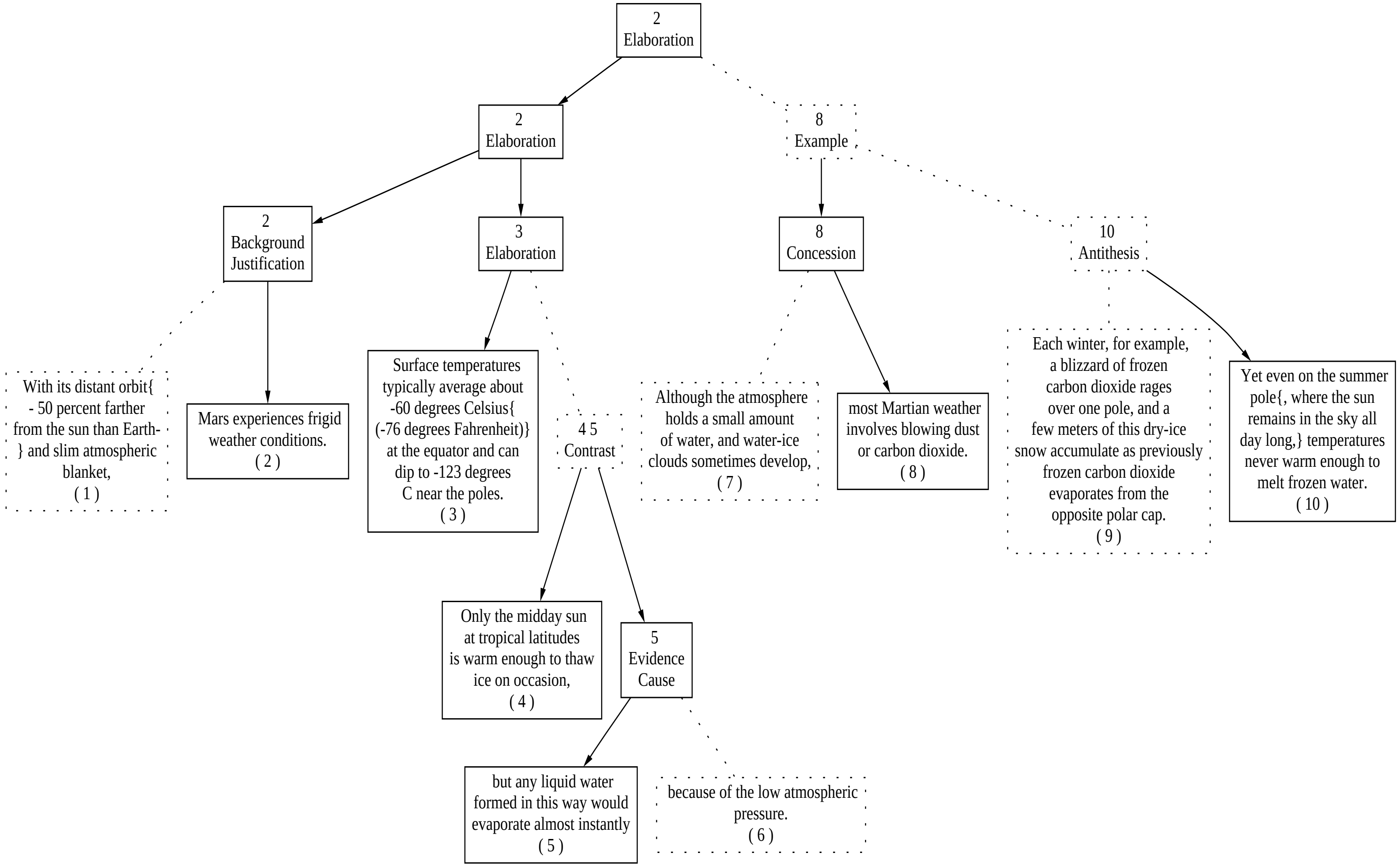

Figure 6.1: The discourse tree of maximal weight that is built by the rhetorical parsing algorithm for text (6.1).

Unit	1	$P1$	2	3	4	5	6	7	8	9	10	$P10$
Score	3	2	6	4	3	3	1	3	5	3	4	2

Table 6.1: The importance scores of the textual units in text (6.1).

promotion unit. The larger the score of a unit, the more important that unit is considered to be in a text. Formula (6.2), 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.

$$(6.2) \qquad score(u, D, d) = \begin{cases} 0 & \text{if } D \text{ is } \textsc{nil}, \\ d & \text{if } u \in promotion(D), \\ d-1 & \text{if } u \in parentheticals(D), \\ max(score(u, leftChild(D), d-1), & \text{otherwise.} \\ \qquad score(u, rightChild(D), d-1)). & \end{cases}$$

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 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 (6.2) to the tree in figure 6.1, which has depth 6, we obtain the scores in table 6.1 for each of the elementary and parenthetical units of text (6.1). 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 (6.2), 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 (6.3), which was induced on the textual units of text (6.1) by the discourse structure in figure 6.1 and formula (6.2).

$$(6.3) \qquad 2 > 8 > 3, 10 > 1, 4, 5, 7, 9 > P1, P10 > 6$$

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.

I. Determine the discourse structure DS of T by means of the rhetorical parsing algorithm in figure 5.1.
II. Determine a partial ordering on the elementary and parenthetical units of DS by means of formula (6.2).
III. Select the first $p\%$ units of the ordering.

Figure 6.2: The discourse-based summarization algorithm

If we are interested in generating a very short summary of text (6.1), 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.

The idea of using discourse structures for constructing text summaries is not new. Researchers in computational linguistics have been 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 (6.2) on the text structures derived by the rhetorical parser is only a precise expression of the original intuition.

6.2.2 A discourse-based summarizer

Given that we can use the rhetorical parser described in chapter 5 to build the discourse structure of any text and that we can use formula (6.2) 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.

The summarization algorithm that I implemented takes two arguments: a text and a number p between 1 and 100 (see figure 6.2). It first uses the rhetorical parsing algorithm in order to determine the discourse structure of the text given as input. It then applies formula (6.2) 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.

6.3 The evaluation of text summaries — general remarks

The evaluation of automatic summarizers has always been a thorny problem: most papers on summarization describe the approach that they use and give some "convincing" samples of the output. In only a very few cases, *the direct output of a summarization program* is compared with a human-made summary or evaluated with the help of human subjects; usually, the results are modest. Unfortunately, evaluating the results of a particular implementation does not enable one to determine what part of the failure is due to the implementation itself and what part to its underlying assumptions.

The position that I take in this thesis is that, in order to build high-quality summarization programs, we need to evaluate not only a representative set of automatically generated outputs (a highly difficult problem by itself), but also the adequacy of the assumptions that these programs use. That way, we are 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.

With few exceptions, automatic approaches to summarization have primarily addressed possible ways to determine the most important parts of a text — much less has been done in finding ways for transforming the selected parts into coherent text (see Paice [1990] for an excellent overview). Determining the salient parts is considered to be achievable because one or more of the following assumptions hold:

- important sentences in a text contain words that are used frequently [Luhn, 1958, Edmundson, 1968];

- important sentences contain words that are used in the title and section headings [Edmundson, 1968];

- important sentences are located at the beginning or end of paragraphs [Baxendale, 1958];

- important sentences are located at positions in a text that are genre dependent — these positions can be determined automatically, through training techniques [Kupiec *et al.*, 1995, Lin and Hovy, 1997, Teufel and Moens, 1997];

- important sentences use *bonus words* such as "greatest" and "significant" or *indicator phrases* such as "the main aim of this paper" and "the purpose of this article", while non-important sentences use *stigma words* such as "hardly" and "impossible" [Edmundson, 1968, Rush *et al.*, 1971, Kupiec *et al.*, 1995, Teufel and Moens, 1997];

- important sentences and concepts are the highest connected entities in elaborate semantic structures [Skorochodko, 1971, Hoey, 1991, Lin, 1995, Barzilay and Elhadad, 1997];

- important and non-important sentences are derivable from a discourse representation of the text [Sparck Jones, 1993b, Ono *et al.*, 1994].

In determining the words that occur most frequently in a text or the sentences that use words that occur in the headings of sections, computers are accurate tools. However, in determining the concepts that are semantically related or the discourse structure of a text, computers are no longer so accurate; rather, they are highly dependent on the coverage of the linguistic resources that they use and the quality of the algorithms that they implement. Although it is plausible that elaborate cohesion- and coherence-based structures can be used effectively in summarization, I believe that when building summarization programs, we should also determine the extent to which these assumptions hold.

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 RS-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.

6.4 From discourse structure to text summaries — an empirical view

6.4.1 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 an attempt to produce an adequate reference set of data, I selected for my experiment five short texts from *Scientific American* that I considered to be well-written. The texts, which are shown in appendix D, ranged in size from 161 to 725 words. The shortest text was the text on Mars that I have used as an example throughout the thesis.

Because my intention was to evaluate the adequacy for summarizing text not only of the program that I implemented but also of the 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 18 textual units, and the longest into 70. Each textual unit was enclosed within square brackets and labelled 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 10 units, as in the case when the discourse-marker and clause-like unit identification algorithm was applied (see text (6.1)), but 18. The text whose minimal units were obtained manually is given in (6.4), below.

(6.4) [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]]

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; I assumed that they had developed adequate comprehension and summarization skills on their own, so no training session was carried out. Table 6.2 presents the scores that were assigned by each judge to the units in text (6.4).

The same texts were also given to two computational linguistics analysts with solid knowledge of Rhetorical Structure Theory. The analysts 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 2 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 (6.2).

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

Table 6.2: The scores assigned by the judges, analysts, and the discourse-based summarizer to the textual units in text (6.4).

Table 6.2 also presents the scores that were derived from the RS-trees that were built by each analyst for text (6.4) 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 6.2 reflects this: all the units that were determined manually and that overlapped an elementary unit determined by the rhetorical parser were assigned the same score. For example, units 1 and 3 in text (6.4) correspond to unit 1 in text (6.1). 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 text (6.4) are assigned the score 3.

6.4.2 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 to possible agreements with the majority opinion. The percent agreements computed for each of the five texts and each level of importance are given in table 6.3. The agreements among judges for my experiment seem to follow the same pattern as those

Text	D.1	D.2	D.3	D.4	D.5	Overall
All units	72.64	73.23	69.23	69.89	70.08	70.67
Very important units	88.46	63.07	64.83	63.73	67.30	65.66
Less important units	51.28	73.07	53.84	46.15	–	58.04
Unimportant units	75.14	82.51	73.07	72.85	71.25	73.86

Table 6.3: Percent agreement with the majority opinion.

described by other researchers in summarization [Johnson, 1970]. That is, the judges are quite consistent with respect to what they perceive as being very important and unimportant, but less consistent with respect to what they perceive as being 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 labelled consistently if a simple majority of the judges (≥ 7) assigned the same score to that unit. Overall, the judges labelled 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 labelled 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 text (6.4), the judges consistently labelled 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 labelling unit 18. If we compute percent agreement figures only for the textual units for which at least 7 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 kinds 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 make 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 that the one computed for the second, I decided to consider the set of 36 textual units labelled 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 text (6.4) belong to this reference set.

6.4.3 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.

The Spearman rank correlation coefficient is an alternative to the usual correlation coefficient. It 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 the Spearman coefficient is that two

Text D.1	Text D.2	Text D.3	Text D.4	Text D.5	Overall
0.645	0.676	0.960	0.772	0.772	0.798

Table 6.4: The Spearman correlation coefficients between the ranks assigned to each textual unit on the basis of the RS-trees built by the two analysts.

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 statistics 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.

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, for text D.1, to 0.960, for text D.3 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.

6.4.4 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 labelled as "very important" by a majority of the judges. For example, for text (6.4), 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 (6.2) 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 6.2). If I had selected units $10, 11$ and 12 as well, I would have ended up selecting 22% of the units in text (6.4), which is farther from 11 than 6. Hence, I determined for each text the set of important units as labelled 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

Text	No. of units that were considered important by judges	First Analyst			
		No. of units that were labelled as important on the basis of the RS-tree built by the analyst	No. of units that were correctly labelled as important on the basis of the RS-tree built by the analyst	Recall	Precision
D.1	2	2	2	100.00	100.00
D.2	9	6	5	55.55	83.33
D.3	7	5	4	57.14	80.00
D.4	12	10	6	50.00	60.00
D.5	6	7	3	50.00	42.85
All	36	30	20	55.55	66.66

Table 6.5: Summarization results obtained by using the text structures built by the first analyst — the clause-like unit case.

RS-trees, with respect to the units labelled important by the judges. 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%, $\sigma = 0.059$. Tables 6.5 and 6.6 show the recall and precision figures for each analyst and each of the five texts.

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 labelled as important textual units 1 to 4 in text (6.4), because they make up a full sentence and because unit 4 was labelled 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. Tables 6.7 and 6.8 show the sentence-related recall and precision figures for each analyst and each of the five texts.

In contrast with the results in tables 6.7 and 6.8, the average recall and precision for the same percentages of units selected randomly 1000 times from the same five texts were 38.4%, $\sigma = 0.048$. 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 as being 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 most of what is needed for selecting accurately the important units in a text. However, the results also suggest that the discourse theory developed in this thesis is not enough by itself if one wants to strive for perfection.

The above results not only provide strong evidence that discourse theories can be used

Text	No. of units that were considered important by judges	Second Analyst			
		No. of units that were labelled as important on the basis of the RS-tree built by the analyst	No. of units that were correctly labelled as important on the basis of the RS-tree built by the analyst	Recall	Precision
D.1	2	1	1	50.00	50.00
D.2	9	8	6	66.66	75.00
D.3	7	5	4	57.14	80.00
D.4	12	7	5	41.66	71.42
D.5	6	9	4	66.66	44.44
All	36	30	20	55.55	66.66

Table 6.6: Summarization results obtained by using the text structures built by the second analyst — the clause-like unit case.

Text	No. of units that were considered important by judges	First Analyst			
		No. of units that were labelled as important on the basis of the RS-tree built by the analyst	No. of units that were correctly labelled as important on the basis of the RS-tree built by the analyst	Recall	Precision
D.1	7	7	7	100.00	100.00
D.2	12	12	12	100.00	100.00
D.3	10	9	8	80.00	88.88
D.4	18	11	8	44.44	72.72
D.5	11	10	5	45.45	50.00
All	58	49	40	68.96	81.63

Table 6.7: Summarization results obtained by using the text structures built by the first analyst — the sentence case.

Text	No. of units that were considered important by judges	Second Analyst			
		No. of units that were labelled as important on the basis of the RS-tree built by the analyst	No. of units that were correctly labelled as important on the basis of the RS-tree built by the analyst	Recall	Precision
D.1	7	7	7	100.00	100.00
D.2	12	11	9	75.00	81.81
D.3	10	9	8	80.00	88.88
D.4	18	11	6	33.33	54.54
D.5	11	13	8	72.72	61.53
All	58	51	38	65.51	74.50

Table 6.8: Summarization results obtained by using the text structures built by the second analyst — the sentence case.

effectively for text summarization, but also suggest strategies that an automatic summarizer might follow. For example, the Spearman correlation coefficient between the judges and the first analyst, the one who did not follow the paragraph structure, was lower than that between the judges and the second analyst. This might suggest that human judges are inclined to use the paragraph breaks as valuable sources of information when they interpret discourse. If the aim of a summarization program is to mimic human behavior, it would then seem adequate for the program to take advantage of the paragraph structure of the texts that it analyzes.

6.5 An evaluation of the discourse-based summarization program

6.5.1 Agreement between the results of the summarization program and the judges with respect to the most important textual units

To evaluate the summarization program, I followed the same method as in section 6.4.4. That is, I used the importance scores assigned by formula (6.2) 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. Tables 6.9 and 6.10 summarize the results.

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 52.77% and the precision was 50%. 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 65.51% and the precision 67.85%. Tables 6.9 and 6.10 show recall and precision results for each of the five

Text	No. of units that were considered important by judges	Discourse-based Summarizer			
		No. of units that were labelled as important on the basis of the tree built by the rhetorical parser	No. of units that were correctly labelled as important on the basis of the tree built by the rhetorical parser	Recall	Precision
D.1	2	2	2	100.00	100.00
D.2	9	8	5	55.55	62.50
D.3	7	8	3	42.85	37.50
D.4	12	14	6	50.00	42.85
D.5	6	6	3	50.00	50.00
All	36	38	19	52.77	50.00

Table 6.9: Summarization results obtained by using the text structures built by the rhetorical parser — the clause-like unit case.

Text	No. of units that were considered important by judges	Discourse-based Summarizer			
		No. of units that were labelled as important on the basis of the tree built by the rhetorical parser	No. of units that were correctly labelled as important on the basis of the tree built by the rhetorical parser	Recall	Precision
D.1	7	7	7	100.00	100.00
D.2	12	14	12	100.00	85.71
D.3	10	9	6	60.00	66.66
D.4	18	20	10	55.55	50.00
D.5	11	6	5	45.45	83.33
All	58	56	38	65.51	67.85

Table 6.10: Summarization results obtained by using the text structures built by the rhetorical parser — the sentence case.

texts that were summarized. The lower recall and precision scores associated with clause-like units seem to be caused primarily by the difference in granularity with respect to the way the texts were broken into subunits: the program does not recover all minimal textual units, and as a consequence, its assignment of importance scores is coarser. When full sentences are considered, the judges and the program work at the same level of granularity, and as a consequence, the summarization results improve significantly.

6.5.2 Comparison of the discourse-based summarizer with the Microsoft Office97 summarization program and a baseline algorithm

I was able to obtain only one other program that summarizes English text — 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%. Tables 6.11 and 6.12 show the recall and precision figures for each of the five texts.

In order to provide a better understanding of the results in this section, I also considered a baseline algorithm that 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. Tables 6.13 and 6.14 show recall and precision results for the baseline, Microsoft Office97, and discourse-based summarizers, as well as the results that would have been obtained if we had applied the score function (6.2) on the discourse trees that were built manually. In tables 6.13 and 6.14, 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 recall and precision results associated with the "Analyst-based Summarizer" in tables 6.13 and 6.14 are averages of the results shown in tables 6.5 and 6.6, and 6.7 and 6.8 respectively.

6.5.3 Discussion

Selecting the most important units in a text

The results presented in this section 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

Text	No. of units considered important by judges	Microsoft Office97 Summarizer			
		No. of units identified	No. of units identified correctly	Recall	Precision
D.1	2	3	1	50.00	33.33
D.2	9	10	5	55.55	50.00
D.3	7	9	3	42.85	33.33
D.4	12	11	1	8.33	9.09
D.5	6	6	0	0.00	0.00
All	36	39	10	27.77	25.64

Table 6.11: Recall and precision figures obtained with the Microsoft Office97 summarizer — the clause-like unit case.

Text	No. of units considered important by judges	Microsoft Office97 Summarizer			
		No. of units identified	No. of units identified correctly	Recall	Precision
D.1	7	8	3	42.85	37.50
D.2	12	12	5	41.66	41.66
D.3	10	11	8	80.00	72.72
D.4	18	20	3	16.66	15.00
D.5	11	11	5	45.45	45.45
All	58	62	24	41.37	38.70

Table 6.12: Recall and precision figures obtained with the Microsoft Office97 summarizer — the sentence case.

Text	Baseline Summarizer	Microsoft Summarizer		Discourse-based Summarizer		Analyst-based Summarizer	
	Recall & Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.
D.1	12.05	50.00	33.33	100.00	100.00	75.00	75.00
D.2	38.01	55.55	50.00	55.55	62.50	61.11	78.57
D.3	36.20	42.85	33.33	42.85	37.50	57.14	57.14
D.4	18.32	8.33	9.09	50.00	42.85	45.83	64.70
D.5	23.06	0.00	0.00	50.00	50.00	58.33	43.75
All	25.7	27.77	25.64	52.77	50.00	55.55	66.66

Table 6.13: Recall and precision figures obtained with the baseline, Microsoft Office97, discourse-based, and analyst-based summarizers — the clause-like unit case.

Text	Baseline Summarizer	Microsoft Summarizer		Discourse-based Summarizer		Analyst-based Summarizer	
	Recall & Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.
D.1	40.12	42.85	37.50	100.00	100.00	100.00	100.00
D.2	50.02	41.66	41.66	100.00	85.71	87.50	91.30
D.3	52.12	80.00	72.72	60.00	66.66	80.00	88.88
D.4	26.91	16.66	15.00	55.55	50.00	38.88	63.63
D.5	42.31	45.45	45.45	45.45	83.83	59.09	56.52
All	38.40	41.37	38.70	65.51	67.85	67.24	78.00

Table 6.14: Recall and precision figures obtained with the baseline, Microsoft Office97, discourse-based, and analyst-based summarizers — the sentence case.

---

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 far from 100%. 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.

A comparison of the discourse-trees built by the analysts and the rhetorical parser reveals some differences. Some of them are caused by the fact that the rhetorical parser makes disjunctive hypotheses about the rhetorical relations that hold between textual units, and sometimes these hypotheses are incorrect. Also, although in some cases the rhetorical parser builds trees that perfectly match the manually built trees, because of its preference for trees that are skewed to the right, it does not select the appropriate ones. This suggests that better heuristics for discourse disambiguation can improve the results. Also, the trees that are built by the rhetorical parser are not as finely grained as those built manually. For example, the rhetorical parser breaks text (6.1) into 10 elementary units; in contrast, the analysts found 18 units for the same text. All these observations suggest that a better rhetorical parser can improve the results of the summarization program.

I turn now to the other possible explanation, the one that concerns the mapping from discourse structures to importance scores. If we examine the results in tables 6.13 and 6.14, we can see that the difference in recall and precision between the discourse-based and analyst-based summarizers is lower than the difference between the analyst-based summarizer and the 100% upper bound. 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, we 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 text D.5 and the units that the judges considered important in the same text. The discourse structure built by the first analyst for text D.5 yielded the lowest recall and precision figures (see table 6.5). Text (6.5), which is given below, replicates text D.5: the elementary units are numbered from 1 to 27 and the units that a majority of the judges agreed to be important are shown in bold.

(6.5) [**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 6.3 shows the discourse structure built by the first analyst. Each elementary unit in the structure is labelled with a number from 1 to 27 as well. The numbers shown in bold that are 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 6.2 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 text (6.5), six units were considered important: those labelled 1, 3, 4, 6, 17, 24. The partial ordering induced by formula (6.2) on the discourse structure of figure 6.3 is that shown in (6.6) below.

(6.6) $$3, 16 > 1, 21, 22, 23, 24 > 2, 4, 5, 6, 17, 18 > 26 > 19, 25, 27 > 8 > 11, 13, 15 > 9, 10 > 12, 14$$

Selecting the first seven units in the partial ordering comes closest to the number of units that were considered important by the judges. As shown in table 6.5, 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.

If we examine the discourse structure of text (6.5) and the units that judges perceived as being important, we notice a couple of very interesting facts. For example, a majority of the judges labelled 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 we examine the text closely, we also 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 (6.2) currently does. By examining the same discourse structure and the importance scores assigned by judges, we 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 (6.2), both pairs of units are assigned the same score. Obviously, mechanisms that are not inherent to

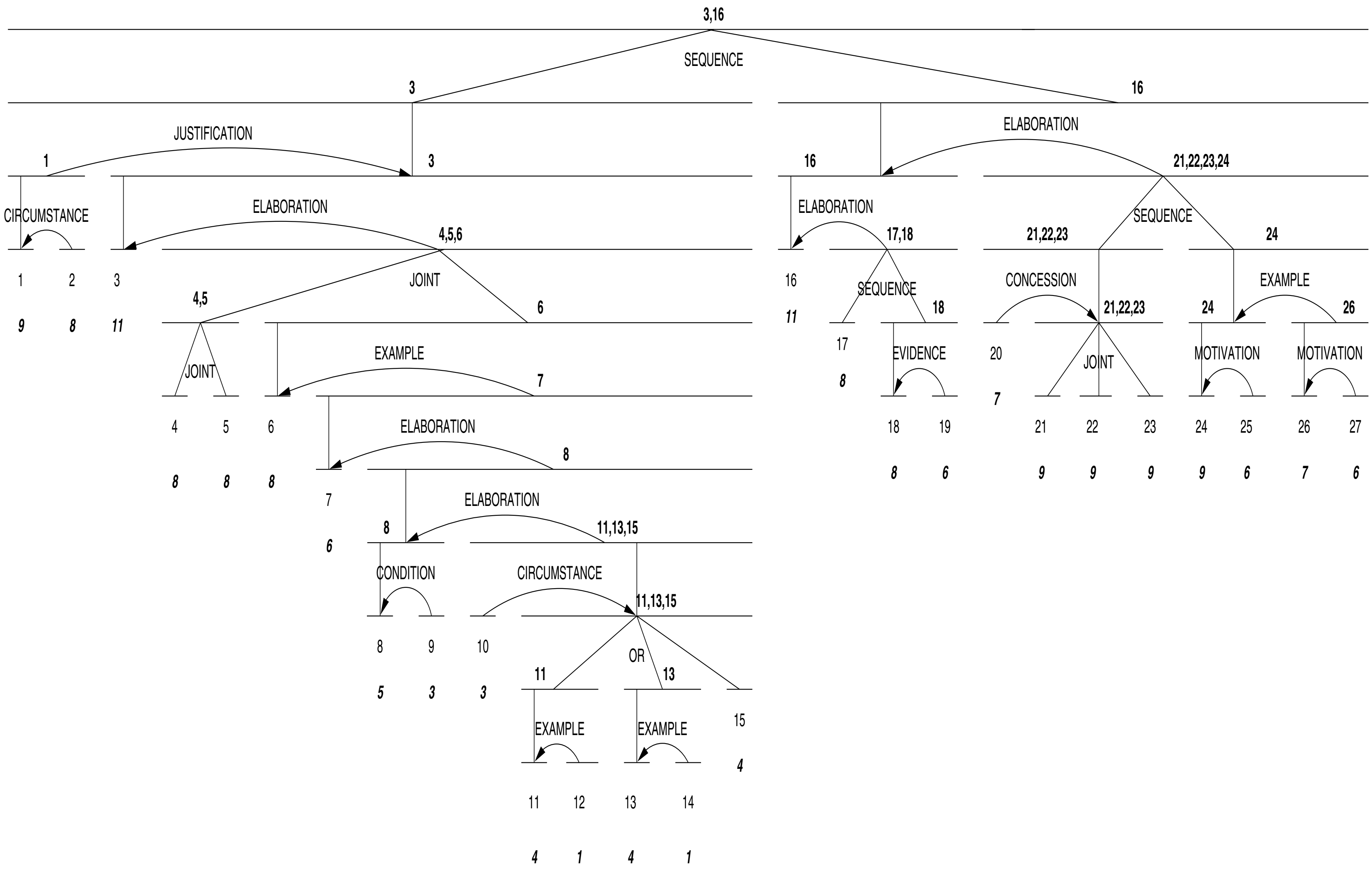

Figure 6.3: The discourse tree that was built for text (6.5) by the first analyst.

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 (6.2) enables a discourse-based summarizer to derive summaries of good quality, there is definitely room for improvement. The experiments described in this chapter 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.

Other issues

Throughout this chapter, we concentrated our attention only on the problem of selecting the most important units in a text. However, this solves only part of the problem, because a complete summarization system will also have to use the selected units in order to produce coherent text. We found that the summaries that are produced by our 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, we have not carried out any readability evaluation. One of the problems that our 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 thesis.

6.6 Related work

6.6.1 Natural language summarization — a psycholinguistic perspective

The empirical experiment described in this chapter confirms the hypothesis that the units that are promoted as important by our text theory correlate with the units considered important by human judges. Given this, it would be interesting to examine how our findings relate to other work in the psycholinguistics of text summarization.

Arguably, the psycholinguistic model of text summarization that has received most attention is that of Kintsch and van Dijk [van Dijk and Kintsch, 1977, Kintsch and van Dijk, 1978, van Dijk, 1980]. This model stipulates that the information to be included in a summary is determined by macrorules (processes of deletion, generalization, and integration) that operate on the propositions of the input text and that incrementally build a macrostructure of that text. Further refinements of Kintsch and van Dijk's model [Garner, 1982, Brown and Day, 1983, Brown *et al.*, 1983] yielded a taxonomy of seven rules that are used consistently by summarizers. Two of the seven rules involve deletion of unnecessary material: material that is trivial and material that is redundant. Two rules concern the substitution of a superordinate term, event, or action for a list of terms or actions. One rule concerns the selection of topic sentences and one rule the invention of topic sentences in the cases in which such sentences are not explicit in the text. The last rule, which has been shown to be used primarily by mature summarizers, concerns the combination of information that was given across paragraphs and the expression of large bodies of texts in a few words. Although Kintsch and van Dijk's model has been criticized as being insufficiently precise in detailing how the macrostructure of text is actually built by readers [Sanford and Garrod, 1981] and as being too specific to narratives [Kintsch, 1982], a number of controlled experiments [Chou Hare and Borchardt, 1984, Sjöstrom and Chou Hare, 1984, Cook and Mayer, 1988] have shown that the teaching of these rules improves the summarization skills of humans.

The fact that the rules proposed by Kintsch, van Dijk, Brown, Day, and others improve the performance of human summarizers suggest that they can be also used in automatic summarization, provided that they can be implemented (see Endres-Niggemeyer [1997] for such a proposal). However, from the perspective of the work described in this chapter, which emphasizes the importance of structure in summarizing text, a different line of research seems to be more relevant.

A set of psycholinguistic experiments 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 signalling. An experiment of Loman and Mayer [1983] has shown that signalling in text increases the recall of conceptual information and helps humans generate high-quality problem solutions. The signalling 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 signalling even across chapters through "preview" and "recall" sentences has a strong effect on readers' recall of prose.

The structure and the explicit signals that pertain to a text can be used to derive general summarization techniques; in fact, our approach relies heavily on that. In some cases, it is, however, useful to exploit the structure of the domain as well. Rumelhart [1972, 1977], for example, has developed a comprehension model of text that is based on readers' application of generic schemata. Rumelhart hypothesized that these schemata help humans not only understand the stories that they read, but also summarize them. An experiment that confirms the role of schemata on text summarization has been carried out by Brooks and Dansereau [1983], who have shown that the teaching of the structural schema of scientific articles improved the recall of important information. A more recent experiment of Dillon [1991], which was carried out in the context of hypertext understanding, has shown that journal readers possess a generic representation of scientific articles that helps them organize isolated pieces of text into a meaningful whole. In fact, it seems that even the abstracts themselves possess an internal structure that can be exploited by means of a schema-based approach [Liddy, 1991].

As we have already seen, the psycholinguistic experiments discussed in this section not only suggest that exploiting the structure of text for the task of automatic summarization bears some cognitive plausibility, but also give hints to further developments that could improve the results that we have obtained so far. Implementing the summarization rules described by Kintsch and van Dijk and using text schemata in specific domains might not be trivial, but might nevertheless lead to better summarization results.

6.6.2 Natural language summarization — a computational perspective

It is very unlikely that in the close future we will be able to support, at a large scale, the development of approaches to natural language summarization that rely heavily on large knowledge resources [Rau *et al.*, 1989, Hahn, 1990]. As a consequence, in this section, I discuss primarily the assumptions and the systems that pertain to the field of domain-independent summarization.

Word-frequency-based systems

The idea that there exists a correlation between, on one hand, the frequency of words and their distribution, and, on the other hand, the significance in texts of the sentences that contain them goes back as far as Luhn [1957, 1958]. In his experiments, Luhn observed that this correlation follows a Bell curve whose minima correspond to words that occur very seldom and very often and whose maximum corresponds to words that occur relatively frequently. The validity of using word-frequency as an indicator of significance has been tested by Edmundson [1968], who showed that it is one of the weakest indicators among a set that also contained title-, position-, and keyword-based indicators: it accounted for only about 36% of the important sentences in a corpus of texts. (A baseline, random indicator recalled 25% of the important sentences in the same collection of documents.) Nevertheless, the word-frequency-based indicator continues to be used even in recent systems [Rau and Brandow, 1993, Manesh, 1997, Leong *et al.*, 1997], most often in connection with other indicators.

Title-based systems

Another assumption that is used frequently in implemented summarization systems is that the words of the title and headings correlate with what is important in texts. Edmundson [1968] showed that the hypothesis that words of the title and heading are positively relevant is statistically valid at the 99 percent level of significance. However, it is able to recall only about 41% of the important sentences in a collection of documents. As in the case of the word-frequency-based method, the title-based method continues to be used in recent systems, such as those described by Preston and Williams [1994], Manesh [1997], and Ochitani et al. [1997].

Position-based systems

An initial experiment of Baxendale [1958] showed that in 85% of 200 individual paragraphs the topic sentence occurred in initial position and in 7% in final position. Although this observation suggests that position may correlate to a high degree with sentence significance, it does not specify how the position indicator scales up to large texts. Edmundson [1968] has shown that the position-based indicator could account for up to 53% of the important sentences in a text. A much more careful study by Lin and Hovy [1997] showed that position of important sentences in a text is genre dependent and that one can derive a partial ordering with respect to their importance by means of training. For newspaper articles announcing computer products, Lin and Hovy have shown that the title of an article is most likely to contain significant topics, followed by the first sentence of the second paragraph, the first sentence of the third paragraph, etc. In contrast, for the *Wall Street Journal*, the order is: the title, the first sentence in the first paragraph, the second sentence in the first paragraph, etc. The position indicator is applied in connection with other indicators in other systems as well, such as those described by Kupiec et al. [1995], Manesh [1997], Teufel and Moens [1997], and Jang and Myaeng [1997].

Keyword-based systems

We have already mentioned that keyword-based systems rely on the assumption that important sentences in a text 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*. Experiments carried out by Edmundson [1968], Rush et al. [1971], Paice [1981], Kupiec et al. [1995], and Teufel and Moens [1997] have repeatedly confirmed that keywords constitute a good indicator of importance, the recall of important sentences being in the range of 40 to 50%. In fact, the keyword method is also used in combination with other methods in systems such as those described by Manesh [1997], Aone et al. [1997], Lehmam [1997], and Jang and Myaeng [1997]. The main characteristic of these systems is that they all use predefined sets of cue phrases.

A similar, but somewhat different line of research is explored by Schwarz [1990], Boguraev and Kennedy [1997], and Szpakowicz et al. [1997], who assume that important sentences are those that contain keyphrases, i.e, noun phrases that are usually generated by term-index identification algorithms. Term identification algorithms, such as that described by Justeson and Katz [1995], usually produce an unordered set of terms. Important sentences are considered to be those that contain these terms [Szpakowicz *et al.*, 1997]. In a more sophisticated approach, Boguraev and Kennedy [1997] use a set of rules that pertain to the linguistic context in which the terms occur in order to assign an importance score to each of them: those with maximal score are considered to be the most salient ones in a text. If desired, one can then build a summary from the sentences that contain the most salient phrases.

Information-extraction-based systems

Information-extraction-based summarization systems [DeJong, 1982, Paice and Jones, 1993, Riloff, 1993, Liddy, 1993, McKeown and Radev, 1995, Gaizauskas and Robertson, 1997] are usually used to generate abstracts that concern very specific aspects, such as the when, who, what, why, etc., of some events. The assumption that they rely upon is that the extraction systems that they use as front-ends are robust and that they select adequately the required information. SUMMONS [McKeown and Radev, 1995], the most sophisticated system in this category, is, in fact, the only system that thoroughly addresses the issue of generating summaries that are not only informative but also coherent and cohesive. A collection of plan operators and templates, which informs much work in natural language generation, is used to combine frames of information that are extracted from a set of documents. The frames are eventually mapped into English using FUF [Elhadad, 1991], a functional linguistic surface generator.

Cohesion-based systems

Another assumption on which summarization systems rely upon is that important words, sentences, and paragraphs are the highest connected entities in elaborate graph-like representations of text. The earliest account of an approach that uses the idea of cohesion is that of Skorochodko [1971]. Given a text, Skorochodko shows how one can associate with it a weighted graph whose nodes are given by individual sentences; weighted links between nodes reflect the semantic overlap between the words of the corresponding sentences. On the basis of the graph, Skorochodko shows how an importance score can be associated with each sentence, a score that depends on the number of arcs that are incident to the node of the sentence under consideration, the total number of nodes in the graph, and the number of sentences in the longest connected fragment of text formed after the removal of the given sentence. Skorochodko's idea was also investigated by Hoey [1991] and implemented by Preston and Williams [1994].

A simple form of cohesion, i.e., term repetition, was exploited by Salton et al. [1995], Salton and Allan [1995], and Salton and Singhal [1996], who applied traditional information retrieval techniques in order to associate with a text a weighted graph whose nodes are given by paragraphs and whose weighted arcs are given by a cosine measure of similarity between the corresponding paragraphs. Subsequent experiments [Mitra *et al.*, 1997] have shown that the degree of overlap between the paragraphs considered important by Salton et al.'s algorithms and the paragraphs considered important by humans is significantly higher than the overlap between the paragraphs considered important by Salton et al.'s algorithms and a set of randomly extracted paragraphs. However, the overlap between the paragraphs considered important by Salton et al.'s algorithms and the paragraphs considered important by humans was lower than the overlap between the paragraphs considered important by Salton et al.'s algorithms and the lead paragraphs.

Another cohesion-based approach to text summarization is that proposed by Barzilay and Elhadad [1997], who explore the use of lexical chains. Lexical chains, as defined by Morris and Hirst [1988, 1991], are sequences of semantically related words, that can be automatically derived using a thesaurus [Morris, 1988] or WordNet [St-Onge, 1995, Hirst and St-Onge, 1997]. Barzilay and Elhadad assign a strength to each chain on the basis of its length and number of elements. They use then various heuristics in order to derive from the chain scores an importance assignment to each sentence in a text.

The relationship between words constitutes the foundation of Mani and Bloedorn [1997a, 1997b] approach as well. The algorithm that they propose first builds a graph for each text, whose nodes are given by words, phrases, and proper names, and whose arcs are both semantic in nature, i.e., they denote relations of synonymy, coreference, etc., and location-based, i.e., they denote adjacency. Using the cohesion graph, a vector of word weights is associated with each document, in the style used by information retrieval systems. Mani and Bloedorn's system also takes as input a topic that is used by a spreading-activation algorithm in order to re-weight the vectors of each document such that words that are "close" to the topic receive higher values. A set of backend algorithms then determine segment boundaries and select the important sentences in a text. An evaluation procedure has shown that the summaries generated in this way can reduce by 20% the time spent by users on a retrieval task.

All the cohesion-based approaches described so far take sentences, paragraphs, and text segments as elementary units. In contrast, the approach described by Lin [1995] and Hovy and Lin [1997] takes concepts as being elementary units and explore the possibility of determining automatically the concepts that subsume those determined important in a text. Determining the subsumers will let one replace, for example, the list *wheel, chain, pedal, saddle, light, frame*, and *handlebars* with *bicycle*, by exploiting a set of part-whole relations defined in WordNet. A similar notion of condensation is explored at the formal level by Reimer and Hahn [1997] in the context of textual information represented in terminological knowledge bases. To a certain extent, even Boguraev and Kennedy's [1997] algorithm for determining the most salient keyphrases in a text can be interpreted as a syntax-based condensation method.

Discourse-based systems

The assumption made by discourse-based summarization systems is that the high-level structure of discourse can be used to determine the most important entities and sentences of a text. Two 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 entities that a text "is about" would be given by the entities 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.

Other issues

In presenting the relevant work in the field, I have chosen a strategy similar to that of Paice [1990], i.e., I reviewed the literature from the perspective of the assumptions that various approaches rely upon. However, as I specified repeatedly, most summarization systems use a combination of methods for determining the most important units in a text. Some of these systems combine the importance scores predicted by various methods using manually crafted heuristics [Edmundson, 1968, Lehmam, 1997], while others rely on various training techniques in order to determine the best way in which the various predictions can be combined [Kupiec *et al.*, 1995, Teufel and Moens, 1997, Jang and Myaeng, 1997]. An approach that takes to an extreme the idea that adequate summarization can be achieved only when a variety of features that range from surface-based to pragmatic-based are accounted for is proposed by Aretoulaki [1996, 1997]. Aretoulaki envisions that one can go from natural language text to fully coherent summaries (this accounts for the process of rewriting the selected important textual units as well) by using both symbolic and connectionist techniques. A collection of morphological, syntactic, semantic, and pragmatic analyzers are supposed to map the input text into surface and rhetorical features. A cascade of neural networks is then supposed to map these features into pragmatic features, which pertain to the goals and plans of the writer and the rhetorical means by which these plans and goals are achieved. One of Aretoulaki's main contributions comes from the experimental side of her work, which suggests that the use of pragmatic features instead of surface features improves the recall and precision of the process that identifies the sentences that are important in a text.

6.7 Summary

In this chapter, I first discussed the importance of evaluating not only the outputs of the summarization programs that we build, but also the adequacy of the assumptions that these programs rely upon; and I claimed that this enables us to distinguish the problems that pertain to a particular implementation from those that pertain to the underlying theoretical framework. To support this claim, I designed an experiment that showed that the theoretical concept of discourse structure can be used effectively for summarizing text. The experiment suggested that discourse-based methods and a simple mapping from discourse trees to importance scores can account for determining the most important units in a text with a recall and precision as high as 70%.

I also showed how the concepts of rhetorical analysis and nuclearity can be treated algorithmically and I compared recall and precision figures of a summarization program that implements these concepts with recall and precision figures that pertain to a baseline algorithm and to a commercial system, the Microsoft Office97 summarizer. The discourse-based summarization program that I propose significantly outperforms both the baseline and the commercial summarizer.

By comparing the recall and precision figures that characterized the important sentences derived from the discourse structures that were built by human analysts and the discourse-based summarizer, I identified and discussed two possible sources of improvement. The first concerns the quality of the discourse structures that are derived by the rhetorical parsing algorithm. The second concerns the mapping between these structures and the importance scores that are assigned to textual units.

Chapter 7

From local to global coherence: A bottom-up approach to text planning

7.1 Motivation

Traditionally, the generation of natural language texts has been modeled as a pipeline of independent processes that assumes a generic architecture similar to that shown in figure 7.1. From this perspective, a natural language generation (NLG) system is supposed to support the following processes, or modules:

Content determination delineates from a given knowledge base the information that is relevant to a certain query or topic;

Content organization determines the way in which this relevant information is structured. The structuring can be done at different levels of refinement:

- **Text planning** pertains to partitioning relevant information into units that consist of similar concepts clustered around an organization focus;
- **Paragraph planning** aims at structuring and ordering text units into clause-like segments so that the outcome is coherent;

Sentence planning aspires at rendering the information encoded in text plans into a linguistically motivated representation; this includes mapping text plans into grammatical relations, generating referring expressions for individual entities, and employing ordering constraints with respect to clauses and sentences;

Realization and Lexical Choice map sentence plans into text and choose the lexical items that are appropriate for conveying the message that is encoded by the sentence

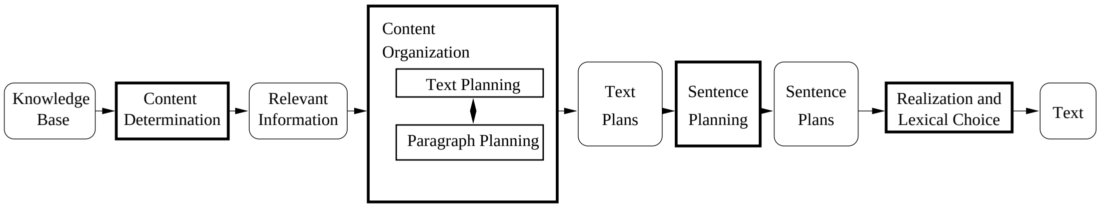

Figure 7.1: Traditional pipeline architecture of an NLG system. Boxes with heavy lines represent processes; boxes with light lines and rounded corners represent intermediate representations that refine a formal representation into a natural text.

---

plans.

All current flexible approaches to text and paragraph planning that assume that the abstract structure of text is a tree-like structure are, essentially, top-down approaches. Some of them define plan operators and exploit hierarchical planning techniques [Hovy, 1993, Moore and Paris, 1993, Moore and Swartout, 1991, Cawsey, 1991, Maybury, 1992] and partial-order planning techniques [Young and Moore, 1994]. Others assume that plans are hierarchically organized sets of frames that can be derived through a top-down expansion process [Nirenburg *et al.*, 1989, Meteer, 1992]. And the recursive application of schemata [McKeown, 1985] can be thought of as a top-down expansion process as well.[1]

One of the major strengths of all these approaches is that, given a high-level communicative goal, they can interleave the task of content organization and content selection, and produce different texts for different knowledge bases and users [McKeown, 1985, Paris, 1991, McCoy and Cheng, 1991, Moore and Swartout, 1991]. Unfortunately, this strength is also a major weakness, because top-down and schema-based approaches are inadequate when the high-level communicative goal boils down to "tell everything that is in this knowledge base" or "tell everything that is in this chosen subset". The reason for this inadequacy is that these approaches cannot ensure that *all* the knowledge that makes up a knowledge pool will be eventually mapped into the leaves of the resulting text plan; after building a partial text plan, which encodes a certain amount of the information found in the initial knowledge pool, it is highly likely that the information that is still unrealized will satisfy none of the active communicative goals. In fact, because the plan construction is plan-operator- or schema-step-driven, top-down approaches cannot even predict what amount of the initial knowledge pool will be mapped into text when a certain communicative goal is chosen. The only way to find a text plan that is maximal with respect to the amount of knowledge that is mapped into text is to enumerate all possible high-level communicative goals and all plans that can be built starting from them, but this is unreasonable.

[1] Section 7.8 discusses in more details the specifics of these approaches.

Given that most NLG systems employ a pipeline architecture in which content determination and content organization are treated as separate processes [Reiter, 1994], I believe that it is critical to provide a flexible solution to the problem of mapping a full knowledge base (or any of its chosen subsets) into text. Previous research in text planning[2] has addressed this issue only for text genres in which the ordering of sentences is very rigid (geographical descriptions [Carbonell and Collins, 1973], stories [Schank and Abelson, 1977], and fables [Meehan, 1977]), has assumed that text plans can be assimilated with linear sequences of textual units [Mann and Moore, 1981, Zukerman and McConachy, 1993], or has employed very restricted sets of rhetorical relations [Zukerman and McConachy, 1993]. Unfortunately, the linear structure of text plans is not sophisticated enough for managing satisfactorily a whole collection of linguistic phenomena such as focus, reference, and intentions, which are characterized adequately by tree-like text plans [Hovy, 1993, Moore and Paris, 1993, Moore and Swartout, 1991, Cawsey, 1991, Paris, 1991, McCoy and Cheng, 1991].

In this chapter, I provide a *bottom-up, data-driven solution for the text planning problem* that relies on the mathematical model of text structures that was proposed in chapter 2. The algorithms that I propose here not only map a knowledge pool into text plans whose leaves subsume all the information given in the knowledge pool, but can also ensure that the resulting plans satisfy multiple high-level communicative goals.

7.2 Foundations of the bottom-up approach to text planning

7.2.1 Introduction

Let us assume that we are given the task of constructing a text plan whose leaves subsume all the information given in a knowledge base (KB). For simplicity, I assume that the KB is represented as a set of semantic units $U = \{u_1, u_2, \dots, u_n\}$. I also assume that rhetorical relations of the kind used throughout this thesis might hold between pairs of semantic units in U. These rhetorical relations can be derived from the KB structure, from the definitions in a library of plan operators, or can be given as input by the creator of the KB. For example, if the semantic units are stored in a description-logic-based KB such as LOOM [MacGregor and Bates, 1987] or CLASSIC [Patel-Schneider *et al.*, 1991, Brachman, 1992], one can derive some rhetorical relations by inspecting the types of links and paths between every pair of semantic units. When the KB consists of a set of frames with clearly defined semantics, such as those produced by systems developed for information extraction tasks [McKeown and Radev, 1995], one can use the underlying semantics of frames to derive rhetorical relations between the information encoded in different slots. For less-structured KBs, one can use the libraries of plan operators that were developed by researchers in hierarchical planning [Hovy, 1993,

[2]In the rest of this thesis, I will adopt the traditional jargon and refer to the task of content organization as "text planning".

Moore and Paris, 1993, Meteer, 1992, Moore, 1995] and derive the set of rhetorical relations that hold between every pair of semantic units. For very rich KBs, such as that used in the HealthDoc Project [Wanner and Hovy, 1996, Hovy and Wanner, 1996, DiMarco and Foster, 1997, DiMarco *et al.*, 1997, Hirst *et al.*, 1997], one can simply extract these relations directly, because they are explicitly represented.

Each of the alternatives described above has been already discussed in the literature to a greater or lesser extent. Therefore, for the purpose of this thesis, I will simply assume that the input for a text planner is a set U of semantic units and the set R_U of rhetorical relations that hold between every pair of units in U. Note that there are no constraints on the number of rhetorical relations that may hold between two semantic units: on one hand, when two units are not related, no rhetorical relation holds between them at all; on the other hand, depending on the communicative goal that one wants to emphasize, more than one relation may hold between two units [Mann and Thompson, 1988, Moore and Pollack, 1992]. In the latter case, I assume that R_U lists all possible relations.

For example, the KB in (7.1) contains four semantic units among which five rhetorical relations hold (7.2).

$$(7.1)\quad U_1 = \begin{cases} \text{A}_1 = \text{“Insulin-dependent diabetes is the less common type of diabetes.”} \\ \text{B}_1 = \text{“The pancreas, a gland found behind the stomach, normally} \\ \qquad\quad \text{makes insulin.”} \\ \text{C}_1 = \text{“With insulin-dependent diabetes, your body makes little or no} \\ \qquad\quad \text{insulin.”} \\ \text{D}_1 = \text{“The condition that you have is insulin-dependent diabetes.”} \end{cases}$$

$$(7.2)\qquad R_{U_1} = \begin{cases} rhet\_rel(\text{ELABORATION}, \text{A}_1, \text{D}_1) \\ rhet\_rel(\text{ANTITHESIS}, \text{A}_1, \text{D}_1) \\ rhet\_rel(\text{ELABORATION}, \text{C}_1, \text{D}_1) \\ rhet\_rel(\text{JUSTIFICATION}, \text{C}_1, \text{D}_1) \\ rhet\_rel(\text{ELABORATION}, \text{B}_1, \text{C}_1) \end{cases}$$

The KB in (7.3) contains three semantic units among which five rhetorical relations hold (7.4).

$$(7.3)\qquad U_2 = \begin{cases} \text{A}_2 = \text{“We can go to the bookstore.”} \\ \text{B}_2 = \text{“We can go to Sam's bookstore.”} \\ \text{C}_2 = \text{“You come home early.”} \end{cases}$$

```
'(asc / ascription
    :tense present
    :domain (cond / abstraction
             :lex condition
             :determiner the
             :process (have / ownership
                        :lex have-possession
                        :tense present
                        :domain (hearer / person)
                        :range cond))
    :range (diab / abstraction
            :lex diabetes
            :determiner zero
            :property-ascription (ins / quality
                                   :lex insulin-dependent)))
```

Figure 7.2: A Sentence Plan Language (SPL) representation of textual unit D_1 in (7.1), "The condition that you have is insulin-dependent diabetes".

(7.4) $$R_{U_2} = \begin{cases} rhet\_rel(\text{ELABORATION}, \text{B}_2, \text{A}_2) \\ rhet\_rel(\text{CONDITION}, \text{C}_2, \text{A}_2) \\ rhet\_rel(\text{CONDITION}, \text{C}_2, \text{B}_2) \\ rhet\_rel(\text{MOTIVATION}, \text{A}_2, \text{C}_2) \\ rhet\_rel(\text{MOTIVATION}, \text{B}_2, \text{C}_2) \end{cases}$$

To increase readability, the semantic units are given in textual form, but one should understand that a chosen formal language is actually used. For example, in HealthDoc, units are represented using the language of sentence plans (SPL), which was developed within the Penman group [Penman Project, 1989, Kasper, 1989] (see figure 7.2). As in the rest of the thesis, the rhetorical relations are represented as first-order predicates whose first argument denotes the name of the rhetorical relation, and whose second and third arguments denote the satellite and the nucleus that pertain to that relation.

In this chapter, I show how one can derive text plans from inputs of the kind shown in (7.1)–(7.2) and (7.3)–(7.4).

7.2.2 Key concepts

The foundations of the bottom-up approach to text planning that I will describe rely on an under-exploited part of Mann and Thompson's Rhetorical Structure Theory [1988] and

Satellite before Nucleus		
	Antithesis	Conditional
	Background	Justify
	Concessive	Solutionhood
Nucleus before Satellite		
	Elaboration	Purpose
	Enablement	Restatement
	Evidence	

Figure 7.3: Canonical orders of text spans for rhetorical relations [Mann and Thompson, 1988, p. 256]

on the formalization of text structures discussed in chapter 2. During the development of RST, Mann and Thompson noticed that rhetorical relations exhibit strong patterns of ordering of their nuclei and satellites, which they called *canonical orderings* (see figure 7.3). The key idea of the bottom-up approach to text planning is to formalize both the strong tendency of semantic units 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, the 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.

7.3 The strengths of the local constraints that characterize coherent texts

The canonical orderings listed by Mann and Thompson (see figure 7.3) do not cover all rhetorical relations and do not provide clear-cut evidence about how "strong" the ordering preferences are. Fortunately, the corpus study discussed in chapter 4 provides empirical data for determining both the ordering preferences of the nucleus and satellite of a much larger set of rhetorical relations and the "strength" of these preferences. The corpus analysis also provides data for determining the strength of the tendency of rhetorically related units to cluster (in some cases, the nucleus and the satellite need not be adjacent).

Using the relational database that encodes the results of the corpus analysis, I computed, for each rhetorical relation, four data, which is explained below: the strength of the preference for the nucleus to precede the satellite, s_o; the normalized average number of sentences that separate the nucleus and satellite, avg_s; the average number of clause-like units that separate the nucleus and satellite, avg_c; the strength of the clustering preference, s_c, which reflects the inclination of rhetorically related units to be realized as adjacent clauses. Table 7.1 presents part of the statistical data that I derived for the rhetorical relations that I use in the examples given in this chapter. Appendix E presents the statistical data for each rhetorical relation that was used in the corpus.

The strength of the ordering of a relation R, $s_o(\text{R})$, is a number between 0 and 1 that reflects the percentage of cases in which the nucleus of a relation was realized before the satellite in the examples found in the corpus. For example, the strengths of the ordering preferences in table 7.1 show that 97% of the ELABORATIONS and 36% of the CONCESSIONS in the corpus realize the nucleus before the satellite. The closer the value is to 1, the more likely it is that rhetorical relation R realizes its nucleus before the satellite. The closer the value is to 0, the more likely it is that rhetorical relation R realizes its nucleus after the satellite.

The second column in table 7.1 represents the normalized average number of sentences that separate the nucleus and satellite of each rhetorical relation in the corpus. The normalized average $avg_s(\text{R})$ is computed using formula (7.5), which is given below.

$$avg_s(\text{R}) = \sum_{\text{R} \in Corpus} \frac{Sentence\_distance_{\text{R}} + 1}{count(\text{R})} \quad (7.5)$$

In formula (7.5), $Sentence\_distance_{\text{R}}$ reflects the content of the field of the same name in the database and $count(\text{R})$ represents the number of examples in the corpus that were labelled with relation R. Since $Sentence\_distance_{\text{R}}$ takes values that are greater than or equal to -1, we add 1 to each value in order to obtain a normalized average that is greater than or equal to 0. The closer the average is to 0, the more likely it is that rhetorical relation R realizes its nucleus and satellite as adjacent clauses within the same sentence. The larger the average is, the more likely it is that rhetorical relation R realizes its nucleus and satellite as sentences that are not even adjacent.

The third column in table 7.1 represents the average number of clause-like units that separate the nucleus and satellite of each rhetorical relation in the corpus. The average $avg_c(\text{R})$ is computed using formula (7.6) below.

$$avg_c(\text{R}) = \sum_{\text{R} \in Corpus} \frac{Clause\_distance_{\text{R}} + Distance\_to\_salient\_unit_{\text{R}} + 1}{count(\text{R})} \quad (7.6)$$

In formula (7.6), $Clause\_distance_{\text{R}}$ and $Distance\_to\_salient\_unit_{\text{R}}$ reflect the content of the fields of the same names in the database. Since $Distance\_to\_salient\_unit_{\text{R}}$ takes values that are greater than or equal to -1, we add 1 to each value in order to obtain a figure that is greater than or equal to 0. The closer the average is to 0, the more likely it is that rhetorical relation R realizes its nucleus and satellite as adjacent clauses.

Since the textual units of interest are clause-like units, the strength of the clustering preference of a relation R, $s_c(\text{R})$, is computed on the basis of the average clause distance between the nucleus and satellite of a rhetorical relation by taking the complement with respect to 1 of the average clause distance. In the cases in which the complement yields a

Rhetorical relation	Strength of the ordering preference (nucleus first) s_o	Average sentence distance between nucleus and satellite avg_s	Average clause distance between nucleus and satellite avg_c	Strength of the clustering preference s_c
ELABORATION	0.97	1.08	0.90	0.10
CONCESSION	0.36	0.11	0.08	0.92
JUSTIFICATION	0.15	0.82	0.53	0.47
CONDITION	0.41	0.07	0.02	0.98
MOTIVATION	0.73	0.64	0.36	0.64

Table 7.1: Ordering and adjacency preferences for a set of rhetorical relations.

negative value, which happens for a few outliers, we assign to the clustering preference the value 0.05 (see formula (7.7)).

(7.7) $$s_c(\mathrm{R}) = \begin{cases} 1 - avg_c(\mathrm{R}) & \text{if } 1 - avg_c(\mathrm{R}) > 0, \\ 0.05 & \text{otherwise.} \end{cases}$$

Values of $s_c(\mathrm{R})$ that are close to 0 reflect no preference for clustering. Values close to 1 reflect a preference for clustering into units that are adjacent. For example, the strengths of the clustering preferences that pertain to CONCESSION and CONDITION reflect a strong tendency of textual units that are related through these relations to be realized as adjacent units. In contrast, the clustering preference associated with the relation of ELABORATION shows a weaker tendency of textual units that are related though this relation to be realized as adjacent units.

The results of the corpus analysis provide strong indications about ways to achieve local coherence. Using the data in table 7.1, one can determine, for example, that if an NLG system is to produce a text that consists of two semantic units for which a CONCESSION relation holds, then it would be appropriate to aggregate the two units into only one sentence and to realize the satellite first. In the case that an ELABORATION relation holds between the two semantic units, it is appropriate to realize the units as two different sentences, with the nucleus being presented first.

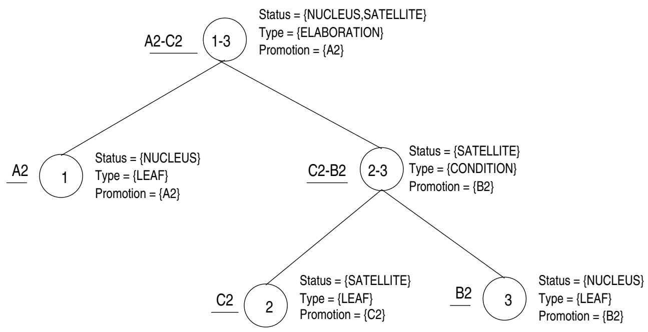

Figure 7.4: Example of a text plan in which units A_2, B_2 are tree-adjacent but not linear-adjacent.

7.4 From local to global coherence

7.4.1 Preamble

One way to formalize these local coherence preferences is as weighted constraints on ordering and adjacency. If one uses this approach, then coherent texts will be those that are characterized by valid text plans that satisfy "most" of the ordering and adjacency constraints. Before fleshing out the mathematics of "most", I believe that it is worthwhile to draw the reader's attention to the fact that a proper treatment of adjacency constraints is not straightforward because the corpus analysis provides data that pertains to a linear structure (the sequence of textual units), whereas text plans are tree-like structures. The position taken here is that a proper treatment of adjacency constraints is one that takes seriously the nuclearity properties that characterize valid discourse trees. When nuclearity is accounted for, two semantic units are considered *tree-adjacent* if they are arguments of a rhetorical relation that connects two subtrees and if the arguments are salient units in those trees. For example, if a certain claim is followed by two evidence units that are connected through a JOINT relation, it is appropriate to assume that both evidence units are tree-adjacent to the claim. Two semantic units are considered *linear-adjacent* if they are adjacent in the text that results from an in-order traversal of the discourse tree. In the text plan shown in figure 7.4, which is a valid text plan for problem (7.3)–(7.4), units A_2, B_2 are tree-adjacent but not linear-adjacent.

In order to provide a mathematical grounding for "most", I associate to each valid discourse tree T, a weight function $w(T)$. The weight of a tree is defined as the sum of the *intrinsic weight*, $w_i(T)$ and the *extrinsic weight*, $w_e(T)$.

$$w(T) = w_i(T) + w_e(T) \tag{7.8}$$

The intrinsic weight is given by a linear combination of the weights of the ordering constraints ($w_{order}(\text{R}, T)$), tree-adjacency constraints ($w_{tree\_adj}(\text{R}, T)$), and linear-adjacency constraints ($w_{lin\_adj}(\text{R}, T)$) that are satisfied by each rhetorical relation R in the discourse structure T that is built (7.9).

$$(7.9) \qquad w_i(T) = \sum_{\text{R} \in T} (w_{order}(\text{R}, T) + 0.5 w_{tree\_adj}(\text{R}, T) + 0.5 w_{lin\_adj}(\text{R}, T))$$

The coefficients in (7.9) reflect the intuition that ordering and clustering are equally important for achieving coherence; nevertheless, extensive experiments could yield different coefficient values. To date, I have not carried out such experiments.

For every relation $\text{R} \in T$ the weights $w_{order}(\text{R}, T)$, $w_{tree\_adj}(\text{R}, T)$, and $w_{lin\_adj}(\text{R}, T)$ are defined as shown in (7.10), (7.11), and (7.12) respectively.

$$(7.10) \qquad w_{order}(\text{R}, T) = \begin{cases} s_o(\text{R}) & \text{if the nucleus of R goes before the satellite,} \\ 1 - s_o(\text{R}) & \text{otherwise.} \end{cases}$$

$$(7.11) \qquad w_{tree\_adj}(\text{R}, T) = s_c(\text{R})$$

$$(7.12) \qquad w_{lin\_adj}(\text{R}, T) = \begin{cases} s_c(\text{R}) & \text{if the nucleus of the satellite of R are adjacent} \\ & \quad \text{in an in-order traversal of the leaves of } T, \\ 0 & \text{otherwise.} \end{cases}$$

If the nucleus of a relation goes before the satellite, then the value of the ordering weight is given by the strength $s_o(\text{R})$ derived from the corpus. If the nucleus goes after the satellite, the value of the ordering weight is given by the complement of $s_o(\text{R})$. Since the rhetorical relation R is used in the tree, it follows that its arguments are tree adjacent. Hence, the value of the tree adjacency weight is given by the strength of the clustering tendency that was derived from the corpus. In the case where the arguments of a rhetorical relation are linear adjacent, the value of the linear adjacency weight is given by the strength of the clustering tendency. If the units are not adjacent, the value is 0.

Since the input to the text-planning problem contains all possible relations between the semantic units given as input, it is likely that the final discourse tree will not use all these relations. However, despite the fact that some relations do not have a direct contribution to the tree that is built, some of their ordering and adjacency constraints may nevertheless be satisfied. I assume that discourse plans that satisfy ordering and adjacency constraints that are not explicitly used in the plans are "better" than those that do not, because the former may enable the reader to derive more inferences.

For a better understanding of this concept, assume, for example, that we are supposed to build a text plan for two units, A and B, between which two rhetorical relations hold: $rhet\_rel(\text{R1}, \text{A}, \text{B})$ and $rhet\_rel(\text{R2}, \text{A}, \text{B})$. Assume that R1 and R2 have the same clustering

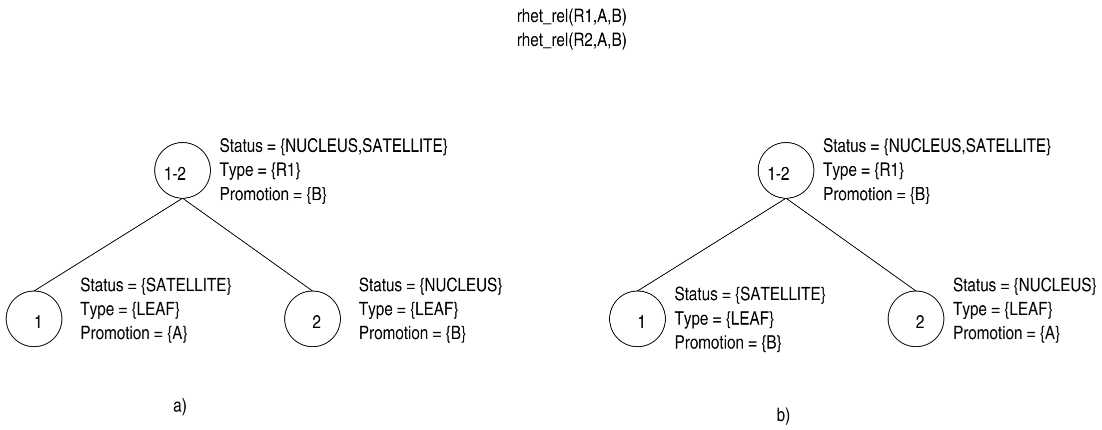

Figure 7.5: Extrinsic and intrinsic weights: an example.

preference and that the ordering preference of R1 is 0.5, while the order preference of R2 is 0.8. That is, relation R1 has no preference for realizing the nucleus or satellite first, but relation R2 has a strong preference for realizing the nucleus first.

Assume now that we use relation R1 to construct a text plan. If we consider only the intrinsic weight of text plans, we have no way to choose between the two solutions of this problem, which correspond to the two different orderings of the units. The text plans associated with these orderings are shown in figure 7.5. Both trees in figure 7.5 have the same weight, because the ordering preference for R1 is 0.5. However, the R2 relation that holds between the same two units has a preference for realizing the nucleus B first. If our purpose is to enable the reader to derive as many inferences as possible, it would be then desirable to choose the text plan in which the ordering preference of the R2 relation is also satisfied. In this case, the text plan shown in figure 7.5.b will be the preferred plan. This position fully embraces Moore and Pollack's [1992] observation that both intentional and informational coherence should be accommodated by a theory of discourse. The mathematical model of text structures that was proposed in chapter 2 does not provide the means to explicitly represent multiple relations in the final discourse plans, but nevertheless, the extrinsic weight favors the plans that enable the reader to recover multiple discourse interpretations.

The extrinsic weight is given by a linear combination of the weights of the ordering and linear-adjacency constraints that are satisfied by each relation R that does not occur in the final text plan:

$$w_e(T) = \sum_{\text{R} \notin T} (0.25 w_{order}(\text{R}, T) + 0.25 w_{lin\_adj}(\text{R}, T)) \tag{7.13}$$

The coefficients that we use in (7.13) reflect the intuition that the extrinsic weight of a text

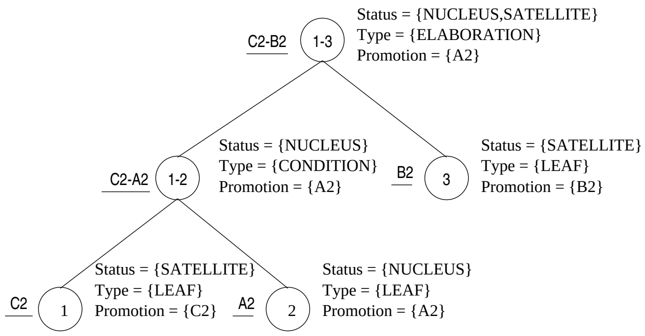

Figure 7.6: Example of a valid text plan for the problem in (7.3)–(7.4).

Relation	Intrinsic weight $w_i(\text{R}, T)$		
	$w_{order}(\text{R}, T)$	$0.5w_{tree\_adj}(\text{R}, T)$	$0.5w_{lin\_adj}(\text{R}, T)$
$rhet\_rel(\text{ELABORATION}, \text{B}_2, \text{A}_2)$	0.97	0.05	0.05
$rhet\_rel(\text{CONDITION}, \text{C}_2, \text{A}_2)$	0.59	0.49	0.49
$rhet\_rel(\text{CONDITION}, \text{C}_2, \text{B}_2)$			
$rhet\_rel(\text{MOTIVATION}, \text{A}_2, \text{C}_2)$			
$rhet\_rel(\text{MOTIVATION}, \text{B}_2, \text{C}_2)$			
$w_p(T)$: 2.64	1.56	0.54	0.54

Table 7.2: The intrinsic weights associated with the discourse tree in figure 7.6. Empty cells have weight zero.

plan is less important than the intrinsic weight (7.9). In the current implementation, the extrinsic weight formula uses coefficients that are half of the values of the coefficients that are used to evaluate the intrinsic weight of a text plan.

A complete example of the extrinsic and intrinsic weights associated with a planning problem and a text plan is given in tables 7.2 and 7.3, which present the weights that pertain to the text plan in figure 7.6.

7.4.2 A precise formulation of the bottom-up approach to text planning

As I specified in section 7.2.2, the key idea of the bottom-up approach to text planning is to formalize both the strong tendency of semantic units to obey a given ordering and the inclination of semantically and rhetorically related information to cluster into larger textual spans. The intrinsic and extrinsic weights that I introduced here provide an objective measure of the ordering and clustering tendencies. Given the discussion above, finding a solution to the text-planning problem corresponds then to finding a discourse tree that is valid, i.e., satisfies the constraints described in chapter 2, and whose weight is maximal. Since the total number of trees that can be built with a set of n units is very large $(n!4^{n-1}/\sqrt{\pi(n-1)^3})(1+O(\frac{1}{n})))$ [Sedgewick and Flajolet, 1996], it is obvious that we cannot merely enumerate all the trees and select then those that are valid and whose weights

Relation	Extrinsic weight $w_c(\text{R}, T)$	
	$0.25 w_{order}(\text{R}, T)$	$0.25 w_{lin\_adj}(\text{R}, T)$
$rhet\_rel(\text{ELABORATION}, \text{B}_2, \text{A}_2)$		
$rhet\_rel(\text{CONDITION}, \text{C}_2, \text{A}_2)$		
$rhet\_rel(\text{CONDITION}, \text{C}_2, \text{B}_2)$	0.147	
$rhet\_rel(\text{MOTIVATION}, \text{A}_2, \text{C}_2)$	0.182	0.160
$rhet\_rel(\text{MOTIVATION}, \text{B}_2, \text{C}_2)$	0.182	
$w_c(T)$: 0.671	0.511	0.160

Table 7.3: The extrinsic weights associated with the discourse tree in figure 7.6. Empty cells have weight zero.

are maximal.

7.4.3 Bottom-up algorithms for text planning

A Cocke-Kasami-Younger-like algorithm for text planning

The simplest way to solve the text-planning problem is to generate all the *valid* trees that can be built given the units in U and return those whose weights are maximal. This can be done by a variation of the Cocke-Kasami-Younger parsing algorithm [Younger, 1967] along the lines described by Brew [1992] (see figure 7.7). The algorithm starts with the initial set of n singleton trees that can be built with the units in U. It then constructs, at each step, all the valid trees that are made of i semantic units, where $i = 2 \ldots n$. Thus, for each i, the algorithm searches for all pairs of trees that have j and $i-j$ semantic units and builds a new tree with i semantic units if possible. The function $CanPutTogether(T_1, T_2, R_U)$ returns all relations $rhet\_rel(\text{R}, s, n) \in R_U$ that have not been used in the construction of any of the two trees T_1 and T_2 and whose arguments s and n belong to the set of salient units of the two trees. Each such relation R is used to enhance the set of valid trees that are associated with the entry $Chart[i]$. The trees that are added to the chart (see line 9 in figure 7.7) comprise both possible orderings in which two subtrees can be assembled.

Theorem 7.1. *Algorithm 7.7 is both sound and complete, i.e., it derives only valid trees and it always derives the valid trees of maximal weight.*

Sketch of the proof. The soundness of the CKY-like algorithm follows immediately from the observation that the trees that are appended to the *Chart* at each step $i \geq 1$ enforce the compositionality criterion and all other characteristics of valid text structures. The completeness of the algorithm follows by induction on the number of units given in the input. If the input contains only one unit, the corresponding tree is derived in line 1 of the algorithm. The CKY-like algorithm considers at each step all possible ways in which two valid trees can be put together to create a larger tree, which has the initial trees as subtrees of the root. Hence, the algorithm derives all the valid trees that can be built with the units

Input: A set $U = \{u_1, \ldots, u_n\}$ of n semantic units;
A set R_U of rhetorical relations that hold among the units in U.
Output: The text plans of maximal weight that can be built with the units in U.

```
Chart[1] := {T(u_1), ..., T(u_n)};
for i := 2 to n
    for j := 1 to i
        for each T_1 ∈ Chart[j]
            for each T_2 ∈ Chart[i - j]
                rels := CanPutTogether(T_1, T_2, R_U);
                if rels ≠ NULL
                  for each r ∈ rels
                      Chart[i] := Chart[i] ∪ newTree(r, T_1, T_2) ∪ newTree(r, T_2, T_1);
Select from Chart[n] the tree of maximal weight.
```

Figure 7.7: A Cocke-Kasami-Younger-like (CKY-like) algorithm for text planning.

in U and the relations R_U. Because it derives all the trees, it follows that it derives the trees of maximal weight as well. □

Although the CKY-like algorithm is both sound and complete, in the worst case it can generate an exponential number of trees.

A greedy Cocke-Kasami-Younger-like algorithm for text planning

If one gives up on completeness, the CKY-like algorithm can be modified so that not all valid discourse trees are generated, but only those that look more promising at every intermediate step. The CKY-like algorithm can be thus modified into a greedy one, which is more efficient because it generates for every pair of trees j and $i - j$ only one tree, that of local maximal weight.

A constraint-satisfaction-based (CS-based) algorithm for text planning

Another way to improve the efficiency of the CKY-like algorithm is by using constraint satisfaction techniques. In this subsection, I describe a CS-based algorithm that first approximates the rich tree-like structure of text plans by a linear sequence. That is, the algorithm determines the sequence of semantic units that is most likely to be coherent, i.e., satisfies most of the linear ordering and adjacency constraints. For some applications, this sequence is sufficient. For other applications, full text plans might be needed. In the latter case, the compilation algorithm described in chapter 3 can be used in order to build a full tree-plan on top of the sequence.

Input: A set $U = \{u_1, \ldots, u_n\}$ of n semantic units;
A set R_U of rhetorical relations that hold among the units in U.
Output: An ordering over U that is most likely to correspond to a coherent text.

1. Create a CSP problem with n variables, each ranging over the set $\{1, 2, \ldots, n\}$.
2. **for each** $rhet\_rel(\text{NAME}, u_i, u_j) \in R_U$
3. Assert weighted ordering and adjacency constraints for the units u_i, u_j.
4. **foreach** pair of units u_i, u_j that are not arguments of the same set of relations
5. Assert one unicity constraint.
6. Find an ordering of the elements in U for which the overall weight of the constraints that are satisfied is maximal.
7.* Use the compilation algorithm in figure 3.11 to build a valid tree structure on top of the sequence obtained at step 6.

Figure 7.8: A CS-based algorithm for text planning.

The CS-based algorithm (see figure 7.8) associates initially to each semantic unit in the input an integer variable whose domain ranges from 1 to n, where n is the cardinality of U. For example, algorithm 7.8 associates to input (7.3) – (7.4) three constraint variables, $v_{A_2}, v_{B_2}, v_{C_2}$, each ranging from 1 to 3.

For each rhetorical relation, the algorithm associates one weighted ordering and one weighted adjacency constraint along the lines described in section 7.4. For example, for the rhetorical relation $rhet\_rel(\text{CONDITION}, \text{C}_2, \text{B}_2)$, the ordering constraint is $v_{C_2} > v_{B_2}$ and has a weight of 0.41, and the adjacency constraint is $(v_{C_2} = v_{B_2} + 1) \vee (v_{C_2} = v_{B_2} - 1)$ and has a weight of 0.98. Hence, the adjacency constraints are formalized by stipulating that the difference between the values of the variables that are associated with the corresponding nucleus and satellite of a rhetorical relation be 1.

Since the CS-based algorithm uses only a linear representation of text plans, it is obvious that the modeling of the adjacency constraints is only an approximation of the way adjacency constraints are accounted for by the CKY-like algorithm. For example, the text plan in figure 7.9 has a greater weight than the weight that results from summing all the weights of the constraints that are satisfied by the linear sequence $\text{C}_2, \text{A}_2, \text{B}_2$. The reason is that, in the linear sequence, the adjacency constraint that pertains to relation $rhet\_rel(\text{MOTIVATION}, \text{B}_2, \text{C}_2)$ is not satisfied because units B_2, C_2 are not adjacent in the linear sequence; however, they *are* adjacent in the resulting tree, due to the nuclearity constraints.

Since in the CS-based approach the initial target is linear, with every pair of variables the algorithm asserts also a unicity constraint; this constraint prevents two semantic units being mapped into the same value. However, if two semantic units occur as arguments

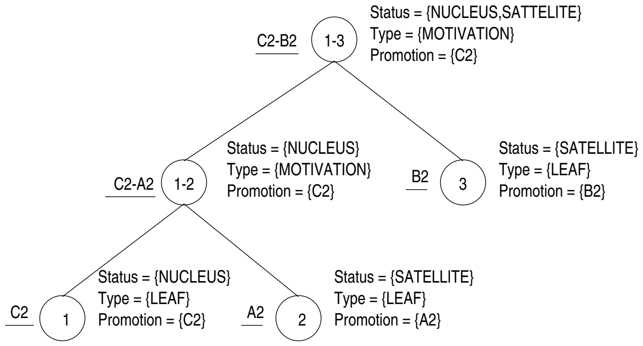

Figure 7.9: Example of a text plan whose weight is different from the weight of the corresponding linear plan.

of the same relations in a set R_U, it is impossible to distinguish between their rhetorical contributions to the text. In these cases, the unicity constraint is not asserted.

Once a constraint satisfaction problem has been derived from the input, any classical CS algorithm can be employed to find out the linear sequence whose overall weight is maximal. The compilation algorithm in figure 3.11 can then be applied to this sequence and full text plans can be obtained. In figure 7.8, the last step of the CS-based algorithm is labelled with a $\star$ symbol, in order to denote this optionality.

The CS-based implementation finds an ordering of the elements in U that maximizes the number of ordering and adjacency constraints that are satisfied. As I have discussed above, the treatment of adjacency constraints is just an approximation of the correct treatment that pertains to the CKY-like algorithm. In addition, in the case of the CS-based algorithm, the contribution of each of the rhetorical relations in the input is not affected by that relation being used or not in the final tree structure of the text. This contrasts again with the treatment in the CKY-like algorithm, where the rhetorical relations that participated directly in the discourse representation contributed more to the final weight of the tree than the relations that were not used in the final discourse structure. Because of these approximations, it is possible that the CS-based algorithm would generate sequences that are different from those derived by the CKY-like algorithm.

7.5 Implementation and experimentation

I implemented in Common Lisp both the Cocke-Kasami-Younger-like and the CS-based algorithms. The constraint-satisfaction based algorithm was also integrated in the Sentence Planner architecture of the HealthDoc Project [Wanner and Hovy, 1996, Hovy and Wanner, 1996, DiMarco and Foster, 1997, DiMarco *et al.*, 1997, Hirst *et al.*, 1997], whose goal is to produce medical brochures that are tailored to specific patients. In fact, the semantic units in (7.1) are members of a large KB that encodes information to be given to diabetic patients.

A knowledge base in HealthDoc, which is called a *Master Document*, encodes all the material that is needed in order to generate customized documents for different types of patients. The semantic units of a Master Document are represented using a variant of the Sentence Plan Language [Penman Project, 1989, Kasper, 1989] and are annotated with information that concerns the suitability of the units for being conveyed to a particular patient, with the rhetorical relations that hold among units, with coreference links, etc. When the system is given as input a set of features that characterize a patient's age, medical history, cultural background, etc., it selects the set of semantic units that are relevant for that patient. After the units have been selected, the CS-based algorithm runs and returns an ordering of the semantic units that is most likely to be coherent. When given, for example, the semantic units in (7.1) among which the rhetorical relations in (7.2) hold, the HealthDoc discourse module that implements algorithm 7.8 proposes that in order to be coherent, the semantic units should be realized in the order $\mathrm{D}_1, \mathrm{A}_1, \mathrm{C}_1, \mathrm{B}_1$, which corresponds roughly to this text:

(7.14) The condition that you have is insulin-dependent diabetes. Insulin-dependent diabetes is the less common type of diabetes. With insulin-dependent diabetes, your body makes little or no insulin. The pancreas, a gland found just behind the stomach, normally makes insulin.

Once the discourse structure for the text has been fixed, other modules operate on the semantic units in the structure. Up to this point, the following modules have been implemented (see [Hovy *et al.*, 1998] for a detailed discussion):

Aggregation — to remove redundancies across neighboring expressions;

Ordering — to place clause constituents in specific positions in the emerging sentence;

Reference — to plan pronominal and other reference.

Each of the modules operates on the sequence of SPL structures that was planned by the discourse module and modifies it in order to increase the quality of the text. After the other modules operate on the structure that was derived by the discourse module, the resulting text is this:

(7.15) The condition that you have is insulin-dependent diabetes, which is the less common type of diabetes. With this condition, your body makes little or no insulin. Insulin is normally made in a gland called the pancreas found just behind the stomach.

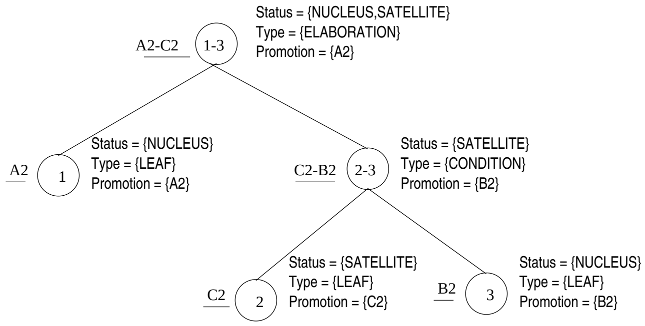

Figure 7.10: The text plan of maximal weight that corresponds to problem (7.3)–(7.4).

Other issues

Although the CKY-like and the CS-based algorithms do not enumerate all the trees that can be built with the units given as input, they are still highly expensive in both space and time. In practice, I have noticed that my current implementations cannot be applied to problems that have more than 20 units in their inputs. Because in the HealthDoc system the text planning algorithms are applied within each section separately, the exponential nature of the problem does not seem to hamper the overall performance of the system. Nevertheless, if these algorithms are to be applied to larger problems, better heuristics would need to be developed in order to enable a faster convergence towards a solution.

The corpus analysis in chapter 4 provides information not only on the ordering and clustering preferences of various relations, but also on the markers that can be used to signal various rhetorical relations. If one simply embeds these markers into the final texts, one can realize, for example, the text plan in figure 7.6 as "*If* you come home early, we can go to the bookstore. We can go to Sam's bookstore". Although implementing such an algorithm is trivial, it seems that a proper account of the discourse markers should take into consideration the local lexicogrammatical constraints as well. For example, in some cases, it would be inappropriate to use a discourse marker twice. In other cases, the use of some discourse markers simply does not sound right. The investigation of the ways in which the modules of the system could interact in order to integrate the markers suggested by the discourse module is beyond the scope of this thesis.[3]

7.6 Generating discourse plans that satisfy multiple communicative goals

By default, the algorithms introduced in this chapter find plans that satisfy the goal "tell everything that is in the KB". For example, when only the default goal is used, algorithm 1 generates for the problem (7.3)–(7.4) one valid tree of maximal weight 3.507 (see figure 7.10) — a possible realization of the text plan in 7.10 is shown in (7.16) below.

(7.16) We can go to the bookstore. If you come home early, we can go to Sam's bookstore.

However, when generating text, it is often useful to specify more than one communicative goal. In some cases, besides informing, we may also want to motivate, persuade, or deceive the reader.

Traditionally, top-down planning algorithms are given as input only one high-level communicative goal. Although we can modify the goal expansion process that characterizes top-down text planners so that the branches that use goals specified in the input are preferred over branches that do not use such goals, we still run into the same problem that we discussed in the beginning of the chapter: there is no way to ensure that all the information that we want to communicate will be eventually included in the final text plan. In addition, the procedure described above assumes that the system can determine the communicative goal that it needs to satisfy first: after all, the system has to start the expansion process from somewhere. In the general case, such an assumption is unreasonable; and enumerating all the possible combinations is too expensive.

In contrast with top-down planning algorithms, the bottom-up text-planning algorithms can be easily adapted to generate plans that satisfy multiple communicative goals. For example, one can specify that besides conveying the information in the KB, another high-level communicative goal is to motivate the reader to come home early ($\text{MOTIVATE}(hearer, \text{C}_2)$). Such a communicative goal can be mapped into an extra constraint that the final discourse plan has to satisfy: in this case, the extra constraint will require that the final discourse plan uses at least one rhetorical relation of MOTIVATION that takes C_2 as nucleus. When such a constraint is specified, there is one tree of maximal weight 3.227 that is returned by the CKY-like algorithm, that shown in figure 7.11. A possible realization of the text plan in figure 7.11 is shown in (7.17) below.

(7.17) Come home early! That way, we can go to the bookstore. We can go to Sam's bookstore.

[3] See [Moser and Moore, 1997, Di Eugenio *et al.*, 1997] for a more sophisticated analysis of the relationship between discourse structure and cue phrases.

Along the lines described here, one can also specify conjunctions and disjunctions of communicative goals and pragmatic constraints that characterize ordering preferences on the linear realization of semantic units. As an example, consider the generation problem given below, i.e, the set of semantic units shown in (7.18) and the corresponding set of rhetorical relations shown in (7.19).

$$(7.18) \quad U_3 = \begin{cases} \text{A}_3 = \text{“About 30\% of the teenagers will become experimental smokers.”} \\ \text{B}_3 = \text{“We know that 3,000 teens start smoking each day.”} \\ \text{C}_3 = \text{“About 90\% of teenagers once thought that smoking was} \\ \qquad \text{something that they'd never do.”} \\ \text{D}_3 = \text{“Of the teenagers who will start smoking, about 90\% will end up} \\ \qquad \text{with a pack and a lighter for the rest of their lives.”} \\ \text{E}_3 = \text{“Teenagers want to stay non-smokers.”} \\ \text{F}_3 = \text{“The pressure to smoke in junior high is greater than it will} \\ \qquad \text{be any other time of one's life.”} \\ \text{G}_3 = \text{“About 75\% of the young adults will pick up a cigarette and let} \\ \qquad \text{curiosity take over.”} \end{cases}$$

$$(7.19) \qquad R_{U_3} = \begin{cases} rhet\_rel(\text{EVIDENCE}, \text{A}_3, \text{F}_3) \\ rhet\_rel(\text{EVIDENCE}, \text{B}_3, \text{F}_3) \\ rhet\_rel(\text{EVIDENCE}, \text{D}_3, \text{F}_3) \\ rhet\_rel(\text{EVIDENCE}, \text{G}_3, \text{F}_3) \\ rhet\_rel(\text{CONCESSION}, \text{A}_3, \text{C}_3) \\ rhet\_rel(\text{CONCESSION}, \text{B}_3, \text{C}_3) \\ rhet\_rel(\text{CONCESSION}, \text{D}_3, \text{C}_3) \\ rhet\_rel(\text{CONCESSION}, \text{G}_3, \text{C}_3) \\ rhet\_rel(\text{JUSTIFICATION}, \text{E}_3, \text{F}_3) \\ rhet\_rel(\text{RESTATEMENT}, \text{C}_3, \text{E}_3) \end{cases}$$

Given the generation problem in (7.18)–(7.19), the CS-based algorithm will create a constraint satisfaction problem with seven variables, $v_{\text{A}_3}, v_{\text{B}_3}, \ldots, v_{\text{F}_3}$, each ranging from 1 to 7. It will associate with these variables the corresponding ordering and adjacency contraints. However, given the set of rhetorical relations (7.19), one can see that the algorithm cannot distinguish between units $\text{A}_3, \text{B}_3, \text{D}_3$, and G_3 because rhetorical relations of EVIDENCE and CONCESSION hold between each of these units and units F_3 and C_3 respectively. Consequently, no unicity constraints are associated with any pairs of variables $v_{\text{A}_3}, v_{\text{B}_3}, v_{\text{D}_3}, v_{\text{G}_3}$; however, unicity constraints are asserted for all the other pairs.

For problem (7.18)–(7.19), algorithm 7.8 generates the partial ordering

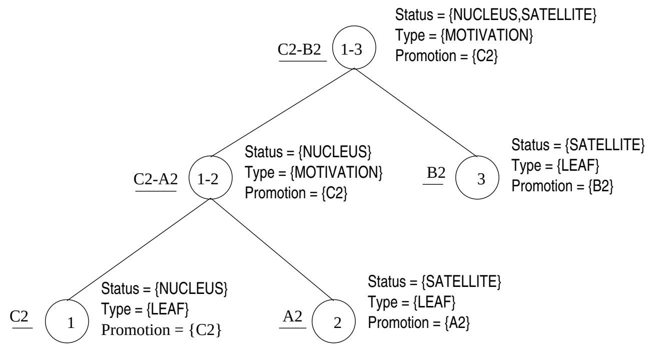

Figure 7.11: A text plan that corresponds to problem (7.3) – (7.4). The text plan satisfies multiple communicative goals.

$E_3 < F_3 < A_3, D_3, B_3, G_3 < C_3$. This partial ordering yields $4! = 24$ total orderings that correspond to 24 different ways in which the units can be realized as coherent text. Texts (7.20) and (7.21) exemplify two of these possible realizations.

(7.20) [No matter how much one wants to stay a non-smoker,[E3]] [the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life:[F3]] [30% of the teenagers will become experimental smokers.[A3]] [Of those who will start smoking, about 90% will end up with a pack and a lighter for the rest of their lives.[D3]] [We know that 3,000 teens start smoking each day[B3]] [and that 75% of the young adults will pick up a cigarette and let curiosity take over,[G3]] [although it is a fact that 90% of them once thought that smoking was something that they'd never do.[C3]]

(7.21) [No matter how much one wants to stay a non-smoker,[E3]] [the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life:[F3]] [75% of the young adults will pick up a cigarette and let curiosity take over,[G3]] [About 30% of them will become experimental smokers.[A3]] [Of those who will start smoking, about 90% will end up with a pack and a lighter for the rest of their lives.[D3]] [We know that 3,000 teens start smoking each day,[B3]] [although it is a fact that 90% of them once thought that smoking was something that they'd never do.[C3]]

From the perspective of coherence, there is no difference between texts (7.20) and (7.21). However, from a pragmatic perspective there is a large difference. Empirical research in communication studies, psychology, and social studies of persuasion [McGuire, 1968, Stiff, 1994] have shown that the likelihood of achieving persuasion grows when arguments are presented in increasing order of their gravity.[4] For example, the units that can play the role of EVIDENCE for the information given in F_3 can be ordered from a pragmatic perspective according to a scale of gravity. On such a scale, G_3 seems to be less serious than A_3, which in turn is less serious than D_3, which is less serious than B_3. If this knowledge is recorded in the initial set of constraints (7.19), $G_3 \prec A_3 \prec D_3 \prec B_3$, it can be used as a direct constraint by the CS-based or the CKY-like algorithm. When this extra constraint is considered, the CS-based algorithm, for example, yields only one ordering of maximal weight, $E_3 < F_3 < G_3 < A_3 < D_3 < B_3 < C_3$. Text (7.21), which corresponds to this ordering, is not only coherent, but is also more likely to convince a teenage reader not to smoke.

7.7 Shortcomings of the bottom-up approach to text planning

The bottom-up approach to text planning that I proposed in this chapter assumes that text coherence can be achieved by satisfying as many of the ordering and clustering constraints as possible. A couple of concerns can be raised in connection with this approach.

- First of all, there are no psycholinguistic experiments to support the assumption. The corpus data provides information only with respect to individual rhetorical relations and it says nothing about their composition.

- Moreover, given the nature of the Brown corpus, which is a collection of texts of various genres, the strengths of the ordering and clustering constraints are not tailored to any specific domain. Being averages over all existing text genres, these scores might not be adequate for a legal or technical domain, for example.

- And most importantly, the ordering of the textual units in a text plan might be influenced by factors that are not captured by our corpus analysis, such as focus, the distribution of given and new information in discourse, and high-level pragmatic and intentional constraints.

[4] Marcu [1996, 1997] reviews empirical research on persuasion in communication studies, psychology, and social studies and discusses its impact on the task of natural language generation from the perspective of content selection, content organization, realization, and lexical choice.

For example, let us take the textual units in text A.4 and the corresponding rhetorical relations, which we reproduce for convenience below.

(7.22) [Farmington police had to help control traffic recently[1]] [when hundreds of people lined up to be among the first applying for jobs at the yet-to-open Marriott Hotel.[2]] [The hotel's help-wanted announcement — for 300 openings — was a rare opportunity for many unemployed.[3]] [The people waiting in line carried a message, a refutation, of claims that the jobless could be employed if only they showed enough moxie.[4]] [Every rule has exceptions,[5]] [but the tragic and too-common tableaux of hundreds or even thousands of people snake-lining up for any task with a paycheck illustrates a lack of jobs,[6]] [not laziness.[7]]

$$(7.23) \quad \begin{cases} rhet\_rel(\textsc{volitional\_result}, 1, 2) \\ rhet\_rel(\textsc{circumstance}, 3, 2) \\ rhet\_rel(\textsc{background}, 2, 4) \\ rhet\_rel(\textsc{evidence}, 6, 4) \\ rhet\_rel(\textsc{concession}, 5, 6) \\ rhet\_rel(\textsc{antithesis}, 7, 6) \end{cases}$$

When we give the rhetorical relations in (7.23) as input to the CKY-like algorithm, the text plan of maximal score that is produced as output corresponds to the ordering shown in (7.24), below.

(7.24) $3 < 2 < 1 < 4 < 7 < 5 < 6$

A possible paraphrase of text plan (7.24) is shown in (7.25).

(7.25) [The Marriot Hotel's help-wanted announcement — for 300 openings — was a rare opportunity for many unemployed.[3]] [When hundreds of people lined up to be among the first applying for jobs at the yet-to-open hotel,[2]] [Farmington police had to help control traffic.[1]] [The people waiting in line carried a message, a refutation, of claims that the jobless could be employed if only they showed enough moxie.[4]] [The tragic and too-common tableaux of hundreds or even thousands of people snake-lining up for any task with a paycheck does not illustrate laziness.[7]] [Every rule has exceptions,[5]] [but the people's snake-lining definitely illustrates a lack of jobs.[6]]

Although text (7.25) is easy to understand and contains all the semantic information that was provided in the original text, it proposes a different ordering of the elementary units.

The first part of the text, i.e., units 1 to 4 seem to convey the same information in spite of being ordered differently than they were in the original text. However, the proposed ordering of units 5 to 7 in text plan (7.24) does not yield a text as nicely balanced as the original. This suggests that in building text plans we should go beyond ordering and clustering constraints.

7.8 Related work

To plan paragraphs, some researchers [Carcagno and Iordanskaja, 1989] start with a tree-like structure that contains all the information that a system could communicate; and then trim and tailor it so that it eventually fits the communicative requirements. Other researchers use recipes [Dale, 1989], house-building plans [Mellish, 1988], mathematical proofs [Zukerman and Pearl, 1986, de Souza and Nunes, 1992, Huang, 1994, Huang and Fiedler, 1997], or the hierarchical structure of tax forms [Weiner, 1980]. In fact, each of these approaches is a direct form of exploitation of the internal structure of some underlying domain. Coherence results as a side-effect of a predetermined internal structure.

In contrast to these approaches, a number of paradigms have been developed to offer more flexibility. In the rest of this section, I review schema-, RST-, and hierachical-planning-based paradigms and discuss their ability to find text plans that subsume all the information given in a knowledge base. I also discuss the ability of these approaches to produce text plans that satisfy multiple high-level communicative goals.

7.8.1 Text plans in schema-based approaches

Schemata, as used by McKeown [1985], are computational devices that have been designed to deal both with content determination and organization. In McKeown's terms, a schema is a compilation of conventional patterns that occur across various expository texts, which is expressed in terms of Grimes's rhetorical predicates [1975]. For example, an *identification* schema captures the strategy that is used for providing definitions. It includes the identification of an item as a member of a generic class, the description of the attributes of the object, analogies, and examples (see figure 7.12). When activated, each predicate in a schema is mapped to a query in the knowledge base, thus determining not only the content of the final text but also the order in which the information retrieved from the knowledge base is realized. The result of traversing a schema is a text plan that has a tree structure of the form given in figure 7.13. The nodes in the plan are either rhetorical predicates or recursively similar schema applications. The tree structure is supposed to be mapped into sentences by a left-to-right traversal of the leaves, which are all rhetorical predicates.

Despite their ability to generate structured paragraphs, useful for, say, simple descriptions of objects or equipment, schemata are not a rich enough representation for

Identification schema

Identification (class & attributive/function)
{Analogy/Constituency/Attributive/Renaming/Amplification}*
Particular-illustration/Evidence+
{Amplification/Analogy/Attributive}
{Particular-illustration/Evidence}

Eltville (Germany) (1) An important wine village of the Rheingau region. (2) The vineyards make wines that are emphatically of the Rheingau style, (3) with a considerable weight for a white wine. (4) Taubenberg, Sonnenberg, and Langenstuck are among vineyards of note.

(1) Identification
(2) Attributive
(3) Amplification
(4) Particular-illustration

Figure 7.12: An identification schema and an example of its use [McKeown, 1985].

text plans since, in this framework, text is modeled as a sequence of rhetorical predicates whose contribution to the ideational, interpersonal, or textual facet of the message cannot be evaluated; the top-level node and the virtual nodes that stand for recursive schema applications are the only nodes that carry information about communicative goals. As algorithmic artifacts, schemata are rigid structures that are not amenable to different orderings of the rhetorical predicates. Further developments [McKeown *et al.*, 1990, Paris, 1991] have shown how a user model can be used to improve the selection of information for different rhetorical predicates, but still, schemata seem conceptually inadequate for the flexibility and richness that text plans have to provide. In other words, schemata are not suitable for a knowledge-driven approach to generation, such as that used in HealthDoc, whose task is to map a segment of a knowledge base into natural text: the success rate of such a mapping could be anywhere between zero and 100% and cannot be predicted unless a schema is actually applied. Since most of the nodes of a text plan that was derived using a schema-based approach are not annotated with intentional or ideational information, it is impossible to generate flexible text plans that satisfy multiple high-level communicative goals.

7.8.2 Text plans in RST-based approaches

The idea of using RST relations in a generation setting is due to Mann [1984] and was first operationalized by Hovy [1988b]. RST-based approaches to text planning employ the constraints on discourse structures that were originally defined by Mann and Thompson [1988]. Hence, although the discourse structures are similar to those formalized in chapter 2, they do not obey all the constraints: for example, some rhetorical structure trees (RS-trees)

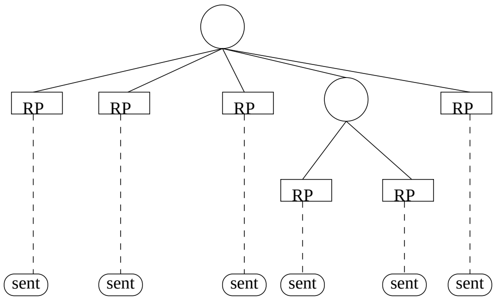

Figure 7.13: Text structure in schema-based approaches. Circles represent virtual nodes that result when schemata are applied recursively. Boxes represent rhetorical predicates that are eventually mapped to individual sentences.

are non-binary structures and none of them employs the compositionality criterion that is essential in the work described here.

Hovy's main contribution [1988b] consists in attaching to each RST relation a communicative intent, and in viewing the combination of relations into paragraphs as a planning process. An operational RST relation contains *growth points* that are used by the text planner to attach more information to the structure that is constructed. The RST tree is built through a hierarchical expansion of the goals in which the growth points can be omitted. The strategy that chooses between different alternatives in the expansion process is rudimentary.

Besides being the first attempt to operationalize RST relations, the main merit of the method consists in the soundness of the algorithm, which ensures that any partial tree reflects the principles of coherence that are captured in the definitions of the operators. However, the approach suffers also from the same symptoms the schema-based approach does: given a specific segment of knowledge, it cannot predict how much of that knowledge will be mapped into text by the application of the goal expansion algorithm. And, although the nodes of the RS-trees contain both intentional and ideational information, there does not seem to be any straightforward way for generating textual plans that satisfy multiple communicative goals. For highly specialized tasks, such as the generation of purpose clauses in instruction settings [Vander Linden *et al.*, 1992, Vander Linden, 1993, Vander Linden, 1994] and generation of automobile maintenance instructions [Rösner and Stede, 1992], alternative methods that are based on systemic network traversals and recursive procedure calls have been proposed for deriving RS-trees. However, these methods do not offer a solution to the problem of mapping a whole knowledge base segment into a text plan because text plans are obtained through the same refinement process that recursively maps an initial goal into a fully fleshed tree structure.

Even if we ignore completely the computational problems that characterize the operationalization of RST, increasing evidence comes to challenge the sufficiency of RST relations for the generation of high quality, coherent text [Hovy, 1990b, Cawsey, 1991, Moore and Pollack, 1992, Moore and Paris, 1993, de Souza and Nunes, 1992, Bateman and Rondhuis, 1994]. The main drawback seems to be the dual ambiguity of the RST relations, since there may be more than one rhetorical structure for a given text, and more than one text for a given rhetorical structure.

As far as I know, the algorithms described in this chapter constitute the first attempt to address the ambiguity that concerns the one-to-many mapping between texts and discourse structures. The bottom-up algorithms described in this chapter can generate not just one plan but many text plans that subsume the information given in a knowledge base. The metric defined in section 7.4 provides the means for determining text plans of maximal weight. The metric accommodates not only coherence constraints, but pragmatic constraints as well. The ambiguity that concerns the one-to-many mapping between text structures and texts is the direct effect of the assumptions that underlie Mann and Thompson's theory: RST is meant to describe relations among segments of text, whether or not they are grammatically or lexically signaled. To account for the multiple ways in which a rhetorical relation can be realized, Wanner [1994] argues for the use of a set of *lexically-biased* discourse structure relations that are derived on the basis of Mel'čuk's *lexical functions* [Mel'čuk and Polguère, 1987]. The lexically-biased discourse relations are meant to provide the mechanism for refining a rhetorical relation into a member of the set that encompasses all the legal sequences of lexical functions through which this rhetorical relation can be realized. Although Wanner's mechanism enumerates all the possible ways in which a rhetorical relation can be realized, it does not provide the criteria that would enable one to choose a preferred alternative.

The work described in this thesis does not address explicitly the second form of ambiguity. However, as I have emphasized quite often throughout this thesis, the formalization of valid text structures and the algorithms proposed here need not work only in conjunction with rhetorical relations of the kind proposed by RST. It is possible that applying the same algorithms in conjunction with a set of rhetorical relations that are closer to the lexicogrammatical resources of a given language could provide a solution for the second form of ambiguity as well.

7.8.3 Text plans in hierarchical-planning-based approaches

Hierarchical text planners [Moore and Paris, 1989, Moore and Swartout, 1989, Cawsey, 1990, Moore and Swartout, 1991, Cawsey, 1991, Maybury, 1992, Moore and Paris, 1993, Maybury, 1993, Mittal, 1993, Mittal and Paris, 1993, Moore, 1995, Reed and Long, 1997b, Reed and Long, 1997a] attempt to make the process of text planning more flexible and to account for a large variety of linguistic phenomena, such as focus, intentions, argumentation, and persuasion.

Maybury [1992, 1993] constructs a set of plan operators that can be used for generating explanatory texts by formalizing within the operators all the possible ways in which one can explain something; these heuristics were discovered through a corpus study of explanatory texts. For example, in order to generate a description, one can use a definition, division, detail, comparison, or analogy. Other text planners [Moore and Paris, 1989, Moore and Swartout, 1989, Cawsey, 1990, Moore and Swartout, 1991, Cawsey, 1991, Moore and Paris, 1993, Mittal, 1993, Mittal and Paris, 1993, Moore, 1995] try to achieve flexibility and generality by encoding in plan operators the intentions, effects, and constraints that were developed within RST. And yet other approaches, such as that taken in Diogenes [Defrise and Nirenburg, 1990, Nirenburg *et al.*, 1989] and Spokesman [Meteer, 1991b, Meteer, 1991a, Meteer, 1992], achieve flexibility by enriching the language of goal refinement that was developed by Hovy [1988b]. In both Diogenes and Spokesman plans are hierarchically organized sets of frames. Each frame contains structural links to its parent and children, partial sentence plans, plan-role relations, locutionary information, etc. A top-down expansion process refines an initial goal into a fully fleshed-out text plan.

In spite of their flexibility and the large variety of linguistic phenomena that they handle, hierarchical planners cannot ensure that *all* the knowledge that makes up a knowledge pool will be eventually mapped into the leaves of the resulting text plan: after building a partial text plan, which encodes a certain amount of the information found in the initial knowledge pool, it is highly possible that the information that is still unrealized will satisfy none of the active communicative goals.

Before ending the discussion on text planning, I would like to discuss a problem of terminology that concerns the inadequacy of using the notion of "hierarchical planning" in conjunction with NLG systems, a notion that was defined originally by Sacerdoti [1974]. The idea behind Sacerdoti's system, Abstrips, was that a problem can be first solved in an abstract space and then refined at levels that are successively more detailed. Consider the case of a robot that is to collect a number of cans that are spread out in three rooms: it seems reasonable to assume that in accomplishing this task, the robot will perform actions such as *move to a given room, move to a specific can,* and *pick up a can.* In order to pick up a can, a robot needs not only to have its arm empty, but also be in the room where the can is. Being in the same room with the can seems to be *more critical* than having the arm empty. Sacerdoti's idea was to assign *criticalities* to plan operators, i.e., numbers that indicate the relative difficulty of satisfying the preconditions of each operator. The planner uses these criticalities to find first an abstract plan that satisfies only the preconditions of the operators with the highest values for criticalities. In the robot example, with appropriate definitions for plan operators and appropriate assignments of criticality values for the preconditions, the robot will first find a plan that will take it to each of the three rooms. Once this plan is built, it can be further refined, so that in each room the robot will pick up the cans.

The first point I want to make is that abstract planning means *finding complete solutions* at different levels of abstraction and that Sacerdoti's approach works in cases where the levels of refinement differ *only* with respect to the preconditions embedded in the plan operators. The second point I want to make is that the very notion of planning requires an ability of the system *to reason about the effects of the actions* that are taken; in order to pick up the cans from, let us say, room two, the robot has to be in room two. *None of these requirements is addressed in most of the natural language approaches to text planning!*

Hovy [1988b, 1990b, 1991, 1992, 1993], Cawsey [1990, 1991], Moore and Paris [1989, 1993], Moore and Swartout [1989, 1991], Moore [1995], Maybury [1992, 1993], and Mittal [1993] all claim that their systems employ hierarchical planning in the style of Abstrips, but none of them satisfy the requirements that I mentioned in the previous paragraph. In fact, all these planners perform goal expansion and goal matching in which partial structures do not constitute complete, abstract solutions for the goal, or even worse, in which different levels of abstractions are mixed more or less randomly; and none of the planners deals seriously with the effects of the actions that are taken. In the best case, the systems monitor the effects that some communicative acts have on the hearer, but none of them reasons about the effects of these actions on the text itself. I include here effects such as *information X has been conveyed*, *concept Y has been constructed and used* or *discourse unit Z has been realized and consequently is available from now on as discourse unit referent.*

As acknowledged by many researchers in planning [Georgeff, 1987, Knoblock, 1992] and computational linguistics [Meteer, 1992, Rubinoff, 1992, Young *et al.*, 1994, Young and Moore, 1994], hierarchical planners impose a homomorphic constraint between the different levels of abstraction: if there is a solution at the ground level, then there exists a corresponding solution at the more abstract levels as well, and vice versa. A strong restriction of hierarchical planners is that they cannot account for effects in different subtrees and that they assume that the preconditions that are determined to be details in the abstraction process are independent [Knoblock, 1992]. Moreover, text planners are unable to distinguish between intended effects and side effects [Young *et al.*, 1994, Young and Moore, 1994], between crucial steps and unimportant ones [Huang, 1994]. Although there is an abundance of approaches that claim to do Abstrips-like planning, only Meteer [1991a, 1992] and Reed and Long [1997b, 1997a] attempt to solve the problems that are enumerated above from the perspective taken in hierarchical planning. In most cases, the requirement of homomorphism is not addressed at all, or is watered down to an interface problem between the abstract text planner and the linguistic realizer, in which the former plans with the constraints that are put forth by the latter (Hovy [1988a, 1988c, 1990a], Rubinoff [1992]).

7.9 Summary

In this chapter, I have presented empirical results that concern the strength of the tendencies of rhetorically related units to obey a given nucleus-satellite ordering and to cluster into large textual spans. I have then shown how these strengths can be used in order to assign a weight to a discourse structure: the larger a weight, the higher the likelihood that the discourse structure is coherent. I have introduced a new, data-driven approach to the text-planning problem and proposed three algorithms that map full knowledge bases into valid discourse trees. I have also shown how these algorithms can be used to generate text plans that satisfy multiple high-level communicative goals and discussed briefly how the text plans produced by the algorithms are further refined into English in HealthDoc, a generation system that produces health-education materials that are tailored to particular audiences.

Chapter 8

Conclusions

This thesis contributes to the understanding of the linguistic and formal properties of the high-level, rhetorical structure of unrestricted texts; the computational means that enable the derivation of this structure in the context of both natural language understanding and generation; and the relationship between text structures and text summaries. In this chapter, I critically review these contributions and suggest future work.

8.1 The linguistic and formal properties of text structures

8.1.1 Contributions

The formulation of the weak and strong compositionality criteria for valid text structures

In chapter 2, I have shown that the current lack of algorithms to derive the high-level structure of unrestricted text can be explained not only by the ambiguity of the definitions of rhetorical relations that are proposed by various theories but also by the lack of a compositionality criterion, one that would explain the relationship between rhetorical relations that hold between large textual units and rhetorical relations that hold between elementary units. In chapter 2, I have first proposed a weak compositionality criterion. This criterion has been proven to be useful for a manual investigation of discourse, but too weak to be formalized using state-of-the-art techniques. To circumvent this problem, I have strengthened the weak criterion, thus obtaining a strong compositionality criterion that can be easily formalized in the language of first-order logic.

The formalization of the mathematical properties of the high-level structure of unrestricted text

I have provided a first-order formalization of the mathematical properties of the high-level structure of unrestricted text. The formalization assumes that texts 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; that trees are a good approximation of the abstract structure of text; and that valid text structures obey the strong compositionality criterion given in proposition 2.2.

The formalization that I have proposed is independent of the taxonomy of rhetorical relations that it relies upon. As an example, I have shown how, by adopting the taxonomy of relations proposed by Mann and Thompson [1988], one can obtain a formalization of Rhetorical Structure Theory.

The melding of Mann and Thompson's and Grosz and Sidner's discourse theories

Taking as a starting point a discussion of Moser and Moore [1996], I have provided a formalization of the relationship between text structures and intentions. More precisely, I have provided a unified formalization of Mann and Thompson's Rhetorical Structure Theory [Mann and Thompson, 1988] and Grosz and Sidner's theory [1986]. The melding of structure- and intention-based constraints enables the derivation of intentional inferences on the basis of the structure of text and provides a means for using intentional judgments for reducing the ambiguity of text structures.

8.1.2 Shortcomings and future work

The main shortcoming of my formal inquiry into the structure of text and the relationship between structures and intentions comes from its simplicity. The formalization presented in this thesis completely 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 thesis is a research direction that promises to be extremely rewarding.

Even if we ignore for the moment the linguistic phenomena that are currently not dealt with in the model, we can still attack the assumptions on which the formalization relies. For example, the formalization assumes that text can be sequenced into elementary units, but as I have discussed in chapter 4, providing an objective definition for this is not trivial. And the same holds with respect to providing objective definitions for a taxonomy of rhetorical relations. And even if these problems are given adequate solutions, in some cases, the tree structures that are formalized here still seem to be insufficient for explicitly representing multiple relations that hold between various textual units. Future research will have to provide means for relaxing the tree-like structure used here in order to enable one textual unit to be related to more than one unit in the formal representation of text.

8.2 The algorithmic derivation of valid text structures

8.2.1 Contributions

The algorithmic derivation of valid text structures — theoretical issues

In chapter 3, I have proposed four paradigms for solving the problem of text structure derivation given in definition 2.2. Two paradigms apply model-theoretic techniques; they yield two algorithms:

- One algorithm maps a text structure derivation problem with N elementary units into a constraint-satisfaction problem with $3\mathrm{N}(\mathrm{N}+1)/2$ variables and $1/12(\mathrm{N}^4+4\mathrm{N}^3+5\mathrm{N}^2+2\mathrm{N}+12)$ constraints.

- The other algorithm maps a text structure derivation problem into a propositional satisfiability problem with at most $O(\mathrm{N}^3)$ variables and $O(\mathrm{N}^5)$ conjunctive-normal-form constraints.

Two paradigms apply proof-theoretic techniques.

- One paradigm yields a proof theory and an algorithm for deriving valid text structures that is both sound and complete.

- The other paradigm maps a text structure derivation problem into a recognition problem with a grammar in Chomsky normal form. An algorithm that uses this paradigm and that is shown to be both sound and complete solves a derivation problem with N elementary units in $O(\mathrm{N}^6)$.

The algorithmic derivation of valid text structures — empirical issues

I have empirically compared the algorithms pertaining to the four paradigms on a set of eight text derivation problems. The comparison has shown that the algorithm that uses grammars in Chomsky normal form outperforms the one that implements straightforwardly the proof-theoretic account, which in turn outperforms the algorithms that use propositional satisfiability, which in turn outperforms the algorithm that applies traditional constraint-satisfaction techniques.

Within the class of algorithms that use propositional satisfiability, I have empirically compared the ability of Davis–Putnam's exhaustive procedure and two greedy procedures, GSAT and WALKSAT to find models of text structure derivation problems. Surprisingly, I found that the Davis–Putnam procedure outperforms the greedy methods.

A non-incremental approach to text structure derivation

The paradigms of text structure derivation that I have proposed in this thesis depart from the incremental approaches to discourse processing that are ubiquitous in the literature. The main advantage of the approach to discourse structure derivation that I have proposed here is that it cannot lead to nonmonotonic interpretations, as incremental approaches can.

8.2.2 Shortcomings and future work

Although computationally efficient, the non-incremental approach to discourse processing that I have proposed in this thesis is not psycholinguistically plausible. By choosing efficiency over cognitive plausibility, I have preempted any possibility of modelling phenomena such as mistakes and re-interpretations, which are common in discourse. Nevertheless, the formalization of text structures in chapter 2 poses no constraints on the algorithms that can derive those structures. Future research can produce produce algorithms that are both efficient and incremental.

The empirical comparison between the ability of Davis–Putnam's, GSAT, and WALK-SAT procedures to find models for propositional theories suggests that greedy methods might not be better than exhaustive methods for satisfiability problems that are highly structured. Unfortunately, the empirical work described in chapter 3 is insufficient to warrant the validity of such conclusion. Further research can nevertheless shed more light on this issue.

8.3 The corpus analysis of cue phrases

8.3.1 Contributions

A comprehensive analysis of cue phrases

The corpus analysis discussed in chapter 4 constitutes the largest empirical study of the relationship between cue phrases, the rhetorical relations that they signal, the rhetorical status and the boundaries of the textual units that are found in their vicinity. It consists of a database of more than 7600 text fragments that contain marked occurrences of more than 450 cue phrases. To my knowledge, this corpus analysis is the first one that encodes not only linguistic information, but also algorithmic information. The blend of linguistic and algorithmic information enables the derivation of algorithms that determine the elementary textual units in a text and that hypothesize rhetorical relations that hold among these units. In other words, the corpus analysis provides empirical grounding for a procedural account of cue phrases [Caron, 1997], one that treats them as instructions that permit the determination of discourse units and the construction of complex text structures.

8.3.2 Shortcomings and future work

The most important shortcoming of the work described in chapter 4 concerns its degree of completion; I was the only analyst of 2100 of the 7600 text fragments in the corpus. To counterbalance this shortcoming, I did not evaluate any of the algorithms that were derived using the data in the corpus against my own subjective judgments but rather against data that did not occur in the corpus and that was analyzed independently by a relatively large number of judges.

Another shortcoming of the corpus analysis is that it relied on non-objective definitions of the notions of elementary textual unit, rhetorical relation, and rhetorical status. However, providing empirically grounded, objective definitions for all these notions amounts to proposing an objective, empirically grounded taxonomy of rhetorical relations, which is beyond the scope of this thesis.

The degree of completion and the lack of objective definitions are not the only shortcomings of the corpus analysis. In fact, I believe that one can find faults with every field in the database and provide many suggestions for improvement. The most obvious suggestion would be to encode full text structures and not merely the relations that are signalled by a certain cue phrase. Or to encode information concerning the part of speech of the words found in the vicinity of cue phrases and to use that information in order to determine whether a cue phrase has a discourse function or not. Or to encode information about the entities that are in focus and study empirically the relationship between focus operations and cue phrases.

8.4 The rhetorical parsing of natural language texts

8.4.1 Contributions

The discourse marker and clause-like unit identification algorithm

In chapter 5, I have proposed an algorithm that determines the elementary units of text and the cue phrases that have a discourse function. To my knowledge, this is the first algorithm that identifies clause-like unit boundaries on the basis of only surface-form methods. The recall and precision figures have been shown to be in the range of 80% and 90% respectively, in the condition in which the input to the algorithm was unrestricted text and not manually encoded sets of features as in the case of the algorithms proposed by Hirschberg and Litman [1993], Litman [1994, 1996], and Siegel and McKeown [1994].

The derivation of valid text structures in the case of disjunctive hypotheses

Discourse markers are an ambiguous indicator of the rhetorical relations that hold among textual units. In order to deal with this ambiguity, I have extended the proof-theoretic techniques that I discussed in chapter 3 so that they can handle disjunctive relations as well. I have designed sound and complete algorithms that derive the structure of text in the case in which the rhetorical relations that are given as input are sets of disjunctive hypotheses and I have proposed a heuristic for determining the "best" discourse tree in a collection of valid trees.

The rhetorical parsing algorithm

In chapter 5, I have proposed the first surface-form-based rhetorical parsing algorithm; the algorithm takes as input an unrestricted English text, determines its elementary units, hypothesizes rhetorical relations among these units, and derives its valid discourse structures. Although the algorithm relies only on cue phrases and word co-occurrences, the methodology it proposes is general: it can easily accommodate more elaborate syntactic and semantic methods both for determining the elementary units and for hypothesizing the rhetorical relations that hold among them.

8.4.2 Shortcomings and future work

In designing a rhetorical parsing algorithm, I had to make a quite large number of choices: I had to choose between using surface-form, syntactic, and part-of-speech tagging methods; I had to choose between assuming or not that paragraph breaks correlate with the high-level structure of discourse; and I had to choose an evaluation function for determining what discourse trees are the "best". As a consequence, it is obvious that the work presented in chapter 5 investigates only one of the many possible ways in which discourse structures can be built. The thesis does not make any claim that the choices that I made would yield the best results.

Another shortcoming pertains to the evaluation of the rhetorical parser. As I have discussed in section 5.8, an adequate evaluation would assume the existence of a significant number of manually built discourse structures. However, until we develop objective definitions for rhetorical relations, it is quite unlikely that we would be able to achieve significant agreement among the analysts that would build these structures; in addition, the resources that would be needed to build a corpus of discourse trees would be quite significant. And even if we assume that we have a corpus of discourse trees, we still need to find appropriate evaluation metrics, similar to those that were developed to evaluate syntactic trees. This thesis did not investigate any of these issues.

The work described in chapter 5 is open to many improvements. For example, one could investigate the use of machine learning techniques for deriving better discourse marker and clause-like unit identification algorithms; or the use of statistical techniques for hypothesizing more precise rhetorical relations among the textual units. Or one could investigate better heuristics for determining the "best" discourse trees of a text.

The extensions suggested above concern "local" improvements. Without diminishing their importance, I would like to point out that the idea of rhetorical parsing can have a significant impact on natural language research, because it provides the ground for many applications that still await adequate solutions. For example, one could investigate the relationship between syntactic and rhetorical parsing and use the rhetorical parser in order to construct accurate syntactic trees. Or one could use the rhetorical parser in order to develop new anaphora resolution algorithms, which put more emphasis on the structure of discourse than current algorithms do. Or one could use the rhetorical parser in order to investigate ways to derive the structure of arguments, the intentions of the writer, etc. The summarization program in chapter 6 is only one of the many possible applications that has the rhetorical parsing algorithm at its foundation.

8.5 The summarization of natural language texts

8.5.1 Contributions

The psycholinguistic investigation of the relationship between text structures and what readers perceive as being important in a text

In chapter 6, I have presented a psycholinguistic experiment that shows that there exists a strong correlation between the nuclei of a text structure and what readers perceive as being important in the corresponding text. Hence, I have shown that the structure of text can be used effectively for determining the most important units in a text.

The discourse-based summarization algorithm

I have also proposed a discourse-based summarization algorithm. The algorithm takes as input a text and a number $p, 1 \leq p \leq 100$. It uses the rhetorical parser in order to derive the text structure of the text, and on that basis, it selects the most important $p\%$ of the units in the text. The discourse-based summarization algorithm has been shown to significantly outperform both a baseline algorithm and Microsoft's Office97 summarization program.

8.5.2 Shortcomings and future work

The main shortcoming of the discourse-based summarization algorithm concerns the mapping between discourse structures and the importance scores that are assigned to the units in the text given as input. In section 6.5.3, I have suggested that in order to improve the quality of the summaries, one would need to use not only the dichotomy between nuclei and satellites but also the types of rhetorical relations that relate certain textual units. Selecting the most important units in a text may also depend on the audience profile and the purpose for which a summary is created — none of these issues have been yet investigated.

And mapping the selected textual units into coherent text is another issue that deserves full attention.

The summarization program presented in chapter 6 can be used as a stand-alone product or can be embedded in a variety of applications. For example, it seems reasonable to use the summarization algorithm as a front-end for a Web indexing engine: instead of indexing full documents, the engine would index only summaries of documents. Intuitively, such an approach would produce better recall and precision results than current search engines do; however, a proper investigation needs to be carried out in order to establish the validity of this intuition.

8.6 The generation of natural language texts

8.6.1 Contributions

The empirical investigation of the strengths of the ordering and clustering preferences of the nuclei and satellites of rhetorical relations

Mann and Thompson [1988] suggested that some rhetorical relations exhibit strong patterns of ordering of their nuclei and satellites. In chapter 7, I use the data in the corpus in order to determine empirically the strengths of the ordering and clustering preferences of the nuclei and satellites of the rhetorical relations in the corpus.

The bottom-up approach to text planning

The strengths of the ordering and clustering preferences of the nuclei and satellites of rhetorical relations provide the empirical grounding for a bottom-up approach to text planning. The approach is particularly suitable when the main communicative goal is "tell everything that is in this knowledge pool", but it can also handle generation problems that involve multiple high-level communicative goals. The bottom-up approach 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 valid. In chapter 7, I have proposed three algorithms that implement the bottom-up approach to text planning and shown how they can be integrated into HealthDoc, a natural language system that generates texts that are tailored to specific audiences.

8.6.2 Shortcomings and future work

One possible criticism of the bottom-up approach to text generation is that it uses strengths of ordering and clustering preferences that were derived from only about a quarter of the data in the corpus. Indeed, it is possible that the values of the preferences that were derived from the analyzed corpus will change when the corpus study will be completed. Although this is likely to happen, I expect that the algorithms will not need to be modified. In fact, the strengths of the ordering and clustering preferences that were obtained so far (see appendix E) are consistent with Mann and Thompson's intuitions that are reflected by the canonical orderings shown in table 7.3.

A more serious criticism concerns the computational properties of the bottom-up text planning algorithms. It is true that the bottom-up text planning algorithms that I have proposed are only exponential, and not undecidable, as the ubiquitous top-down planning algorithms are, but still if the bottom-up algorithms are to be applied for large scale problems, better solutions will have to be identified.

Another direction for future work concerns the integration of the text planning algorithms into the HealthDoc architecture. The current implementation applies aggregation, reordering, and reference repairs to the text plans generated by the bottom-up algorithm. Ideally, all the modules would perform repairs concurrently, thus also enabling an appropriate signalling with discourse markers of the discourse relations that pertain to a given text plan.

Appendix A

Text examples

Text A.1 [Mann and Thompson, 1988, p. 252].

[The next music day is scheduled for July 21 (Saturday), noon-midnight.[1]] [I'll post more details later[2]] [but this is a good time to reserve the place on your calendar.[3]]

$$\left\{ \begin{array}{l} rhet\_rel(\textsc{concession}, 2, 3), rhet\_rel(\textsc{elaboration}, 2, 1) \\ rhet\_rel(\textsc{justification}, 1, 2), rhet\_rel(\textsc{justification}, 1, 3) \end{array} \right.$$

Text A.2

[No matter how much one wants to stay a non-smoker,[1]] [the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life.[2]] [We know that 3,000 teens start smoking each day,[3]] [although it is a fact that 90% of them once thought that smoking was something that they'd never do.[4]]

$$\left\{ \begin{array}{l} rhet\_rel(\textsc{justification}, 1, 2), rhet\_rel(\textsc{justification}, 4, 2) \\ rhet\_rel(\textsc{evidence}, 3, 2), rhet\_rel(\textsc{concession}, 3, 4) \\ rhet\_rel(\textsc{restatement}, 4, 1) \end{array} \right.$$

Text A.3 [Hirst, 1994]

[A suspected bank robber was in fair condition in hospital last night[1]] [after being hit in the face with a shotgun blast fired by police on a west-end Toronto street.[2]]

[Police said a detective armed with a 12-gauge shotgun fired one shot at the van[3]] [when the man pulled a handgun.[4]] [The pellets went through the van, shattering both van windows.[5]]

$$\left\{\begin{array}{l} rhet\_rel(\textsc{sequence}, 2, 1), rhet\_rel(\textsc{non\_volitional\_result}, 2, 1) \\ rhet\_rel(\textsc{sequence}, 5, 2), rhet\_rel(\textsc{non\_volitional\_result}, 5, 2) \\ rhet\_rel(\textsc{non\_volitional\_result}, 3, 2), rhet\_rel(\textsc{sequence}, 3, 5) \\ rhet\_rel(\textsc{volitional\_result}, 3, 5), rhet\_rel(\textsc{sequence}, 4, 3) \\ rhet\_rel(\textsc{non\_volitional\_result}, 4, 3) \end{array}\right.$$

Text A.4 [Mann and Thompson, 1988, p. 253].

[Farmington police had to help control traffic recently[1]] [when hundreds of people lined up to be among the first applying for jobs at the yet-to-open Marriott Hotel.[2]] [The hotel's help-wanted announcement — for 300 openings — was a rare opportunity for many unemployed.[3]] [The people waiting in line carried a message, a refutation, of claims that the jobless could be employed if only they showed enough moxie.[4]] [Every rule has exceptions,[5]] [but the tragic and too-common tableaux of hundreds or even thousands of people snake-lining up for any task with a paycheck illustrates a lack of jobs,[6]] [not laziness.[7]]

$$\left\{\begin{array}{l} rhet\_rel(\textsc{volitional\_result}, 1, 2), rhet\_rel(\textsc{circumstance}, 3, 2) \\ rhet\_rel(\textsc{background}, 2, 4), rhet\_rel(\textsc{evidence}, 6, 4) \\ rhet\_rel(\textsc{concession}, 5, 6), rhet\_rel(\textsc{antithesis}, 7, 6) \end{array}\right.$$

Text A.5

[No matter how much one wants to stay a non-smoker,[1]] [the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life:[2]] [75% of young adults will pick up a cigarette and let curiosity take over,[3]] [About 30% of will become experimental smokers.[4]] [Of those who will start smoking, about 90% will end up with a pack and a lighter for the rest of their lives.[5]] [We know that 3,000 teens start smoking each day,[6]] [although it is a fact that 90% of them once thought that smoking was something that they'd never do.[7]]

$$\begin{cases} rhet\_rel(\textsc{justification}, 1, 2), rhet\_rel(\textsc{evidence}, 3, 2) \\ rhet\_rel(\textsc{evidence}, 4, 2), rhet\_rel(\textsc{evidence}, 5, 2) \\ rhet\_rel(\textsc{evidence}, 6, 2), rhet\_rel(\textsc{concession}, 3, 7) \\ rhet\_rel(\textsc{concession}, 4, 7), rhet\_rel(\textsc{concession}, 5, 7) \\ rhet\_rel(\textsc{concession}, 6, 7), rhet\_rel(\textsc{restatement}, 7, 2) \end{cases}$$

Text A.6 [Mann and Thompson, 1988, p. 261].

[What if you're having to clean floppy drive heads too often?[1]] [Ask for Syncom diskettes, with burnished Ectype coating and dust-absorbing jacket liners.[2]] [As your floppy drive writes or reads,[3]] [a Syncom diskette is working four ways[4]] [to keep loose particles and dust from causing soft errors, drop-outs.[5]] [Cleaning agents on the burnished surface of the Ectype coating actually remove build-up from the head,[6]] [while lubricating it at the same time.[7]] [A carbon additive drains away static electricity[8]] [before it can attract dust or lint.[9]] [Strong binders hold the signal-carrying oxides tightly within the coating.[10]] [And the non-woven jacket liner[11]] [more than just wiping the surface[12]] [provides thousands of tiny pockets to keep what it collects.[13]] [To see which Syncom diskette will replace the ones you're using now,[14]] [send for our free Flexi-Finder selection guide and the name of the supplier nearest you.[15]]

$$\begin{cases} rhet\_rel(\textsc{purpose}, 5, 4), rhet\_rel(\textsc{circumstance}, 3, 4) \\ rhet\_rel(\textsc{circumstance}, 7, 6), rhet\_rel(\textsc{purpose}, 9, 8) \\ rhet\_rel(\textsc{antithesis}, 12, 11), rhet\_rel(\textsc{joint}, 6, 8) \\ rhet\_rel(\textsc{joint}, 8, 10), rhet\_rel(\textsc{joint}, 10, 11) \\ rhet\_rel(\textsc{elaboration}, 10, 4), rhet\_rel(\textsc{elaboration}, 11, 4) \\ rhet\_rel(\textsc{enablement}, 15, 14), rhet\_rel(\textsc{purpose}, 13, 14) \\ rhet\_rel(\textsc{motivation}, 4, 2), rhet\_rel(\textsc{enablement}, 14, 2) \\ rhet\_rel(\textsc{solutionhood}, 1, 2) \end{cases}$$

Text A.7

[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]]

$$\left\{\begin{array}{l}
rhet_rel(\textsc{elaboration}, 2, 1), rhet_rel(\textsc{joint}, 1, 3)\\
rhet_rel(\textsc{justification}, 1, 4), rhet_rel(\textsc{justification}, 3, 4)\\
rhet_rel(\textsc{joint}, 5, 6), rhet_rel(\textsc{elaboration}, 5, 4)\\
rhet_rel(\textsc{contrast}, 7, 8), rhet_rel(\textsc{non_volitional_result}, 9, 8)\\
rhet_rel(\textsc{elaboration}, 7, 5), rhet_rel(\textsc{concession}, 10, 12)\\
rhet_rel(\textsc{concession}, 11, 12), rhet_rel(\textsc{joint}, 10, 11)\\
rhet_rel(\textsc{example}, 13, 12), rhet_rel(\textsc{example}, 14, 12)\\
rhet_rel(\textsc{joint}, 13, 14), rhet_rel(\textsc{non_volitional_result}, 15, 13)\\
rhet_rel(\textsc{non_volitional_result}, 15, 14), rhet_rel(\textsc{elaboration}, 17, 16)\\
rhet_rel(\textsc{elaboration}, 17, 18), rhet_rel(\textsc{joint}, 16, 18)\\
rhet_rel(\textsc{elaboration}, 12, 4), rhet_rel(\textsc{antithesis}, 16, 4)\\
rhet_rel(\textsc{antithesis}, 18, 4)
\end{array}\right.$$

Text A.8 [Martin, 1992, p. 259].

[Governments were committed to inflation[1]] [because they were themselves part of the system which required it.[2]] [Modern capitalism thrives on expansion and credit[3]] [and without them it shrivels.[4]] [Equally however it requires the right context,[5]] [which is an expanding world economy:[6]] [a national economy is distinct and severable from other national economies in some senses but not all.[7]] [If the total economy of which it is part does not expand,[8]] [then the inflation in the particular economy ceases to be fruitful[9]] [and becomes malignant.[10]] [Furthermore, the more the particular economy flourishes,[11]] [the more dependent is it upon the total economy to which it is directing a part of its product,[12] [and the more dangerous is any pause in its alimentation[13]] [— the easier it is to turn from boom to bust.[14]] [Finally, any government operating within such a system becomes overwhelmingly committed to maintaining it,[15]] [more especially when symptoms of collapse appear[16]] [as they did in the last decade of our period[17]] [when governments felt compelled to help out not only lame ducks but lame eagles too.[18]] [All this was inflationary.[19]]

$$\left\{\begin{array}{l}
rhet_rel(\textsc{evidence}, 2, 1), rhet_rel(\textsc{joint}, 3, 4), rhet_rel(\textsc{evidence}, 3, 1)\\
rhet_rel(\textsc{evidence}, 4, 1), rhet_rel(\textsc{elaboration}, 6, 5), rhet_rel(\textsc{joint}, 11, 13)\\
rhet_rel(\textsc{comparison}, 3, 5), rhet_rel(\textsc{antithesis}, 3, 5), rhet_rel(\textsc{joint}, 3, 5)\\
rhet_rel(\textsc{evidence}, 5, 1), rhet_rel(\textsc{sequence}, 5, 8), rhet_rel(\textsc{condition}, 8, 9)\\
rhet_rel(\textsc{condition}, 8, 10), rhet_rel(\textsc{joint}, 9, 10), rhet_rel(\textsc{condition}, 11, 12)\\
rhet_rel(\textsc{joint}, 8, 11), rhet_rel(\textsc{condition}, 13, 14), rhet_rel(\textsc{elaboration}, 7, 5)\\
rhet_rel(\textsc{joint}, 8, 11), rhet_rel(\textsc{circumstance}, 18, 17), rhet_rel(\textsc{joint}, 11, 15)\\
rhet_rel(\textsc{restatement}, 19, 1), rhet_rel(\textsc{conclusion}, 12, 19)\\
rhet_rel(\textsc{conclusion}, 14, 19), rhet_rel(\textsc{conclusion}, 15, 19)
\end{array}\right.$$

Appendix B

Cue phrases

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
above all	9	12	9	12
accordingly	18	10	10	10
actually	33	117	10	20
add to this	0	1	0	1
additionally	3	2	3	2
admittedly	0	3	0	3
after	211	685	10	20
after a time	2	3	2	3
after all	22	33	10	20
after that	8	15	8	15
after this	7	5	7	5
afterwards	4	6	4	6
again	35	356	8	22
again and again	1	7	1	7
and again	1	28	1	10
never again	1	5	1	5
once again	7	21	7	20
then again	3	3	3	3
all in all	3	3	3	3
all right	14	36	10	20
all the same	0	5	0	5
all this time	0	2	0	2

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
already	14	227	10	20
also	52	841	10	20
also because	1	5	1	5
but also	0	65	0	15
and also	2	54	2	15
not only	13	173	10	20
alternatively	0	2	0	2
analogously	1	1	1	1
although	116	179	10	20
and	349	7823	30	60
and another	3	21	3	20
and then	34	215	10	20
another time	0	1	0	1
anyhow	3	8	3	8
anyway	6	22	6	20
apart from	9	19	9	19
arguably	0	0	0	0
as	354	3476	10	20
as a consequence	2	4	2	4
as a corollary	0	1	0	1
as a logical conclusion	0	0	0	0
as a matter of fact	6	5	6	5
as a result	21	36	10	20
as against	0	7	0	7
as evidence	1	9	1	9
as far as	5	24	5	15
as for	26	17	15	10
as if	8	129	8	20
as it happened	0	1	0	1
as it is	3	25	3	15
as it turned out	2	1	2	1
as long as	10	45	10	20
as soon as	9	31	9	20
as such	2	13	2	13

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
as though	0	67	0	20
as to	4	158	4	20
as we shall	0	5	0	5
as we will	0	1	0	1
as well	0	257	0	20
aside from	9	8	9	8
at any rate	5	5	5	5
at first	19	34	10	20
at last	15	28	10	20
at least	22	239	10	20
at once	7	55	7	20
at that	15	39	10	20
at that moment	4	4	4	4
at that time	8	17	8	17
at the moment	5	15	5	15
at the outset	2	6	2	6
at the same time	27	47	10	20
at this date	1	0	1	0
at this moment	5	6	5	6
at this point	5	14	5	14
at this stage	1	2	1	2
at which	0	37	0	20
back	12	694	10	20
because	62	641	10	20
because of	25	179	10	20
because of this	5	2	5	2
before	58	744	10	20
before long	0	2	0	2
before that	2	5	2	5
before then	1	1	1	1
besides	34	22	10	20
besides that	1	2	1	2
briefly	1	34	1	20
but	1020	2064	10	20

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
but also	0	65	0	20
but then	17	9	10	9
but then again	0	1	0	1
by	140	3068	10	20
by all means	1	2	1	2
by and large	3	2	3	2
by comparison	2	4	2	4
by contrast	2	4	2	4
by that time	1	4	1	4
by the same	2	7	2	7
by the time	16	16	10	16
by the way	1	9	1	9
by then	3	6	3	6
certainly	24	110	10	20
clearly	9	103	9	20
conceivably	1	9	1	9
consequently	9	17	9	17
considering	7	38	7	20
contrariwise	0	0	0	0
conversely	5	2	5	2
correspondingly	0	2	0	2
decidedly	0	3	0	3
definitely	0	20	0	20
despite	36	63	10	20
despite this	2	1	2	1
doubtless	1	12	1	12
each time	5	6	5	6
earlier	3	122	3	20
either	10	235	10	20
either case	0	3	0	3
either event	0	2	0	2
either way	3	2	3	2
else	0	141	0	20
elsewhere	1	28	1	20

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
equally	3	57	3	20
especially	6	147	6	20
essentially	3	41	3	20
even	150	800	10	20
even after	2	10	2	10
even before	6	10	6	10
even if	16	40	10	20
even so	13	6	10	6
even then	4	6	4	6
even though	12	58	10	20
even when	7	24	7	20
eventually	8	41	8	20
ever since	5	14	5	14
every time	3	13	3	13
everywhere	7	29	7	20
evidently	2	22	2	20
except	8	152	8	20
except that	1	19	1	19
except when	1	4	1	4
excuse me	1	0	1	0
finally	52	119	10	20
fine	8	123	8	20
first	96	977	10	20
first of all	10	6	10	6
following	15	185	10	20
for	358	4565	10	20
for example	49	85	10	20
for fear that	0	2	0	2
for instance	12	31	10	20
for one	10	50	10	20
for that	6	40	6	20
for that matter	2	7	2	7
for that reason	2	1	2	1
for this	20	133	10	20

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
for this reason	7	6	7	6
formerly	1	24	1	20
fortunately	13	3	13	3
from now on	2	2	2	2
from then on	2	0	2	0
further	21	177	10	20
furthermore	29	2	10	2
given	4	330	4	20
given that	0	2	0	2
having said	0	1	0	1
hence	27	26	10	20
here	122	456	10	20
heretofore	1	7	1	7
hitherto	0	3	0	3
however	135	292	10	20
however that may be	1	0	1	0
I mean	30	0	20	0
if	547	1058	10	20
if ever	1	3	1	3
if not	2	44	2	20
if only	8	10	8	10
if so	5	6	5	6
in addition	78	33	10	20
in any case	13	10	10	10
in case	1	14	1	14
in comparison	0	9	0	9
in conclusion	2	0	2	0
in consequence	0	4	0	4
in contrast	13	10	10	10
in doing	3	6	3	6
in doing so	2	4	2	4
in fact	48	84	10	20
in general	11	32	10	20
in order to	19	85	10	20

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
in other respects	0	1	0	1
in other words	14	7	10	7
in particular	5	18	5	18
in place of	2	8	2	8
in point of fact	2	2	2	2
in short	12	11	10	11
in so doing	1	1	1	1
in spite of	19	29	10	20
in such a	6	18	6	18
in such an	0	3	0	3
in sum	2	0	2	0
in that	13	114	10	20
in that case	1	2	1	2
in the beginning	0	6	0	6
in the case of	5	23	5	20
in the end	5	16	5	16
in the event	2	8	2	8
in the first place	8	8	8	8
in the hope that	0	1	0	1
in the meantime	4	3	4	3
in the same way	3	10	3	10
in this case	8	21	8	20
in this connection	5	3	5	3
in this respect	4	6	4	6
in this way	8	12	8	12
in truth	0	2	0	2
in turn	3	38	3	20
in which	0	330	0	20
in which case	0	4	0	4
incidentally	5	7	5	7
including	1	157	1	20
incontestably	0	0	0	0
incontroversialy	0	0	0	0
indeed	37	98	10	20

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
indisputably	0	1	0	1
indubitably	0	0	0	0
initially	5	12	5	12
insofar	2	5	2	5
instantly	1	12	1	12
instead	40	112	10	20
instead of	24	90	10	20
it can be concluded that	0	0	0	0
it stands to reason that	0	0	0	0
it follows	5	7	5	7
it is because	2	6	2	6
it is only	4	12	4	12
it may seem that	0	1	0	1
just	97	637	10	20
just as	23	91	10	20
just before	5	13	5	13
just then	2	1	2	1
largely	1	61	1	20
last	35	540	10	20
lastly	2	1	2	1
later	38	290	10	20
lest	0	16	0	16
let us	47	50	10	20
let us assume	2	1	2	1
let us consider	0	2	0	2
like	44	929	10	20
likewise	4	12	4	12
luckily	1	1	1	1
mainly	2	27	2	20
meanwhile	22	12	10	12
merely	1	118	1	20
merely because	0	2	0	2
moreover	54	16	10	16
most likely	0	12	0	12

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
more accurately	0	1	0	1
more importantly	1	0	1	0
more precisely	0	2	0	2
more specifically	0	3	0	3
more to the point	0	0	0	0
much as	3	66	3	20
much later	0	2	0	2
naturally	13	48	10	20
needless	6	3	6	3
neither	32	95	10	20
never again	1	5	1	5
nevertheless	32	26	10	20
next	29	304	10	20
next moment	0	2	0	2
next time	0	9	0	9
no doubt	14	38	10	20
no matter	19	32	10	20
nonetheless	2	6	2	6
nor	34	149	10	20
not	108	2587	10	20
not because	0	13	0	13
not only	13	173	10	20
not that	13	28	10	20
notably	0	15	0	15
notwithstanding	1	3	1	3
now	213	790	10	20
now that	7	29	7	20
obviously	22	79	10	20
of course	75	181	10	20
okay	10	5	10	5
on a different note	0	0	0	0
on account of	0	12	0	12
on another	0	5	0	5
on balance	1	1	1	1

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
on condition	0	1	0	1
on one side	0	8	0	8
on the assumption	0	4	0	4
on the bases	0	1	0	1
on the basis	8	45	8	20
on the contrary	8	6	8	6
on the grounds	0	9	0	9
on the one hand	3	10	3	10
on the other hand	33	20	10	20
on the other side	4	16	4	16
on this basis	2	0	2	0
on top of it	0	0	0	0
on which	0	58	0	20
once	71	337	10	20
once again	7	21	7	20
once more	6	19	6	19
only	85	1297	10	20
only after	0	7	0	7
only because	0	12	0	12
only if	0	13	0	13
only when	6	20	6	20
oops	0	1	0	1
or	51	2404	10	20
or again	0	1	0	1
or else	1	7	1	7
originally	2	20	2	20
other than	1	49	1	20
otherwise	12	57	10	20
overall	0	10	0	10
parenthetically	0	1	0	1
particularly	3	130	3	20
particularly when	1	5	1	5
perhaps	76	188	10	20
plainly	0	17	0	17

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
possibly	3	53	3	20
presently	6	25	6	20
presumably	5	32	5	20
previously	2	50	2	20
provided	2	112	2	20
provided that	0	6	0	6
providing that	0	2	0	2
put another way	0	1	0	1
quite likely	0	3	0	3
rather	16	312	10	20
regardless	6	30	6	20
returning to	2	12	2	12
second	32	287	10	20
secondly	3	1	3	1
seemingly	2	14	2	14
similarly	12	19	10	20
simply	5	153	5	20
simply because	1	7	1	7
simultaneously	5	26	5	20
since	151	388	10	20
so	176	1343	10	20
so far	12	47	10	20
so that	3	211	3	20
some time	4	25	4	20
soon	20	153	10	20
speaking of	6	6	6	6
specifically	1	35	1	20
still	43	597	10	20
subsequently	1	9	1	9
such as	2	180	2	20
such that	0	20	0	20
suddenly	20	113	10	20
summarizing	0	3	0	3
summing up	0	1	0	1

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
suppose	20	58	10	20
suppose that	3	17	3	17
supposedly	0	11	0	11
sure enough	1	2	1	2
surely	7	33	7	20
that	229	4950	10	20
that done	1	1	1	1
that is	51	199	10	20
that is all	0	1	0	1
that is how	1	1	1	1
that is to say	2	4	2	4
that is why	9	2	9	2
that reminds me	0	0	0	0
that way	4	33	4	20
the end	0	165	0	20
the fact is	7	1	7	1
the first time	5	53	5	20
the last time	2	8	2	8
the latter	25	73	10	20
the moment	4	52	4	20
the more	9	92	9	20
the next time	3	3	3	3
the thing is	1	0	1	0
then	276	777	10	20
then again	3	3	3	3
thereafter	3	14	3	14
thereby	0	33	0	20
therefore	39	125	10	20
thereupon	2	3	2	3
third	17	145	10	20
this means	10	7	10	7
this time	20	75	10	20
though	61	326	10	20
thus	152	138	10	20

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
thus far	2	9	2	9
to add	1	23	1	20
to be sure	4	20	4	20
to begin with	3	1	3	1
to clarify	0	6	0	6
to close	0	9	0	9
to comment	0	2	0	2
to conclude	0	5	0	5
to explain	0	29	0	20
to get back	0	7	0	7
to illustrate	1	8	1	8
to interrupt	0	3	0	3
to note	0	13	0	13
to open	0	18	0	18
to repeat	0	8	0	8
to start with	0	1	0	1
to stop	1	32	1	20
to sum up	0	1	0	1
to summarize	2	0	2	0
to the degree that	0	2	0	2
to the extent	9	14	9	14
to this end	2	2	2	2
to wit	0	1	0	1
too	28	611	10	20
true	14	173	10	20
ultimately	4	17	4	17
under the circumstances	3	4	3	4
under these circumstances	0	2	0	2
undeniably	0	1	0	1
undoubtedly	6	17	6	17
unfortunately	16	12	10	12
unless	12	81	12	20
until	25	380	10	20
until then	0	4	0	4

Cue phrase	Number of occurrences in the Brown corpus (Beginning of sentences)	Number of occurrences in the Brown corpus (Middle or end of sentences)	Number of selected occurrences (Beginning of sentences)	Number of selected occurrences (Middle or end of sentences)
unquestionably	2	9	2	9
up to now	1	4	1	4
up to this	2	4	2	4
very likely	1	5	1	5
well	115	616	10	20
what is more	2	1	2	1
whatever	22	80	10	20
when	456	1264	10	20
whenever	13	24	10	20
where	72	672	10	20
whereas	10	26	10	20
whereby	0	19	0	19
wherein	0	5	0	5
whereupon	0	5	0	5
wherever	2	23	2	20
whether	26	205	10	20
whether or not	5	14	5	14
which	18	2322	10	20
which is why	0	1	0	1
which means	0	7	0	7
whichever	0	5	0	5
while	105	462	10	20
who	51	1523	10	20
whoever	8	5	8	5
with regard to	2	11	2	11
with respect to	11	45	10	20
with that	4	32	4	20
with this	16	50	10	20
without	36	453	10	20
yet	125	232	10	20
you know	27	52	10	20
Total	9599	69884	2140	5461

Appendix C

Rhetorical relations used in the corpus analysis

Rhetorical relation	Number of occurrences in the first 2100 text fragments of the corpus
ADDITIVE-EMPHASIS	17
ALTERNATIVE	4
ANTI-SEQUENCE	18
ANTITHESIS	67
ANTITHESIS-SEQUENCE	2
ARGUMENTATION	36
BACKGROUND	70
BROKEN-INTENTION	1
CIRCUMSTANCE	156
COMPARISON	36
CONCESSION	76
CONCLUSION	14
CONCURRENCY	6
CONDITION	41
CONTINUATION	6
CONTRAST	120
COUNTER-EVIDENCE	1
DETAIL	5
DURATION	1

Rhetorical relation	Number of occurrences in the first 2100 text fragments of the corpus
ELABORATION	236
ENABLEMENT	4
EVALUATION	2
EVIDENCE	108
EXAMPLE	2
EXPLANATION	48
FORWARD-REFERENCE	2
FINAL-STEP	1
INTERPRETATION	70
INTRODUCTION	1
JOINT	214
JUSTIFICATION	34
MEANS	2
MOTIVATION	11
NARRATION	2
NON-EVIDENCE	1
NON-EXPLANATION	1
NONVOLITIONAL-CAUSE	74
NONVOLITIONAL-RESULT	14
NONVOLITIONAL-CAUSE-RESULT	5
OR	2
OTHERWISE	20
OUTCOME	2
PARENTHETICAL	60
PROBLEM-SOLUTION (SOLUTIONHOOD)	1
PURPOSE	10
QUESTION-ANSWER	5
REASON	10
REFUTATION	1
RESTATEMENT	9
SEQUENCE	160
SUMMARY	6

Rhetorical relation	Number of occurrences in the first 2100 text fragments of the corpus
TOPIC-SHIFT	25
VOLITIONAL-CAUSE	28
VOLITIONAL-RESULT	2

Appendix D

The texts that were used in the summarization experiment

Text D.1

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

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

Text D.2

[Cars account for half the oil consumed in the U.S., about half the urban pollution and one fourth the greenhouse gases.] [They take a similar toll of resources in other industrial nations and in the cities of the developing world.] [As vehicle use continues to increase in the coming decade,] [the U.S. and other countries will have to address these issues] [or else face unacceptable economic, health-related and political costs.] [It is unlikely that oil prices will remain at their current low level] [or that other nations will accept a large and growing

U.S. contribution to global climatic change.]

[Policymakers and industry have four options:] [reduce vehicle use,] [increase the efficiency and reduce the emissions of conventional gasoline-powered vehicles,] [switch to less noxious fuels,] [or find less polluting propulsion systems.] [The last of these] [— in particular the introduction of vehicles powered by electricity —] [is ultimately the only sustainable option.] [The other alternatives are attractive in theory] [but in practice are either impractical] [or offer only marginal improvements.] [For example, reduced vehicle use could solve congestion woes and a host of social and environmental problems,] [but evidence from around the world suggests that it is very difficult to make people give up their cars to any significant extent.] [In the U.S., mass-transit ridership and carpooling have declined since World War II.] [Even in western Europe,] [with fuel prices averaging more than $1 a liter (about $4 a gallon)] [and with pervasive mass transit and dense populations,] [cars still account for 80 percent of all passenger travel.]

Text D.3

[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.] [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.] [Fraze succeeded] [because he conceived of a simple and economical rivet to hold the tab in place.] [Unlike previous approaches, the rivet was formed from the surface of the can top itself.]

[Since the mid-1960s, the pop top has experienced dozens of refinements.] [Sharp edges that might cut the person who drinks from the can are gone.] [And the tab remains fixed to the top after opening,] [so that park maintenance workers no longer spend hours scouring the grounds to remove the metal scraps.] [The development of the technology, in fact, continues.] [Today one pound of aluminum yields 1,000 tabs,] [a fourfold increase over the amount produced per pound of metal in 1965.] [The simple manufacture of snap, tap and pop may pose a challenge to the ingenuity of the engineering community for years to come.]

Text D.4

[Understanding how training builds the strength and stamina needed for Olympic events requires basic knowledge of how the body produces energy.] [All human motion depends on the use and resynthesis of adenosine triphosphate (ATP),] [a high-energy molecule consisting of a base (adenine), a sugar (ribose) and three phosphate groups.] [The breaking of the bond between two phosphate units releases energy that powers muscle contractions and other cellular reactions.] [Humans have a very limited capacity for storing ATP.] [At a maximum rate of work, the five millimoles of ATP available for each kilogram of muscle is completely depleted in a few seconds.] [To sustain activity, the body has three interrelated metabolic processes] [for continually resupplying the molecule.] [Which one predominates depends on the muscles' power requirements at a given moment and on the duration of the activity.]

[The most immediately available source for reconstructing ATP is phosphocreatine,] [itself a high-energy, phosphate-bearing molecule.] [The energy released by the breakdown of the phosphocreatine molecule is used to resynthesize ATP.] [The phosphocreatine system can recharge ATP for only a short while] [— just five to 10 seconds during a sprint.] [When the supply of this molecule is exhausted,] [the body must rely on two other ATP-generating processes] [— one that does not require oxygen (anaerobic)] [and one that does (aerobic).]

[The anaerobic process,] [also known as glycolysis,] [is usually the first to kick in.] [Cells break down specific carbohydrates] [(glucose or glycogen in muscle)] [to release the energy for resynthesizing ATP.] [Unfortunately for the athlete, the anaerobic metabolism of carbohydrates can yield a buildup of lactic acid,] [which accumulates in the muscles within two minutes.] [Lactic acid and associated hydrogen ions cause burning muscle pain.] [But lactic acid and its metabolite,] [lactate,] [which accumulates in muscle,] [do not always degrade performance.] [Through training, the muscles of elite competitors adapt] [so that they can tolerate the elevated levels of lactate produced during high-intensity exercise.]

[Even so, lactic acid and lactate eventually inhibit muscles from contracting.] [So anaerobic glycolysis can be relied on only for short bursts of exercise.] [It cannot supply the ATP needed for the sustained activity in endurance events.] [That task falls to aerobic metabolism] [— the breakdown of carbohydrate, fat and protein in the presence of oxygen.] [In contrast with anaerobic glycolysis, the aerobic system cannot be switched on quickly.] [At least one to two minutes of hard exercise must pass until the increase in breathing and heart rate ensures delivery of oxygen to a muscle cell.] [During that interval, the athlete depends on a combination of stored ATP, the phosphocreatine system or anaerobic glycolysis to provide energy.] [With the activation of the aerobic processes,] [these other systems function at a lower level.] [In the aerobic phase, for instance, lactic acid and lactate are still produced,] [but they are consumed by less active muscles] [or metabolized in the liver] [and so do not accumulate.]

[Although the aerobic system is highly efficient,] [its ability to supply the muscles with energy reaches an upper threshold.] [If still more ATP is needed,] [the muscles must step up the use of various other energy sources.] [A soccer player in the middle of a 45-minute half, for example, would depend mostly on aerobic metabolism.] [But if he needed to sprint briefly at full speed,] [his body would immediately call on stored ATP] [or ATP reconstituted by the phosphocreatine system] [to supplement the aerobic system.] [Similarly, if this high-intensity sprint continued for five to 15 seconds,] [the player would experience a rapid increase in the rate of anaerobic glycolysis.] [As the play ended, the body would return to its reliance on the aerobic metabolic system,] [while the capacities of the other energy systems regenerated themselves.]

[Coaches must understand the requirements of their sports] [and adjust the intensity or duration of training] [to improve an athlete's aerobic or anaerobic functioning.] [The fundamental principle of training is that sustained activity will result in adaptation of the muscles to ever increasing levels of stress] [— an idea sometimes referred to as the stimulus-response model.] [Over time, training will induce physiological changes,] [which are adapted to the needs of a specific sport.] [The distance runner's training, for example, focuses on enhancing the capabilities of the aerobic system.] [In contrast, a weight lifter would concentrate on strength and power] [instead of the endurance requirements of the distance events.]

Text D.5

[Smart cards are becoming more attractive] [as the price of microcomputing power and storage continues to drop.] [They have two main advantages over magnetic-stripe cards.] [First, they can carry 10 or even 100 times as much information] [— and hold it much more robustly.] [Second, they can execute complex tasks in conjunction with a terminal.] [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.] [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] [without revealing the actual balance or the account number.] [Depending on the importance of the information involved,] [security might rely on a personal identification number] [such as those used with automated teller machines,] [a midrange encipherment system,] [such as the Data Encryption Standard (DES),] [or a highly secure public-key scheme.]

[Smart cards are not a new phenomenon.] [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.] [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.] [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.] [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.]

Appendix E

Ordering and clustering preferences of the nuclei and satellites of the rhetorical relations in the corpus

Rhetorical relation	Strength of the ordering preference (nucleus first) s_o	Average sentence distance between nucleus and satellite avg_s	Average clause distance between nucleus and satellite avg_c	Strength of the clustering preference s_c
ADDITIVE-EMPHASIS	1.00	0.00	0.06	0.94
ALTERNATIVE	1.00	2.50	3.50	0.05
ANTI-SEQUENCE	0.39	0.33	0.28	0.72
ANTITHESIS	0.15	0.76	0.87	0.13
ANTITHESIS-SEQUENCE	0.00	0.00	0.00	1.00
ARGUMENTATION	0.72	0.61	0.50	0.50
BACKGROUND	0.03	1.06	0.76	0.24
BROKEN-INTENTION	0.00	0.00	0.00	1.00
CIRCUMSTANCE	0.28	0.17	0.12	0.88
COMPARISON	0.97	0.25	0.00	1.00

Rhetorical relation	Strength of the ordering preference (nucleus first) s_o	Average sentence distance between nucleus and satellite avg_s	Average clause distance between nucleus and satellite avg_c	Strength of the clustering preference s_c
CONCESSION	0.36	0.11	0.08	0.92
CONCLUSION	0.00	1.86	2.57	0.05
CONCURRENCY	0.67	0.83	0.83	0.17
CONDITION	0.41	0.07	0.02	0.98
CONTINUATION	1.00	4.83	6.33	0.05
CONTRAST	0.98	0.47	0.34	0.66
COUNTER-EVIDENCE	1.00	1.00	0.00	1.00
DETAIL	0.80	0.00	0.00	1.00
DURATION	1.00	0.00	0.00	1.00
ELABORATION	0.97	1.08	0.90	0.10
ENABLEMENT	0.5	0.50	0.50	0.50
EVALUATION	0.00	0.50	0.00	1.00
EVIDENCE	0.80	0.79	0.69	0.31
EXAMPLE	1.00	0.50	0.00	1.00
EXPLANATION	0.67	0.31	0.25	0.75
FORWARD-REFERENCE	0.00	1.00	1.00	0.05
FINAL-STEP	0.00	1.00	0.00	1.00
INTERPRETATION	0.70	0.81	0.39	0.61
INTRODUCTION	0.00	0.00	0.00	1.00
JOINT	0.99	0.73	0.81	0.19
JUSTIFICATION	0.15	0.82	0.53	0.47
MEANS	0.50	0.00	0.00	1.00
MOTIVATION	0.27	0.64	0.36	0.64
NARRATION	1.00	0.50	1.00	0.05
NON-EVIDENCE	1.00	0.00	0.00	1.00
NON-EXPLANATION	1.00	0.00	0.00	1.00
NONVOLITIONAL-CAUSE	0.39	0.51	0.45	0.55

Rhetorical relation	Strength of the ordering preference (nucleus first) s_o	Average sentence distance between nucleus and satellite avg_s	Average clause distance between nucleus and satellite avg_c	Strength of the clustering preference s_c
NONVOLITIONAL-RESULT	0.86	0.71	0.43	0.57
NONVOLITIONAL-CAUSE-RESULT	1.00	0.20	0.20	0.80
OR	1.00	0.00	0.00	1.00
OTHERWISE	1.00	0.55	0.90	0.10
OUTCOME	1.00	1.00	0.00	1.00
PARENTHETICAL	1.00	0.03	0.03	0.97
PROBLEM-SOLUTION (SOLUTIONHOOD)	0.00	1.0	1.0	0.05
PURPOSE	0.70	0.00	0.00	1.00
QUESTION-ANSWER	1.00	1.00	0.00	1.00
REASON	0.80	0.40	0.30	0.70
REFUTATION	0.00	1.00	0.00	1.00
RESTATEMENT	0.89	0.89	0.89	0.11
SEQUENCE	0.98	1.25	1.13	0.05
SUMMARY	0.00	0.83	0.00	1.00
TOPIC-SHIFT	0.88	1.64	0.88	0.12
VOLITIONAL-CAUSE	0.32	0.36	0.39	0.61
VOLITIONAL-RESULT	1.00	1.50	0.50	0.50

Bibliography

[Anscombre and Ducrot, 1983] J.C. Anscombre and O. Ducrot. *L'argumentation dans la langue*. Pierre Mardaga, Bruxelles, 1983.

[Aone *et al.*, 1997] Chinatsu Aone, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen. A scalable summarization system using robust NLP. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 66–73, Madrid, Spain, July 11 1997.

[Aretoulaki, 1996] Maria Aretoulaki. *COSY–MATS: A Hybrid Connectionist-Symbolic Approach to the Pragmatic Analysis of Texts for Their Automatic Summarization*. PhD thesis, Department of Language Engineering, University of Manchester Institute of Science and Technology, March 1996.

[Aretoulaki, 1997] Maria Aretoulaki. COSY–MATS: An intelligent and scalable summarization shell. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 74–81, Madrid, Spain, July 11 1997.

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

[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, New Mexico, June 27-30 1994.

[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.* The MIT Press, 1985.

[Barzilay and Elhadad, 1997] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, July 11 1997.

[Bateman and Rondhuis, 1994] John Bateman and Klaas Jan Rondhuis. Coherence relations: analysis and specification. Technical Report Deliverable R1.1.2:a,b, Esprit Basic Research Project 6665, October 1994. DANDELION Discourse Functions and Discourse Representation: An Empirically and Linguistically Motivated, Interdisciplinary-Oriented Approach to Natural Language Texts.

[Baxendale, 1958] P.B. Baxendale. Machine-made index for technical literature — an experiment. *IBM Journal of Research and Development*, 2:354–361, 1958.

[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. 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, Pennsylvania, 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, October 1995.

[Boguraev and Kennedy, 1997] Branimir Boguraev and Christopher Kennedy. Salience-based content characterisation of text documents. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 2–9, Madrid, Spain, July 11 1997.

[Brachman, 1992] Ron Brachman. Reducing CLASSIC to practice: Knowledge representation theory meets reality. In *Proceedings of the Conference on Knowledge Representation*, pages 247–258, 1992.

[Brew, 1992] Chris Brew. Letting the cat out of the bag: Generation for Shake-and-Bake MT. In *Proceedings of the International Conference on Computational Linguistics, COLING-92*, Nantes, August 23–28 1992.

[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, California, June 1996.

[Brooks and Dansereau, 1983] Larry W. Brooks and Donald F. Dansereau. Effects of structural schema training and text organization on expository prose processing. *Journal of Educational Reading*, 75(6):811–820, 1983.

[Brown and Day, 1983] Ann L. Brown and Jeanne D. Day. Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior*, 22:1–14, 1983.

[Brown *et al.*, 1983] Ann L. Brown, Jeanne D. Day, and R.S. Jones. Development of plans for summarizing texts: the development of expertise. *Child Development*, 54:968–979, 1983.

[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, Massachusetts, July 25-28 1990.

[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, 1993.

[Carbonell and Collins, 1973] J.R. Carbonell and A.M. Collins. Natural semantics in artificial intelligence. In *Proceedings of the Third International Joint Conference on Artificial Intelligence (IJCAI-73)*, pages 344–351, 1973.

[Carcagno and Iordanskaja, 1989] D. Carcagno and L. Iordanskaja. Content determination and text structuring in Gossip. In *Extended Abstracts of the Second European Natural Language Generation Workshop (ENLG-89)*, pages 15–22, University of Edinburgh, April 6-8 1989.

[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. Lawrence Erlbaum Associates, 1997.

[Cawsey, 1990] Alison Cawsey. Generating explanatory discourse. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation*, pages 75–101. Academic Press, New York, 1990.

[Cawsey, 1991] Alison Cawsey. Generating interactive explanations. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI–91)*, volume 1, pages 86–91, July 14–19 1991.

[Chomsky, 1965] Noam Chomsky. *Aspects of the theory of syntax*. The MIT Press, Cambridge, Massachusetts, 1965.

[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] W.G. 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*. PhD 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.

[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.

[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 (ACL/EACL–97)*, pages 88–95, Madrid, Spain, July 7–12 1997.

[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.* Hartcourt Brace, 1994.

[Dale, 1989] Robert Dale. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL–89)*, pages 68–75, University of British Columbia, Vancouver, June 26–29 1989.

[Davis and Putnam, 1960] M. Davis and H. Putnam. A computing procedure for quantification theory. *Journal of the Association for Computing Machinery*, 7(3):201–215, 1960.

[de Souza and Nunes, 1992] Clarisee S. de Souza and Maria V. Nunes. Explanatory text planning in logic based systems. In *Proceedings of the International Conference on Computational Linguistics (COLING-92)*, pages 742–748, Nantes, France, August 23–28 1992.

[de Villiers, 1974] P.A. de Villiers. Imagery and theme in recall of connected discourse. *Journal of Experimental Psychology*, 103:263–268, 1974.

[Deaton and Gernsbacher, 1997] J.A. Deaton and M.A. Gernsbacher. Causal conjunctions and implicit causality cue mapping in sentence comprehension. *Journal of Memory and Language*, 1997.

[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, July 8–12 1985.

[Defrise and Nirenburg, 1990] Christine Defrise and Sergei Nirenburg. Meaning representation and text planning. In *Proceedings of the International Conference on Computational Linguistics (COLING–90)*, volume 2, pages 219–224, Helsinki, 1990.

[DeJong, 1982] G. DeJong. An overview of the FRUMP system. In W.G. Lehnert and M.H. Ringle, editors, *Strategies for Natural Language Processing*, pages 149–172. Lawrence Erlbaum, London, 1982.

[Delin *et al.*, 1994] J. Delin, A. Hartley, C. Paris, D. Scott, and K. Vander Linden. Expressing procedural relationships in multilingual instructions. In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 61–70, Kennebunkport, Maine, June 1994.

[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.

[Di Eugenio, 1992] Barbara Di Eugenio. Understanding natural language instructions: the case of purpose clauses. In *Proceedings 30th Annual Meeting of the Association for Computational Linguistics (ACL–92)*, pages 120–127, 1992.

[Di Eugenio, 1993] Barbara Di Eugenio. *Understanding Natural Language Instructions: A Computational Approach to Purpose Clauses*. PhD 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 (ACL/EACL–97)*, pages 80–87, Madrid, Spain, July 7-12 1997.

[Dillon, 1991] Andrew Dillon. Readers' models of text structures: the case of academic articles. *International Journal of Man-Machine Studies*, 35:913–925, 1991.

[DiMarco and Foster, 1997] Chrysanne DiMarco and Mary Ellen Foster. The automated generation of web documents that are tailored to the individual reader. In *Proceedings of the AAAI–97 Spring Symposium on Natural Language Processing for the World Wide Web*, Stanford, CA, March 1997.

[DiMarco *et al.*, 1997] Chrysanne DiMarco, Graeme Hirst, and Eduard Hovy. Generation by selection and repair as a method for adapting text for the individual reader. In *Proceedings of the Workshop on Flexible Hypertext, Eighth ACM International Hypertext Conference*, Southampton, UK, April 1997.

[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, 1991] Michael Elhadad. FUF User manual — version 5.0. Technical Report CUCS-038-91, Department of Computer Science, Columbia University, 1991.

[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, 1990.

[Endres-Niggemeyer, 1997] Brigitte Endres-Niggemeyer. SimSum: simulation of summarizing. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 89–96, Madrid, Spain, July 11 1997.

[Fillenbaum, 1977] S. Fillenbaum. Mind your p's and q's: The use of content and context in some uses of *and*, *or*, and *if*. In G. Bower, editor, *The psychology of learning and motivation*, volume 11, pages 41–100. Academic Press, New York, 1977.

[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.

[Gaizauskas and Robertson, 1997] Robert Gaizauskas and Alexander M. Robertson. Coupling information retrieval and information extraction: A new text technology for gathering information from the web. In *Proceedings of the 5th RIAO Computer-Assisted Information Searching on Internet*, pages 356–370, Montreal, Canada, June 25–27 1997.

[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, 1992.

[Gardent, 1994] Claire Gardent. Discourse multiple dependencies. Technical Report CLAUS-Report Nr. 45, Universität des Saarlandes, Saarbrücken, October 1994.

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

[Garey and Johnson, 1979] Michael R. Garey and David S. Johnson. *Computers and Intractability. A Guide to the Theory of NP-Completeness.* W.H. Freeman and Company, New York, 1979.

[Garner, 1982] Ruth Garner. Efficient text summarization: costs and benefits. *Journal of Educational Research*, 75:275–279, 1982.

[Georgeff, 1987] M.P. Georgeff. Planning. *Annual Reviews in Computer Science*, 2:359–400, 1987. Also published in *Readings in Planning*, J. Allen, J. Hendler, and A. Tate eds., Morgan Kaufmann Publishers, Inc., 1990, 5–25.

[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. 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. John Benjamins, Philadelphia, 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 in Typological Studies of Language, pages 59–115. John Benjamins, 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, 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.

[Green, 1997] Stephen J. Green. *Automatically generating hypertext by computing semantic similarity.* PhD thesis, Department of Computer Science, University of Toronto, 1997.

[Grice, 1975] H.P. Grice. Logic and conversation. In Cole P. and Morgan J.L., editors, *Syntax and Semantics, Speech Acts*, volume 3, pages 41–58. Academic Press, 1975.

[Grimes, 1975] J.E. Grimes. *The Thread of Discourse.* Mouton, The Hague, Paris, 1975.

[Grosz and Hirschberg, 1992] Barbara Grosz and Julia Hirschberg. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, 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.*, 1995] Brigitte Grote, Nils Lenke, and Manfred Stede. Ma(r)king concessions in english and german. In *Proceedings of the Fifth European Workshop on Natural Language Generation*, pages 11–32, Leiden, The Netherlands, May 20-22 1995.

[Hahn, 1990] Udo Hahn. Topic parsing: Accounting for text macro structures in full-text analysis. *Information Processing and Management*, 26(1):135–170, 1990.

[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 (ACL/EACL–97)*, pages 104–111, Madrid, Spain, July 7–12 1997.

[Halliday, 1994] Michael A.K. Halliday. *An Introduction to Functional Grammar*. Second Edition, Edward Arnold, London, England, 1994.

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

[Harabagiu and Moldovan, 1995] Sanda M. Harabagiu and Dan I. Moldovan. A marker-propagation algorithm for text coherence. In *Working Notes of the Workshop on Parallel Processing in Artificial Intelligence*, pages 76–86, Montreal, Canada, August 1995.

[Harabagiu and Moldovan, 1996] Sanda M. 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, Massachusetts, 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, Las Cruces, New Mexico, 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.

[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. Lawrence Erlbaum Associates, 1997.

[Hirschberg, 1991] Julia B. Hirschberg. *A Theory of Scalar Implicature*. Garland Publishing, Inc., 1991.

[Hirschberg and Litman, 1987] Julia B. Hirschberg and Diane 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, 1987.

[Hirschberg and Litman, 1993] Julia Hirschberg and Diane 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, California, June 24–27 1996.

[Hirst, 1994] Graeme Hirst. Introduction to computational linguistics. Lecture Notes, 2501F, 1994.

[Hirst and St-Onge, 1997] Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database and some of its applications*. The MIT Press, Cambridge, MA, 1997.

[Hirst *et al.*, 1993] G. Hirst, S. McRoy, P. Heeman, P. Edmonds, and D. 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.

[Hirst *et al.*, 1997] Graeme Hirst, Eduard Hovy, Chrysanne DiMarco, and Kimberley Parsons. Authoring and generating health-education documents that are tailored to the needs of the individual patient. In *Proceedings of the Sixth International Conference on User Modeling*, Sardinia, Italy, June 1997.

[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, 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)*, 1995.

[Hobbs, 1990] Jerry R. Hobbs. *Literature and Cognition*. CSLI Lecture Notes Number 21, 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, 1995] Jerry R. Hobbs. Why is discourse coherent? In M.A. Gernsbacher and T. Givón, editors, *Coherence in spontaneous text*, volume 31 of *Typological Studies in Language*, pages 29–70. John Benjamins Publishing Company, 1995.

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

[Holmes and Gallagher, 1917] H.W. Holmes and O. Gallagher. *Composition and Rhetoric*. D. Appleton and Co., New York, 1917.

[Horacek, 1992] Helmut Horacek. An integrated view of text planning. In *Aspects of Automated Natural Language Generation; 6th International Workshop on Natural Language Generation*, number 587 in Lecture Notes in Artificial Intelligence, pages 29–44. Springer-Verlag, Trento, Italy, April 1992.

[Hovy, 1988a] Eduard H. Hovy. *Generating Natural Language Under Pragmatic Constraints*. Lawrence Erlbaum Associates, 1988.

[Hovy, 1988b] 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, June 27–30 1988.

[Hovy, 1988c] Eduard H. Hovy. Two types of planning in language generation. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (ACL–88)*, pages 179–186, State University of New York at Buffalo, June 27-30 1988.

[Hovy, 1990a] Eduard H. Hovy. Pragmatics and natural language generation. *Artificial Intelligence*, 43:153–197, 1990.

[Hovy, 1990b] 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. Academic Press, New York, 1990.

[Hovy, 1993] Eduard H. Hovy. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63(1–2):341–386, October 1993.

[Hovy and Arens, 1991] Eduard H. Hovy and Yigal Arens. Automatic generation of formatted text. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI–91)*, volume 1, pages 92–97, July, 14–19 1991.

[Hovy and Lin, 1997] Eduard Hovy and Chin Yew Lin. Automated text summarization in SUMMARIST. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 18–24, Madrid, Spain, July 11 1997.

[Hovy and Maier, 1997] Eduard H. Hovy and Elisabeth Maier. Parsimonious or profligate: How many and which discourse structure relations? In *Discourse Processes*. 1997. To appear.

[Hovy and Wanner, 1996] Eduard H. Hovy and Leo Wanner. Managing sentence planning requirements. In *Proceedings of the ECAI–96 Workshop, Gaps and Bridges: New Directions in Planning and Natural Language Generation*, pages 53–38, Budapest, Hungary, August 1996.

[Hovy *et al.*, 1992] Eduard Hovy, Julia Lavid, Elisabeth Maier, Vibhu Mittal, and Cécile Paris. Employing knowledge resources in a new text planner architecture. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation; 6th International Workshop on Natural Language Generation*, number 587 in Lecture Notes in Artificial Intelligence, pages 56–72, Trento, Italy, April 1992. Springer-Verlag.

[Hovy *et al.*, 1998] Eduard H. Hovy, Leo Wanner, and Daniel Marcu. Microplanning in text generation. In preparation, 1998.

[Huang, 1994] Xiaorong Huang. Planning argumentative texts. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING–94)*, volume 1, pages 329–334, Kyoto, Japan, August 1994.

[Huang and Fiedler, 1997] Xiaorong Huang and B. Fiedler. Proof verbalization as an application of NLG. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI–97)*, Nagoya, Japan, August 23-29 1997.

[Jang and Myaeng, 1997] Dong-Hyun Jang and Sung Hyun Myaeng. Development of a document summarization system for effective information services. In *Proceedings of the 5th RIAO Computer-Assisted Information Searching on Internet*, pages 101–111, Montreal, Canada, June 25–27 1997.

[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 Joshi. An introduction to tree adjoining grammar. In Alexis Manaster-Ramer, editor, *Mathematics of Language*. John Benjamins, 1987.

[Justeson and Katz, 1995] J.S. Justeson and S.M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27, 1995.

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

[Kamp, 1981] Hans Kamp. A theory of truth and semantic interpretation. In J.A.G. Groenendijk, T.M.V. Janssen, and M.B.J. Stokhof, editors, *Formal Methods in the Study of Language*, Mathematical Centre Tracts 135, pages 277–322. Mathematisch Centrum, Amsterdam, 1981.

[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.* Kluwer Academic Publishers, London, Boston, Dordrecht, 1993. Studies in Linguistics and Philosophy, Volume 42.

[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.* Springer Verlag, Berlin, 1979.

[Kasper, 1989] Robert Kasper. Sentence planning language. Unpublished technical document, USC Information Sciences Institute, 1989.

[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, Oregon, August 4–8 1996.

[Keller, 1992] Bill Keller. A logic for representing grammatical knowledge. In *Proceedings of the Tenth European Conference on Artificial Intelligence (ECAI-92)*, 1992.

[Keller, 1993] Bill Keller. *Feature Logics, Infinitary Descriptions and Grammar.* Number 44. CSLI Lecture Notes, Center for the Study of Language and Information, 1993.

[Kintsch, 1977] Walter Kintsch. On comprehending stories. In Marcel Just and Patricia Carpenter, editors, *Cognitive processes in comprehension.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

[Kintsch, 1982] Walter Kintsch. Text representation. In Wayne Otto and Sandra White, editors, *Reading Expository Material*, pages 87–101. Academic Press, New York, 1982.

[Kintsch and van Dijk, 1978] Walter Kintsch and Teun A. van Dijk. Toward a model of text comprehension and production. *Psychological Review*, 85:363–394, 1978.

[Knoblock, 1992] Craig A. Knoblock. An analysis of ABSTRIPS. In J. Hendler, editor, *Proceedings of the First International Conference on Artificial Intelligence Planning Systems*, pages 126–135, College Park, Maryland, June 15–17 1992.

[Knott, 1995] Alistair Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations.* PhD thesis, University of Edinburgh, 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. 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, 1996.

[Krippendorff, 1980] Klaus Krippendorff. *Content analysis: An Introduction to its Methodology.* Sage Publications, Beverly Hills, CA, 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, Washington, 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.

[Langleben, 1983] Maria Langleben. An approach to the microcoherence of text. In Fritz Neubauer, editor, *Coherence in natural-language texts*, volume 38 of *Papers in textlinguistics*, pages 71–98. Helmut Buske Verlag, Hamburg, 1983.

[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, 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*, number 587 in Lecture Notes in Artificial Intelligence, pages 167–182, Trento, Italy, April 1992. Springer-Verlag.

[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, 1992.

[Lehmam, 1997] Abderrafih Lehmam. Automatic summarization on the WEB? A system for summarizing using indicating fragments: RAFI. In *Proceedings of the 5th RIAO Computer-Assisted Information Searching on Internet*, pages 112–122, Montreal, Canada, June 25–27 1997.

[Leong *et al.*, 1997] H. Leong, S. Kapur, and O. de Vel. Text summarization for knowledge filtering agents in distributed heterogeneous environments. In *Working Notes of the AAAI–97 Spring Symposium on Natural Language Processing Tools for the World-Wide-Web*, pages 87–94, Stanford, CA, March 1997.

[Liddy, 1991] Elizabeth DuRoss Liddy. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing and Management*, 27(1):55–81, 1991.

[Liddy, 1993] Elizabeth DuRoss Liddy. Development and implementation of a discourse model for newspaper texts. In *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication*, Dagstuhl, Germany, December 13–17 1993.

[Lin, 1995] Chin-Yew Lin. Knowledge-based automatic topic identification. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL–95)*, pages 308–310, Cambridge, Massachusetts, June 26–30 1995.

[Lin and Hovy, 1997] Chin-Yew Lin and Eduard Hovy. Identifying topics by position. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP–97)*, pages 283–290, Washington, DC, March 31 – April 3 1997.

[Litman, 1994] Diane J. Litman. Classifying cue phrases in text and speech using machine learning. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI–94)*, volume 1, pages 806–813, Seattle, July 31 – August 4 1994.

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

[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.

[Lloyd, 1987] John Wylie Lloyd. *Foundations of Logic Programming.* Springer Verlag, Second edition, 1987.

[Longacre, 1979] Robert E. Longacre. The paragraph as a grammatical unit. In Talmy Givón, editor, *Syntax and semantics*, volume 12, pages 115–134. Seminar Press, New York, 1979.

[Longacre, 1983] Robert E. Longacre. *The Grammar of Discourse.* Plenum Press, New York, 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, 1957] H.P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, October 1957.

[Luhn, 1958] H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, April 1958.

[MacGregor and Bates, 1987] R.M. MacGregor and R. Bates. The LOOM knowledge representation language. Technical Report RS-87-188, Information Sciences Institute, 1987.

[Mackworth, 1977] Alan K. Mackworth. Consistency in networks of relations. *Artificial Intelligence*, 8:99–118, 1977.

[Maier, 1993] Elisabeth A. 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.

[Manesh, 1997] Kavi Manesh. Hypertext summary extraction for fast document browsing. In *Working Notes of the AAAI–97 Spring Symposium on Natural Language Processing Tools for the World-Wide-Web*, pages 95–103, Stanford, CA, March 1997.

[Mani and Bloedorn, 1997a] Inderjeet Mani and Eric Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 622–628, Providence, Rhode Island, July 27–31 1997.

[Mani and Bloedorn, 1997b] Inderjeet Mani and Eric Bloedorn. Summarizing similarities and differences among related documents. In *Proceedings of the 5th RIAO Computer-Assisted Information Searching on Internet*, pages 373–387, Montreal, Canada, June 25–27 1997.

[Mann, 1984] William C. Mann. Discourse structures for text generation. In *Proceedings of the 22nd Annual Meeting of the Association for Computational Linguistics (ACL–84)*, 1984. Also available as ISI Report RR-84-127.

[Mann and Moore, 1981] Willliam C. Mann and James A. Moore. Computer generation of multiparagraph english text. *American Journal of Computational Linguistics*, 7(1):17–29, January-March 1981.

[Mann and Thompson, 1987] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, Information Sciences Intitute, 4676 Admiralty Way, Marina del Rey, California 90290-6685, June 1987.

[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, 1996] Daniel Marcu. The conceptual and linguistic facets of persuasive arguments. In *Proceedings of the ECAI–96 Workshop, Gaps and Bridges: New Directions in Planning and Natural Language Generation*, pages 43–46, Budapest, Hungary, August 12th 1996.

[Marcu, 1997] Daniel Marcu. Perlocutions: The Achilles' Heel of Speech Act Theory. In *The Proceedings of the AAAI-97 Fall Symposium on Communicative Action in Humans and Machines*, MIT, Cambridge, Massachussetts, November 8–10 1997.

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

[Maybury, 1993] Mark T. Maybury. Communicative acts for generating natural language arguments. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93)*, pages 357–364, Washington, DC, July 11–15 1993.

[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. Kluwer Academic Publishers, 1991.

[McGuire, 1968] William J. McGuire. The nature of attitudes and attitude change. In G. Lindzey and E. Aronson, editors, *The Handbook of Social Psychology*, volume 3, pages 136–314. Addison-Wesley, second edition, 1968.

[McKeown, 1985] Kathleen R. McKeown. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Cambridge University Press, 1985.

[McKeown and Radev, 1995] Kathleen R. McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. In *Proceedings of the Seventeenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, Seattle, Washington, 1995.

[McKeown *et al.*, 1990] Kathleen R. McKeown, Michael Elhadad, Yumiko Fukumoto, Jong Lim, Christine Lombardi, Jacques Robin, and Frank A. Smadja. Natural language generation in COMET. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current*

Research in Natural Language Generation, pages 103–139. Academic Press, New York, 1990.

[McRoy, 1993] Susan W. McRoy. *Abductive Interpretation and Reinterpretation of Natural Language Utterances*. PhD thesis, Department of Computer Science, University of Toronto, 1993.

[Meehan, 1977] James R. Meehan. TALE-SPIN, an interactive program that writes stories. In *Proceedings of the Fifth International Conference on Artificial Intelligence (IJCAI–77)*, 1977.

[Mellish, 1988] Chris Mellish. Natural language generation from plans. In Michael Zock and Gérard Sabah, editors, *Advances in Natural Language Generation: An Interdisciplinary Perspective*, volume 1, chapter 7, pages 131–145. Ablex Publishing Corporation, Norwood, NJ, 1988. Also appears as CSRP Tech Report 031, University of Sussex.

[Mel'čuk and Polguère, 1987] Igor Mel'čuk and Alan Polguère. A formal lexicon in the meaning-text theory (or how to do lexica with words). *Computational Linguistics*, 13(3-4):276–289, 1987.

[Meteer, 1991a] Marie W. Meteer. Bridging the generation gap betweenttext planning and linguistic realization. *Computational Intelligence*, 7:296–304, 1991.

[Meteer, 1991b] Marie W. Meteer. The implications of revisions for natural language generation. In Cécile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 155–177. Kluwer Academic Publisher, 1991.

[Meteer, 1992] Marie W. Meteer. *Expressibility and the Problem of Efficient Text Planning*. 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.

[Mitra *et al.*, 1997] Mandar Mitra, Amit Singhal, and Chris Buckley. Automatic text summarization by paragraph extraction. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 39–46, Madrid, Spain, July 11 1997.

[Mittal, 1993] Vibhu O. Mittal. *Generating Natural Language Descriptions with Integrated Text and Examples*. PhD thesis, USC/Information Sciences Institute, September 1993.

[Mittal and Paris, 1993] Vibhu O. Mittal and Cécile L. Paris. Generating natural language descriptions with examples: Differences between introductory and advanced texts. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI–93)*, pages 271–276, Washington, DC, July 11–15 1993.

[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, The Netherlands, 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, 1990.

[Moore, 1995] Johanna D. Moore. *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context.* The MIT Press, 1995. ACL-MIT Press series in natural language processing.

[Moore and Paris, 1989] Johanna D. Moore and Cécile L. Paris. Planning text for advisory dialogues. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL–89)*, pages 203–211, University of British Columbia, Vancouver, June 26-29 1989.

[Moore and Paris, 1993] Johanna D. Moore and Cécile 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, 1989] Johanna D. Moore and William R. Swartout. A reactive approach to explanation. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI–89)*, volume 2, pages 1504–1510, Detroit, MI, August 20-25 1989.

[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. Kluwer Academic Publisher, 1991.

[Morris, 1988] Jane Morris. Lexical cohesion, the thesaurus, and the structure of text. Master's thesis, Department of Computer Science, University of Toronto, 1988. Also published as Technical Report CSRI-219.

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

[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, June 1994.

[Nirenburg *et al.*, 1989] Sergei Nirenburg, Victor Lesser, and Eric Nyberg. Controlling a language generation planner. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI–89)*, volume 2, pages 1524–1530, Detroit, MI, August 20-25 1989.

[Noordman and Vonk, 1997] Leo G.M. Noordman and Wietske Vonk. Toward a procedural approacch 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. Lawrence Erlbaum Associates, 1997.

[Norvig, 1992] Peter Norvig. *Paradigms of Artificial Intelligence Programming: Case Studies in Common Lisp*. Morgan Kaufmann Publishers, 1992.

[Nunberg, 1990] G. Nunberg. *The linguistics of punctuation*. CSLI Lecture Notes 18, Stanford, CA. University of Chicago Press, 1990.

[Oberlander and Knott, 1996] Jon Oberlander and Alistair Knott. Issues in cue phrase implicature. In *Working Notes of the AAAI Spring Symposium on Computational Implicature: Computational Approaches to Interpreting and Generating Conversational Implicature*, pages 78–85, Stanford, March 25–27 1996.

[Ochitani *et al.*, 1997] Ryo Ochitani, Yoshio Nakao, and Fumihito Nishino. Goal-directed approach for text summarization. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 47–50, Madrid, Spain, July 11 1997.

[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, Japan, 1994.

[Paice, 1981] Chris D. Paice. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, and P.W. Williams, editors, *Information Retrieval Research*, pages 172–191. Butterworths, 1981.

[Paice, 1990] Chris D. Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26(1):171–186, 1990.

[Paice and Jones, 1993] Chris D. Paice and P.A. Jones. The identification of important concepts in highly structured technical papers. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 69–78, June 1993.

[Paley, 1981] V. Paley. *Wally's Stories*. Harvard University Press, Cambridge, Massachusetts, 1981.

[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.

[Paris, 1991] Cécile L. Paris. Generation and explanation: Building an explanation facility for the explainable expert systems framework. In Cécile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 49–82. Kluwer Academic Publishers, 1991.

[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, California, June 1996.

[Passonneau, 1997a] Rebecca J. Passonneau. Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In Ellen Prince, Aravind Joshi, and Marilyn Walker, editors, *Proceedings of the Workshop on Centering Theory in Naturally Occuring Discourse*. Oxford University Press, 1997. To appear.

[Passonneau, 1997b] Rebecca J. Passonneau. Using centering to relax informationnal constraints on discourse anaphoric noun phrases. *Language and Speech*, 39(1-2), 1997. Special Issue devoted to Discourse, Syntax, and Information. To appear.

[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, Ohio, June 22-26 1993.

[Passonneau and Litman, 1997a] Rebbeca J. Passonneau and Diane J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–140, March 1997.

[Passonneau and Litman, 1997b] Rebecca J. Passonneau and Diane J. Litman. Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence, and linguistic devices. In Eduard Hovy and Donia Scott, editors, *Interdisciplinary Perspectives on Discourse*. Springer Verlag, 1997. To appear.

[Patel-Schneider *et al.*, 1991] P. Patel-Schneider, D. McGuinness, R. Brachman, R. Resnik, and A. Borgida. The CLASSIC knowledge representation system: Guiding principles and implementation rationale. *SIGART Bulletin*, 2(3):108–113, 1991.

[Penman Project, 1989] Penman Project. Penman Documentation: The Primer, The User Guide, The Reference Manual, The Nigel Manual. Technical report, Information Sciences Institute, November 1989.

[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, 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, 1996.

[Preston and Williams, 1994] Keith R. Preston and Sandra H. Williams. Managing the information overload. *Physics in Business*, June 1994.

[Prince, 1978] Ellen F. Prince. A comparison of *it*-clefts and *wh*-clefts in discourse. *Language*, 54:883–906, 1978.

[Quirk *et al.*, 1985] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Usage.* Longman, 1985.

[Rau *et al.*, 1989] Lisa F. Rau, Paul S. Jacobs, and Uri Zernick. Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing and Management*, 25(4):419–428, 1989.

[Rau and Brandow, 1993] Lisa F. Rau and Ron Brandow. Domain-independent summarization of news. In *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication*, Dagstuhl, Germany, December 13–17 1993.

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

[Reed and Long, 1997a] Chris Reed and Derek Long. Content ordering in the generation of persuasive discourse. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI–97)*, pages 1022–1027, Nagoya, Japan, August 23–29 1997.

[Reed and Long, 1997b] Chris Reed and Derek Long. Ordering and focusing in an architecture for persuasive discourse planning. In *Proceedings of the Sixth European Workshop on Natural Language Generation (ENLG–97)*, Duisburg, Germany, March 23–26 1997.

[Reimer and Hahn, 1997] Ulrich Reimer and Udo Hahn. A formal model of text summarization based on condensation operators of a terminological logic. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 97–104, Madrid, Spain, July 11 1997.

[Reiter, 1994] Ehud Reiter. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on*

Natural Language Generation (INLG-94), pages 163–170, Kennebunkport, Maine, June 1994.

[Richmond *et al.*, 1997] Korin Richmond, Andrew Smith, and Einat Amitay. Detecting subject boundaries within text: A language independent statistical approach. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, pages 47–54, Providence, Rhode Island, August 1–2 1997.

[Riloff, 1993] Ellen Riloff. A corpus-based approach to domain-specific text summarization: A proposal. In *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication*, Dagstuhl, Germany, December 13–17 1993.

[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*. PhD 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*, number 587 in Lecture Notes in Artificial Intelligence, pages 199–214, Trento, Italy, April 1992. Springer-Verlag.

[Rubinoff, 1992] Robert Rubinoff. Integrating text planning and linguistic choice by annotating linguistic structures. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation; 6th International Workshop on Natural Language Generation*, number 587 in Lecture Notes in Artificial Intelligence, pages 29–44, Trento, Italy, April 1992. Springer-Verlag.

[Rumelhart, 1972] D.E. Rumelhart. Notes on a schema for stories. In D.G. Bobrow and A. Collins, editors, *Representation and Understanding*. Academic Press, New York, 1972.

[Rumelhart, 1977] D.E. Rumelhart. Understanding and summarizing brief stories. In D. LaBerge and S.J. Samuels, editors, *Basic Processes in Reading: Perception and Understanding*, pages 265–303. Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

[Rush *et al.*, 1971] J.E. Rush, R. Salvador, and A. Zamora. Automatic abstracting and indexing. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of American Society for Information Sciences*, 22(4):260–274, 1971.

[Russell and Norvig, 1995] Stuart Russell and Peter Norvig. *Artificial Intelligence. A Modern Approach.* Prentice Hall, Englewood Cliffs, New Jersey, 1995.

[Sacerdoti, 1974] Earl D. Sacerdoti. Planning in a hierarchy of abstraction spaces. *Artificial Intelligence*, 5:115–135, 1974. Also published in *Readings in Planning*, J. Allen, J. Hendler, and A. Tate eds., Morgan Kaufmann Publishers, Inc., 1990, 98–108.

[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 and Singhal, 1996] Gerard Salton and Amit Singhal. Automatic text decomposition and structuring. *Information Processing and Management*, 32(2):127–138, 1996.

[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 *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.

[Sanford and Garrod, 1981] Anthony J. Sanford and S.C. Garrod. *Understanding written language: explorations of comprehension beyond the sentence.* Wiley, New York, 1981.

[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, California, 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, 1988.

[Schank and Abelson, 1977] Roger C. Schank and Robert P. Abelson. *Scripts, plans, goals, and understanding.* Lawrence Erlbaum Associates, 1977.

[Schiffrin, 1987] Deborah Schiffrin. *Discourse Markers.* 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. Lawrence Erlbaum Associates, 1997.

[Schwarz, 1990] C. Schwarz. Content based text handling. *Information Processing and Management*, 26(2):219–226, 1990.

[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. Academic Press, New York, 1990.

[Sedgewick and Flajolet, 1996] Robert Sedgewick and Philippe Flajolet. *An Introduction to the Analysis of Algorithms.* Addison-Wesley Publishing Company, 1996.

[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. 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, California, 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, Washington, 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, California, 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. MIT Press, Cambridge, Massachusetts, 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, July 31 – August 4 1994.

[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, 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.

[Sjöstrom and Chou Hare, 1984] Colleen L. Sjöstrom and Victoria Chou Hare. Teaching high school students to identify main ideas in expository text. *Journal of Educational Research*, 78:114–118, 1984.

[Skorochodko, 1971] E.F. Skorochodko. Adaptive method of automatic abstracting and indexing. In *Information Processing*, volume 2, pages 1179–1182. North-Holland Publishing Company, 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.

[St-Onge, 1995] David St-Onge. Detecting and correcting malapropisms with lexical chains. Master's thesis, Department of Computer Science, University of Toronto, 1995. Also published as Technical Report CSRI-319.

[Stark, 1988] H.A. Stark. What do paragraph markings do? *Discourse processes*, 11:275–303, 1988.

[Stiff, 1994] James B. Stiff. *Persuasive Communication.* The Guilford Press, 1994.

[Sumita *et al.*, 1992] K. Sumita, K. Ono, T. Chino, T. Ukita, and S. Amano. A discourse structure analyzer for Japanese text. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, volume 2, pages 1133–1140, 1992.

[Szpakowicz *et al.*, 1997] Stan Szpakowicz, K. Barker, T. Copeck, J.F. Delannoy, and S. Matwin. Preliminary validation of a text summarization algorithm. Technical report, University of Ottawa, 1997.

[Talmy, 1983] L. Talmy. How language structures space. In H. Pick and L. Acredolo, editors, *Spatial Orientation: Theory, Research, and Application.* Plenum Press, New York, 1983.

[Teufel and Moens, 1997] Simone Teufel and Marc Moens. Sentence extraction as a classification task. In *Proceedings of the ACL'97/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.* Kluwer Academic Publishers, 1985.

[Toulmin *et al.*, 1979] Stephen Toulmin, Richard Rieke, and Allan Janik. *An Introduction to Reasoning.* Macmillan Publishing Co., Inc., 1979.

[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, 1996.

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

[van Dijk, 1979] Teun A. van Dijk. Pragmatic connectives. *Journal of Pragmatics*, 3:447–456, 1979.

[van Dijk, 1980] Teun A. van Dijk. *Macrostructures. An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1980.

[van Dijk and Kintsch, 1977] Teun A. van Dijk and Walter Kintsch. Cognitive psychology and discourse: Recalling and summarizing stories. In W.U. Dressler, editor, *Trends in textlinguistics.* New York: De Gruyter, 1977.

[Vander Linden, 1993] Keith Vander Linden. *Speaking of Actions: Choosing Rethorical Status and Grammatical Form in Instructional Text Generation.* PhD thesis, University of Colorado at Boulder, July 1993. Also published as Technical Report CU-CS-654-93, Department of Computer Science, University of Colorado at Boulder.

[Vander Linden, 1994] Keith Vander Linden. Generating precondition expressions in instructional text. In *Proceedings 32nd Annual Meeting of the Association for Computational Linguistic (ACL–94)*, pages 42–49, Las Cruces, New Mexico, USA, 27-30 June 1994.

[Vander Linden and Martin, 1995] Keith Vander Linden and J.H. Martin. Expressing rhetorical relations in instructional text: A case study of the purpose relation. *Computational Linguistics*, 21(1):29–58, March 1995.

[Vander Linden *et al.*, 1992] Keith Vander Linden, Susanna Cumming, and James Martin. Using system networks to build rhetorical structures. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation; 6th International Workshop on Natural Language Generation*, number 587 in Lecture Notes in Artificial Intelligence, pages 183–198, Trento, Italy, April 1992. Springer-Verlag.

[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] Marylin A. Walker. Limited attention and discourse structure. *Computational Linguistics*, 22(2):255–264, 1996.

[Walker, 1997] Marylin A. Walker. Centering, anaphora resolution, and discourse structure. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering in Discourse*. Oxford University Press, 1997. To appear.

[Wanner, 1994] Leo Wanner. Building another bridge over the generation gap. In *Proceedings of the Seventh International Workshop on Natural Language Generation (INLG–94)*, pages 137–144, Kennebunkport, Maine, June 1994.

[Wanner and Hovy, 1996] Leo Wanner and Eduard Hovy. The HealthDoc sentence planner. In *Proceedings of the Eighth International Natural Language Generation Workshop (INLG–96)*, pages 1–10, Herstmonceux, UK, June 12–15 1996.

[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, 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.

[Weiner, 1980] J.L. Weiner. BLAH: A system which explains its reasoning. *Artificial Intelligence*, 15:19–48, 1980.

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

[Wing and Scholnick, 1981] C.S. Wing and E.K. Scholnick. Children's comprehension of pragmatic concepts expressed in *because*, *although*, *if*, and *unless*. *Journal of Child Language*, 8:347–365, 1981.

[Winograd, 1984] Peter N. Winograd. Strategic difficulties in summarizing texts. *Reading Research Quaterly*, 19(4):404–425, Summer 1984.

[Yaari, 1997] Yaakov Yaari. Segmentation of expository texts by hierarchical agglomerative clustering. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP–97)*, Bulgaria, 1997.

[Young *et al.*, 1994] R. Michael Young, Johanna D. Moore, and Martha E. Pollack. Towards a principled representation of discourse plans. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, Atlanta, August 1994.

[Young and Moore, 1994] R. Michael Young and Johanna D. Moore. DPOCL: A principled approach to discourse planning. In *Proceedings of the Seventh International Workshop on Natural Language Generation (INLG–94)*, pages 13–20, Kennebunkport, Maine, June 1994.

[Younger, 1967] D.H. Younger. Recognition of context-free languages in time n^3. *Information and Control*, 10:189–208, 1967.

[Zadrozny and Jensen, 1991] Wlodek Zadrozny and Karen Jensen. Semantics of paragraphs. *Computational Linguistics*, 17(2):171–210, June 1991.

[Zock, 1985] Michael Zock. *Le Fil D'Ariane ou Les grammaires de texte comme guide dans l'organisation et l'expression de la pensée en langue maternelle et/ou étrangère.*. Rapport pour L'Unesco, Section Education, Juin 1985.

[Zukerman and McConachy, 1993] Ingrid Zukerman and Richard McConachy. An optimizing method for structuring inferentially linked discourse. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93)*, pages 202–207, Washington, DC, July 1993.

[Zukerman and Pearl, 1986] Ingrid Zukerman and Judea Pearl. Comprehension-driven generation of meta-technical utterances in math tutoring. In *Proceedings of the Fifth National*

Conference on Artificial Intelligence (AAAI-86), pages 606–611, Philadelphia, PA, August 1986.