A computational model of lexical cohesion analysis and its application to the evaluation of text coherence

by

Marzena H. Makuta

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 1997

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

In this thesis, we discuss how to apply the analysis of lexical cohesion in texts to the problem of evaluating text coherence. We have two objectives. The first one is to create a computational model to represent the lexical cohesion of a given text. In order to store this information we design a new data structure — the lexical graph — with lexical items as nodes and lexical relations between those items, such as synonymy, represented as arcs. This structure is particularly suitable for short texts. For longer texts, we propose a different but related data structure, the collapsed lexical graph, with paragraphs as nodes and lexical bonds as arcs.

Next, we show how to apply our model for the representation of cohesion to the problem of evaluating text coherence, for texts of arbitrary length. We present hypotheses on how to detect the sites of possible coherence problems based on the cohesion information supplied by our model. We also describe an experiment which we conducted to confirm the validity of our model, comparing the predictions of the model with text evaluations performed by human judges.

In addition, we discuss the areas of application for the model, commenting on how detecting sites of possible incoherence can be of value to problems such as text critiquing and second language learning and proposing new improvements to automated procedures such as natural language generation and machine translation.

The thesis therefore provides important new research within the field of computational linguistics on how a representation of the cohesion of a text provides an understanding of the coherence of that text.

Acknowledgements

This thesis would never have been written without help and support of many people. First, I would like to thank my supervisor Robin Cohen for all the help, for gentle strength, and for being on my side.

I would also like to thank the members of my committee: Cecile Paris, Frank Tompa, Fahiem Bacchus, and Neil Randall.

Peter Vanderheyden helped tremendously with rather mundane logistics after I moved out of Waterloo, making my life a lot easier. Thanks, Peter.

I would like to acknowledge the help of Giovanni Merola in setting up the statistical analyses of my experiments.

Susan Williams has been a wonderful friend throughout all the thesis fighting process. She talked sense into me when it was needed, and commiserated or made me laugh, whichever worked better at a moment. Here's a toast to you, Susan. And also a toast to Mario Gauthier, particularly for his help with a prosaic car breakdown on the day of the defense.

Another person that deserves special thanks is Mert Cramer. His loud realism at a difficult time was, and is, greatly appreciated.

Thanks to Wendy Rush, for listening, and for asking the right questions that forced me to think through my assumptions.

My parents have always believed in me, and their support means a lot to me. My dad actually traveled thousands of kilometers to attend a defense in a language he couldn't understand, so that he could lend his moral support. Dziekuje.

It makes me very sad that one very important person will never see this thesis. Serious illness claimed Slawek Franaszek before the work was completed, but his kindness and wisdom remain with me forever.

It is perhaps unusual to thank non-humans, but I feel Sava should be mentioned here. My gentle Furry Creature deserves the largest pig ear I can find.

In loving memory of Slawek

Contents

List of Tables

List of Figures

Chapter 1

Motivation

The purpose of this thesis is to develop a computational model to analyze and represent lexical cohesion of texts of an arbitrary length. The model is intended to be used as a means for evaluating the coherence of the analyzed texts.

1.1 Basic concepts

What makes a text? What are the properties that determine if a collection of sentences fits together? Meaning certainly influences textuality, but to say that meaning binds a text to form a unified whole is simplistic. Clearly, a text does not happen just because several sentences are about the same thing. There are linguistic phenomena other than semantics that create the unity of a text. These phenomena bind components of a text into an organized, harmonious unit.

Before we can discuss these phenomena further, we need to decide what constitutes a text. One approach considers a text to be a rigorously defined structure with sentences as basic constituents. Hence, one can design a text or story grammar which

defines what is an acceptable text, in a similar way as it is done for sentences. Van Dijk [1972] and others are the main proponents of text grammars. These grammars seem to work in some genres with strictly defined rules, such as romance novels, in which the structure of a text is perfectly predictable. However, outside these genres, the usefulness of story grammars is uncertain.

Halliday and Hasan [1976] (and also independently other researchers, such as Hoey [1991]) believe that it is a mistake to consider a text to be a 'super-sentence', that is, a higher-order structure in which there are relationships between sentences defined in the same strict way as the syntactic relationships between sentence constituents within one sentence. It is therefore their opinion that a text is not a grammatical unit, and that in general it is impossible to construct a grammar for it. Rather, Halliday and Hasan view a text as a unit of meaning, where the relationships between constituents are not syntactic, but purely semantic and functional. A text is realized as a string of related sentences. Since there is no limit on the number of those sentences, text sizes can vary considerably.

We accept the view of text as a string of related sentences with special relationships between text constituents. However, we believe that in addition to semantic relationships there also exist lexical connections between text constituents. In this thesis we in fact discuss the nature of these connections in more detail.

It is also important to develop a more precise notion of text for computational purposes. We therefore consider a text to be a string of related sentences which also has the property of *textual unity*.

Textual unity can be achieved in many ways[1]. Most typically, several devices work together to convey the sense of unity of a text. A uniform, harmonious style is one

[1]All languages have resources to create textual unity, but we will focus our attention on English for now. These language-specific resources make it possible to create a unit of meaning out of

example of such a device. Another good example, a lower-level one, is cohesion.

Cohesion is a property of text which is achieved by making connections between text constituents. There are two kinds of cohesion — syntactic (signaled by the use of certain parts of speech or grammatical constructions), and lexical (concerned with a certain class of lexical relationships among individual words and phrases). It is important to note that cohesion is not a binary property of a text — rather, different texts may display different degrees of cohesion. Moreover, cohesion is not uniform across a text: some parts may have more connections than others.

It is also important to understand that textual devices such as cohesion, if used correctly, are usually transparent to the reader. In other words, a reader does not pay much attention to the way the text is constructed and unified. However, the lack of cohesion, particularly in longer texts, produces a prose that is "choppy" and difficult to understand. Yet, cohesion is not a necessary condition for text understanding. It is entirely possible for a text to lack cohesion and still be understood. This is the result of coherence, a property of text that helps the reader make logical sense of a sequence of phrases or sentences. Coherence is achieved by establishing *semantic* relationships between text constituents.

Psycholinguistic research into coherence [Blakemore, 1987] shows that people have a strong motivation to make sense of even loosely connected sentences. The fact that it is difficult to find an example of a totally incoherent text supports the validity of Blakemore's claim. Consider, for example, the following sentences:

(1) I arrived today. Nellie tripped over her trunk.

separate sentences. Without them, a 'text' is not really a text, because it lacks unity and does not function as a unit of discourse.

Taken as presented, these sentences are loosely associated and, without a context, it is difficult to see how they could form a text. However, with a little creativity, it is not that difficult to supply a context in which these sentences will make sense as a part of a discourse. For example, the sentences can form part of a narration, and Nellie might have been mentioned in a preceding sentence. Or perhaps Nellie is a common acquaintance of both the speaker and the hearer, and the speaker brought her trunk with him (perhaps to return it to her) and placed it clumsily.

Cohesive and incoherent texts are even rarer. Consider the following nonsense text (from [Morris and Hirst, 1991]):

(2) Wash and core six apples. Use them to cut out the material for your new suit. They tend to add a lot to the color and texture of clothing. Actually, maybe you should use five of them instead of six, since they are quite large.

This text is cohesive because of the use of the pronouns *them* and *they*, which must refer to the word *apples*. The words *suit* and *clothing* provide another cohesive connection, and *five* and *six* yet another. But it is difficult to make sense of the text as a whole.

This particular example is artificially constructed, however. In this thesis we will examine real texts which have been written for actual readers, but which still may have problems with coherence.

1.2 The aim of the thesis

As stated previously, our aim is twofold: first, to develop a model to represent lexical cohesion of texts; and second, to show how those representations can be used to evaluate coherence of the underlying texts.

To date, there have been only a few proposals for representing lexical cohesion of texts. The first, [Morris and Hirst, 1991], although a useful starting point, was not implemented. It also was somewhat limited in its scope. In subsequent work, [Hirst and St-Onge, 1995], a variation of this model was implemented, but still had similar limitations.

There has also been work on representing coherent structure of text (e.g. [Hobbs, 1976] [Mann and Thompson, 1983]). The latter has been intended for natural language generation rather than analysis. But all this work assumes that the text is coherent to begin with.

With our research we are extending the state of the art in both lexical cohesion and in coherence. We have specific algorithms for constructing our representation of lexical cohesion. Various design decisions are made, presented and discussed. We also have specific hypotheses for interpreting these representations to draw conclusions about potential sites of incoherence in texts. This constitutes part of an evaluation procedure for text coherence. Moreover, this work contributes to the understanding of coherence itself.

Next we discuss the application of our work, for example, as a method for critiquing written texts, providing advice on where the writing should perhaps be repaired. We also have suggestions for improving various tasks in computational linguistics, such as machine translation and natural language generation, which may need to critique and re-evaluate how texts are constructed.

1.3 Overview of the thesis

The method we use in this thesis is as follows. First, we study background work on cohesion and coherence. This background is presented in chapter 2.

Next, we design a model for constructing the representation of the lexical cohesion of a text. We present our data structures, the lexical graph and the collapsed lexical graph, together with an algorithm for creating them. We present the hypotheses we have formulated after studying various texts, both well written and problematic. We also apply our method to sample texts. These results are discussed in chapter 3.

The model is implemented. This implementation is described in chapter 4.

We then conduct an experiment to compare the results of our method of evaluating coherence with those of human judges. Chapter 5 presents this experiment and discusses other defense of our findings.

We also propose specific applications for which our work might be useful. Chapter 6 contains a discussion of these applications. These include such areas as text critiquing, natural language generation, machine translation, information retrieval, and detecting incoherence in conversations.

We then step back to record various contributions and directions for future research. These are discussed in chapter 7.

Chapter 2

Background work

Text connectedness is a vital characteristic of a text. For this reason, there has been keen interest in the topic, and many researchers have attempted to work on it from various angles. Although considerable progress has been made, we still don't understand it fully.

Cohesion has been also researched, and, because it is a simpler phenomenon, it is much better understood. What is not clear is the relationship between coherence and cohesion. In subsequent chapters, we will give one possible account for the connection. But before we do, in this chapter we will discuss the text connectedness work that has been completed until now. We will begin with psycholinguistic research, to gain a broader perspective. Following that, we will turn our attention to computational approaches.

There are two directions of work on computational aspects of text connectedness: the discourse structure approach and the topic continuity approach. Both contribute to our understanding of coherence, but neither fully explains the phenomenon. In addition, these two views are not really complementary, but rather competing. This

research is outlined in section 2.2.

In this work, we take the first steps towards reconciling these two conflicting views. Since we recognize that coherence is a difficult problem, we don't claim to have the last word about it. Rather, we see this work as a step in the direction of better understanding the unity of text. We do this by focusing on the interpretation of lexical cohesion. In order to situate our particular approach to representing cohesion, we include a general discussion of cohesion in English and a brief summary of existing work on computational models of cohesion (in section 2.3).

2.1 Psycholinguistic research on text connectedness

Until fairly recently, most psycholinguistic research did not recognize coherence and cohesion as separate phenomena. Rather, the research was concentrated on text connectedness, covering both coherence and cohesion. For this reason, in this section we do not make the distinction either. However, where it is possible to determine whether a particular study was more significant for coherence or for cohesion, the distinction is maintained.

From the psycholinguistic point of view, there are three explanations of text connectedness. The first one is propositional. The earliest propositional model was developed by Kintsch and van Dijk [1977]. Using the idea of cyclic processing, they described text processing as follows. In order to determine where each new proposition of a text fits with respect to the text processed so far, a selected set of previous propositions is used. A special procedure selects new propositions from the working memory. This so-called leading-edge strategy takes into account the text hierarchy,

that is, the structure of the text, and the recency of the proposition, that is, the distance from the current to the selected previously processed one.

After the whole text is processed using this model, the result demonstrates the connectedness. In addition, Kintch and Van Dijk discuss what to do when there are certain problems with their method, such as when there are no propositions left in working memory. The first approach, reinstatement, involves re-introducing a proposition that has already been processed and removed from working memory. The second approach, inference, involves using real world knowledge to establish a link between the part of the text that has been processed and the remaining text. Inferences are rare. In fact, in Kintsch and van Dijk's model, inference is a way to gracefully resolve the otherwise unresolvable cohesive links.

The second of the main psycholinguistic approaches proposes a mental model to account for text understanding. The mental model approach is more psychologically plausible than the propositional model because it does not treat a text as a stream of propositions; rather, the aim now is to extract the text semantics in one pass. Johnson-Laird [1983] proposed a distinction between a propositional representation and a representation by mental models. The propositional representation is built by extracting from text facts that are stated explicitly. By contrast, the mental model does contain the propositions that occur in the text, but is augmented by a considerable amount of world knowledge not directly linked to the text (this idea is not original; cf. scripts as in Schank and Abelson [1977]).

The mental model is capable of capturing not only the explicit but also the implicit links present in a text. This is achieved by considering the text as a dynamic structure that interacts with the knowledge of the speaker and the listener.

The idea of mental models has been supported by psychological experiments.

However, some of those experiments have difficulties. The biggest problem is that experiments rely on researchers making underspecified off-line observations about their subjects' performance, rather than relying on objective measurements. This is a cause of concern that perhaps mental models are a reconstructive tool, rather than an actual text processing tool.

To remedy this problem, Sanford and Garrod [1981] proposed the theory of scenarios. Scenarios are essentially mental models, but are more precisely defined to contain three elements. First, scenarios include knowledge about social situations and their components (this is analogous to Schank's scripts, such as dining at a restaurant, for example). Second, scenarios include not only knowledge of people involved in a particular type of situation, but also knowledge of their particular social relationships with each other (such as waiters and guests in a restaurant). Third, scenarios contain knowledge about typical actions that occur in certain situations (for example, the most probable actions of a waiter at a restaurant).

The Sanford and Garrod model assumes that not all parts of the processing system are equally active all the time. They propose static and dynamic modules to differentiate between the more and the less active parts. Another useful feature of the model is the distinction made between the information contained in the text and information derived from world knowledge. This is analogous to the distinction between semantic and episodic memory proposed by psycholinguistics.

The model includes two foci. The explicit focus is that part of the world to which the text refers explicitly. The implicit focus involves the part of the world derived by the hearer or the speaker from the knowledge contained in the scenarios.

According to the mental model theory, the structure of the mental model is used as the basis for reconstructing the semantics of the text. For example, processing a novel

in this way explains the connecting roles of the major characters: they stay implicitly in focus throughout the novel, while the secondary characters disappear from implicit focus immediately after completion of processing of the current scenario. This does not mean that the author cannot refer to the minor characters once the scenario is processed. However, to refer to such characters requires the reactivation of the particular scenario in which they occurred.

The main contribution of the mental model approach is the evidence that there are certain texts for which it is impossible to fully explain text connectedness without considering the world knowledge of the speaker and the hearer.

The third, functional, approach to text connectedness is represented by Rickheit and Strohner [1986]. They propose to view text processing in terms of communicative functions, such as explaining or convincing. In fact, they claim that comprehension is not necessarily a result of following text processing procedures, but perhaps a creative action that does not have a clearly marked solution.

To explain the functional approach, we must first define the smallest particle of linguistic analysis. Rickheit and Strohner adopt Hörmann's [1981] definition[1] of the action unit. Using this unit, they analyze discourse into a sequence of goal-directed actions. This analysis is possible because linguistic activity has its source in other social and physical activities, and hence is not completely independent. As an example of such social activity, consider a speaker and a hearer cooperating on some task, such as renovating the basement. Let also one person be more knowledgeable about the renovation process. The cooperation starts with an agreement on the communicative goal, in this case, explaining. Since both speaker and hearer share the goal, they are both interested in achieving it. The necessity of renovation becomes the motivating

[1] As discussed in [Rickheit, 1991].

factor that guides the communication. (This is similar to the work on task-oriented dialogues by Grosz and Sidner [1986]).

Now let us turn our attention to the discourse itself. Clark and Marshall [1983] proposed a set of principles that a speaker uses to construct a text. The principles describe the ways to structure a text so that the hearer is able to create a coherent mental model of the message contained in it. If the same content is structured differently, the hearer has difficulties in constructing the appropriate mental model. Moreover, the specific difficulties form quite predictable patterns [Rickheit and Stroher, 1986]. Rickheit and Strohner describe the influences of several factors that could lead to problems with mental model construction. These include culture, knowledge and attitudes of the listener or reader, communicative situation, medium, and textual characteristics.

Certainly all these factors are important, but since we discuss machine translation as an application of our work in chapter 6, let us examine culture more closely (for the full discussion see [Rickheit and Stroher, 1986]). Psycholinguistic research shows that communicational conventions are culture-dependent. Since the conventions are applied to discourse processing, it is no surprise that ignoring them leads to misunderstandings (Friedriksen [1981] was among the first to demonstrate this). In addition, the schemas and scenarios discussed above are very definitely culture-specific. Obviously then, knowing the appropriate schema will be a considerable aid in text processing. An experiment by Kintsch and Greene [1978] shows that subjects from Western cultures were better able to understand stories from *The Decameron* than Alaskan Indian myths. This effect occurred not on the level of individual sentences and propositions, but followed from the different overall organizations of the stories. Another experiment paired Grimm fairy tales and Apache Indian tales; the subjects' task was to repeat the story they had heard.

Grimm stories have a conventional structure, so they were easy to process and the subjects' recollections were almost perfect. In contrast, the Apache tales had a loose structure in which the episodes were not time-related. Consequently, the subject recollections were very poor. It would be interesting to repeat the experiments on the Apache Indian subjects, but we can speculate that the situation would be reversed. Even with just this one experiment, the conclusion is that general text processing would require some modeling of cultural differences.

In this section, we discussed text connectedness as a whole. We will now return to our division of connectedness into simpler phenomena and consider them separately. In the next section, we will concentrate our discussion on coherence.

2.2 Coherence work

Coherence determines what makes sense in an utterance. There are two main, rival directions in the research on text coherence. One of them investigates *referential*, or *topic, continuity* and is concerned mainly with the content of the discourse. From this point of view, a discourse is coherent if it refers to the same, perhaps abstract, entity. The reference may take various forms, such as a stereotypical situation [Schank and Abelson, 1977], or semantic links between discourse segments (e.g. [Polanyi, 1988]).

The other direction, called the *discourse structure approach,* is concerned not with what goes on inside discourse segments, but rather with the segments themselves, and with their interrelationships. These interrelationships are thought to be context-independent. The main proponents of this approach include Grosz and Sidner [1986] (who proposed two types of relations), Mann and Thompson [1983] [1986] (who consider numerous relations), and multiple variations based on these two main ideas, including a taxonomy of coherence relations described by Sanders *et al.* [1992].

In this section, we will discuss the most prominent and most consequential of these works, many of them computational. We will examine their strengths and weaknesses. In chapter 3, we will present our own approach, which attempts to unify the above conflicting views. We begin, however, with psycholinguistic accounts.

2.2.1 Discourse coherence and referential continuity

Blakemore [1987] describes how context determines semantic constraints which then limit what can be said next and still be considered coherent. In other words, her approach is to constrain the part of discourse that follows the current focus so that the whole discourse is coherent.

One important lesson about discourse coherence follows from Grice's maxims. Consider the following example:

(3) A: Would you like some chocolate?
B: I'm on a diet.

Without Grice's Maxim of Quality (i.e., say only what is relevant), B's response would seem incoherent. But considering that the response is appropriate for the question, we can work out its intended meaning.

Normally, B's reply is taken as a negative answer. However, as Blakemore shows, no one would accuse B of lying if she took a chocolate. Moreover, if this were the case, her response would no longer be understood as declining the offer. Hence, the context is often the only way to determine what the utterance actually means.

Another interesting psycholinguistic model of text coherence is given by Givón [1992]. The difference here is that Givón studies coherence from the hearer's point of

view. Hence, the same text may be considered incoherent by one hearer and perfectly coherent by another. Givón proposes to use abstract mental entities, called *files*, to store all the knowledge a hearer has acquired. Assuming that during the conversation new knowledge comes only from the current discourse, the discourse processing proceeds as follows. The incoming information is classified by the hearer as new or already known. If the hearer perceives the information as known, then he mentally activates a particular file that holds this information. If, however, the incoming information is new, then the hearer must decide how important the information is. If the information is important, a new file is created. This new file is then activated and the information is placed there. But if the information is not considered important, then it is placed in the currently active file and no other action is required.

There are two difficulties with the model. The first difficulty is the lack of a procedure for deciding whether the information is new. Conceptually, this is not a serious problem, since one can say that the hearer knows what he knows and can determine that subconsciously. But in the computational context this requires the full user-model knowledge base to be searched and is therefore impractical. To simplify the matter, one can assume for example that all the knowledge presented in the text is new. However, it may well be the case that the distinction is crucial.

Another, and perhaps more difficult, problem is to decide whether the information is important. Givón proposes a grammar of *topic markers* to differentiate among important and less important information. The idea is based on the observation that each text contains both lexical and syntactic clues that could be useful for the construction of such a grammar. For example, sentence topics, called *referents*, usually contain the most important information in the sentence. The main topic of a sentence is usually its subject, with the direct object being a secondary topic, while all the remaining roles are non-topics. However, a formal description of these

clues would require a full computational theory that does not yet exist. In particular, such a description would have to disregard the sentence boundaries, since "... the topic is only important if it remains 'talked about' through a number of consecutive sentences." ([Givón, 1992, 12])

2.2.2 Discourse structure approach

In the previous section we discussed the topic continuity approach. Now, we consider the discourse structure approach. The main characteristic of this approach is the underlying assumption that significant information about the discourse can be discovered without considering the meaning of the entire discourse. Rather, the proponents of this approach study discourse constituents and the relations among them.

An early example of work in this field is the model of Reichman [1978]. Her model considers text as a dynamic structure that interacts with the knowledge of the speaker and the listener (incorporating the idea of a mental model described in section 2.1).

A discourse is divided into context spaces used for recording the structure of a text. Context spaces are recognized by identifying certain clue words[2] (i.e., words and phrases which signal discourse connections, such as *therefore*, and *for example*). Reichman's rules for context spaces in discourse constitute, as well, a characterization of what are coherent configurations of discourse segments.

Reichman relies heavily on clue words. A work which addresses clue word interpretation but is not reliant on it is the model of Cohen [1987], used to process a specific kind of discourse — arguments.

Cohen's work on understanding natural language arguments presents a specific characterization of coherent argument structures, used to limit the analysis of these

[2]alternatively called clue phrases or cue words

texts. Particular strategies of presenting claims and evidence in arguments are identified and labeled as coherent; then, the overall analysis procedure is restricted to searching only for these kinds of structures. This is one approach to characterizing coherence for texts in terms of acceptable orderings of propositions within texts (but only for a very restricted kind of text, i.e., arguments).

For the discourse structure approach, research then emerged which focused on characterizing possible relations between discourse constituents. We will begin our discussion with work that proposes the smallest number of discourse structure relations and progress towards more complex theories.

Grosz and Sidner

One of the first discourse structure theories is that of Grosz and Sidner [1986]. They try to formalize discourse structure, but only for a particular type of discourse, the task-oriented dialogue. Even with this limiting assumption, however, they were able to discover that discourse can be logically divided into smaller parts, called *discourse segments*. Each such segment can be further subdivided. The division of discourse into segments is determined by three separate, but inter-related, components: the linguistic structure, the intentional structure, and the attentional state.

The linguistic structure is defined simply as the particular sequence of utterances. The utterances are grouped into discourse segments in such a way that two consecutive utterances may or may not be in the same discourse segment. The segments are related to each other by embedding relationships, which are a surface manifestation of the intentional structure. Some linguistic expressions, such as phrases (*e.g., in the first place*), intonation, changes in tense or aspect, all carry important clues about the boundaries between discourse segments.

The intentional structure addresses the purpose of the discourse. Recall that Grosz and Sidner restrict themselves to goal-oriented discourse. The individual discourse segments have their own purposes, usually compatible with the overall discourse purpose.

Grosz and Sidner claim that there are only two structural relations in the discourse structure. These are *dominance* and *satisfaction precedence*. The dominance relation is used to relate segments when the purpose of one segment contributes to the satisfaction of the purpose of the other segment. The satisfaction precedence relation determines the sequence in which discourse segment purposes must be satisfied.

The attentional state is an abstract representation of the foci of attention of the discourse participants. It is modeled by a set of *focus spaces*. A focus space is associated with each discourse segment and represents the objects and relations currently in focus. However, in the Grosz and Sidner model the attentional state is a property of the discourse itself, not the discourse participants. For this reason, focus may shift from one discourse segment to the next.

There are limitations concerning focus shifts from one segment to another. These limitations are imposed by the *focus stack*, a data structure that stores focus spaces introduced in previous parts of the discourse. Shifts in focus determine some linguistic characteristics, such as the use of anaphora and personal pronouns. But, more importantly for us, shifts in focus determine the structure of the discourse.

If we consider texts to be particular types of discourse where the author addresses an intended audience, and view text structure in terms of discourse segments, we can begin to see how the model of Grosz and Sidner can be used to explain some aspects of text coherence. The differences between text and discourse are discussed again in section 6.8.

While Grosz and Sidner's work is generally descriptive rather than processing-oriented, it does have some important consequences for us. First, it shows that the discourse can be partitioned into smaller, simpler segments. Another important contribution of this work is that there are relationships between discourse segments. Finally, the work shows that understanding focus shifts is necessary for generating well-written texts.

It is possible to disagree on the basis of partitioning, on the choice of relations, and on the mechanism for shifting focus. But it is still useful to employ all of these in a discourse processing model.

Segmenting the discourse

Grosz and Sidner do not discuss how to segment the discourse. This question has been investigated by other researchers. One interesting account has been given by Passoneau and Litman [1993]. They performed an empirical study on a corpus of transcripted spoken narrations to determine where, according to human readers, the segment boundaries occur. They compared their findings against three simple algorithms, based on referential noun phrases, cue words, and pauses.

The three algorithms are very simple. The first one uses noun phrases. If the phrase in the current clause location refers to the current segment, then no boundary exists at that sentence. However, the texts need some pre-processing to identify the referents, which is not trivial.

The cue words treatment is simpler. The cue words at the beginning of a prosodic phrase are taken as denoting the beginning of a new segment.

As for pauses, phrases that begin segments are correlated with duration of preceding pauses, while phrases ending the segment are correlated with subsequent pauses.

Although these algorithms help to provide a basis for segmentation, they are still limited. In particular, this work did not attempt to perform the hierarchic decomposition of the text. Passoneu and Litman claim that human subjects found the process difficult, and the results were unreliable.

Other work on discourse segmentation (using lexical cohesion) is discussed in section 2.3.

Understanding shifts of focus

The idea that focus spaces can be organized into a focus stack was used by McKeown [1985] and later by McCoy and Cheng [1988] for generation of coherent texts.

The TEXT generation model, developed by McKeown [1985], makes use of an explicit focus to identify the appropriate order of sentences. In order to do this, she loosely defines a set of rhetorical predicates, such as analogy or inference. Using these predicates, she defines schemas that are then used as templates for generating various texts of paragraph length. Such generation is possible because the schemas can be applied recursively.

To explain focus shifts, McKeown defines four possibilities. The simplest possibility is to retain focus from one proposition to the next. Alternatively, focus can be shifted to an item mentioned in the previous proposition. It is also possible to return to a topic that has been discussed before. However, this return closes the previous topic for further generation. It is then difficult to return to the closed topic gracefully. Finally, one can select a proposition with the greatest number of implicit links to the previous proposition.

The TEXT system was able to generate simple coherent paragraphs. However, it was very restrictive about the form of the generated text and allowed little variation

because schemata were used. Therefore, it can determine the coherence of only a limited class of texts.

Another work based on the idea that focus constrains what can be said next is the work of McCoy and Cheng [1988]. They observe that none of the previously proposed mechanisms can handle all focusing phenomena. For this reason, they tried to extend the treatment of focus and proposed a new data structure, called a *discourse focus tree*. The tree is constructed and traversed, one node at a time, as the discourse proceeds. The tree can contain several types of nodes to account for various types of focusing phenomena. Although all nodes represent essentially the same thing—the topic of the conversation—they are stored differently because they are implemented as pointers to entities of different types in the knowledge base. Hence, the types of objects focused on depend on the ontology.

Just because an object was mentioned in a conversation, it does not necessarily mean that one can talk about anything pertaining to that object. The context places additional restrictions on what can be discussed next. To implement the restrictions, the focus of attention is calculated using not only the node itself, but also its ancestors and siblings. To obtain additional information, McCoy and Cheng use Cohen's [1987] idea of clue words to facilitate building the discourse focus trees. Phrases, such as *in addition*, or *in particular*, changes of tense, use of anaphor, and switching of pronouns all carry significant meaning that can be used to determine focus.

This work would benefit from several extensions. First, it lacks formal specifications. For this reason, it is difficult to duplicate. In fact, it is not clear how general the proposed solution is. Testing on several different domains would perhaps establish its greater applicability. Finally, there is a stopping problem: a discourse focus tree can grow infinitely large, since in the model presented there is no limit on its size. However, in real discourse one does not talk indefinitely; there always comes a

time when the speaker decides that he has said enough. Hence, there is no need to construct nodes of the focus tree that will never be used.

Coherence relations — Hobbs

The work of Grosz and Sidner described above uses only two relations: dominance and precedence. However, many researchers believe this approach to be limiting. One alternative is to use more relations and to apply them to a text in a systematic way in order to arrive at the semantic interpretation of the text.

Much of this work is inspired by a much earlier approach developed by Hobbs [1976] (and later in [Hobbs, 1985]). His aim was to precisely define coherence relations in order to analyze the coherence structure of text. The basic idea is to design a small set of coherence relations that are powerful enough to express the coherence relationships between text constituents. In order to recognize these relationships, Hobbs emphasizes that rich world knowledge is necessary.

Hobbs proposes several types of coherence relations: overlapping temporal succession, cause, contrast, elaboration, example, and parallel. He also provides a formal definition for each relation. In addition, he recognizes a disguised relative clause, or a set of sentences that really should be one sentence.

There are two immediately apparent difficulties. First, it is not always clear how to choose text constituents. This is the problem of the last relation, the disguised relative clause, which requires an arbitrary classification. Hobbs gives the outline of the process, but it is somewhat vague and no implementation is discussed. Second, it is not clear that one of the relations will always apply. There is a description of how to choose the relation, but no procedure to follow if no relation fits.

Hobbs uses a sentence as the basic unit of analysis. Observing that redundancy

is very high in language, he uses this redundancy to match some aspects of items in consecutive sentences in order to determine the coherence relation that best fits the given pair. However, as the relations are applied, the earlier relation has no bearing on the subsequent choice of the relations, so that the decisions about coherence are all local.

Agar and Hobbs [1981] recognize three kinds of coherence. Global coherence is seen in terms of global goals. Based on these goals, one develops a plan, breaks it into subgoals, and produce an utterance. Hence, the realization of global coherence follows a top-down design.

Local coherence arises when the speaker needs to expand a subplan. For example, if the plan is to tell about an event, the speaker may need to supply some background information first. Hence, local coherence is realized in a bottom-up way.

The third type of coherence is themal, and it roughly corresponds to topic continuity.

Rhetorical Structure Theory

Another approach somewhat similar to that of Hobbs is Rhetorical Structure Theory (RST). To formally describe how sentence constituents relate to one another, RST ([Mann and Thompson, 1983], [1986]) uses various types of relationships between individual sentences as well as between constituents within one sentence. The basic unit of description, the RST schema, consists of a nucleus and a satellite, bound together by a relationship (see Figure 2.1 for the list of original RST relations). Both the nucleus and the satellite are subject to constraints particular for the given relationship. In addition, the theory claims that a specific rhetorical effect results when a particular nucleus is bound with a particular satellite.

1. circumstance;	2. solutionhood;
3. elaboration;	4. background;
5. enablement;	6. motivation;
7. evidence;	8. justify;
9. volitional cause;	10. non-volitional cause;
11. volitional result;	12. non-volitional result;
13. purpose;	14. antithesis;
15. concession;	16. condition;
17. otherwise;	18. interpretation;
19. evaluation;	20. restatement;
21. summary;	22. sequence;
23. contrast;	24. join;

Figure 2.1: The list of the original RST relations.

RST was originally designed for text generation. To generate a text, one particular RST relation is chosen to span the entire text. Since any relationship has a nucleus and a satellite, the generator needs to create them. Each nucleus and each satellite can in turn itself be a relation. Hence, the process is recursive. The process stops when all the basic units of generation are included in the resulting tree.

RST has certain shortcomings. First, there is no possibility of representing the overall structure of a text within the RST framework, except, perhaps, in the case of the simplest texts, because in general it is extremely difficult to find just one relationship that will span the whole text. For example, it is not possible to generate circular arguments using RST alone[3].

Another serious problem is that the effects that are assigned to relationships are treated in isolation. In real texts, there is an interplay of subtle effects between text constituents. Therefore, describing the relationships between pairs of text constituents does not fully describe the rhetorical effects of a whole text. And RST is even more limited than this: it does not allow for a description of relationships between all text constituents, only consecutive ones.

This latter observation gave rise to a proposal by Moore and Pollock [1992] to modify RST to provide two types of information. The first one is related to the informational structure, and the second to the intentional one. This would require two types of relations sometimes holding simultaneously between text constituents. This idea seems to provide some help to increase the expressive power of RST.

There are, however, other problems with RST. It is non-reproducible — i.e., it does not explain why the particular relationships have been chosen, or why they are needed in the first place. RST is not complete, so it is not known how many new

[3]One might argue that circular arguments perhaps are not the best kinds of text to aim for. Still, they exist and are sometimes appropriate.

relationships will be needed. Judging by the past performance of RST, where new relationships were added whenever the 'theory' was not powerful enough, the number may continue to grow indefinitely.

There have been some attempts to apply RST to text analysis [Marcu, 1996]. This poses yet another problem: the analysis phase must decide which relationship best fits a particular text chunk. It turns out that this is largely a matter of taste, since the relationships are not formalized. In many cases, there are not even clear rules about which part should be a nucleus and which a satellite. Both assignments are not only legal, but also logical. However, one can imagine that particular assignments are important, since two different assignments would presumably create different rhetorical effects. Currently, the problem is pushed aside and the implementation relies on outside help, an oracle, to find the relationship that best fits a given pair of text constituents.

Making sense of the maze: Classification of relationships

Interesting accounts of text coherence based on RST-style relations link coherence with cognitive relations. We will describe one such approach, proposed by Sanders, Spooren, and Nordman [1992]. The working assumption is that understanding discourse amounts to a construction of a mental representation (cf. our discussion of mental models in section 2.1). Sanders *et al.* claim that coherence relations must be considered as cognitive entities. This claim leads to the conclusion that coherence relations affect discourse understanding. While current experimental research on discourse understanding seems to support this claim, it is not clear if what is actually being observed is the result of a representation created at the time of understanding, or a structure created to facilitate the recall. For this reason, we remain skeptical about the evidence, although the claim seems perfectly reasonable.

To classify coherence relations, Sanders *et al.* use a relational criterion. The criterion is defined as the property of a coherence relation which adds more information to the discourse that cannot otherwise be easily deduced from the contents of the discourse. In other words, the discourse structure itself carries important information about how the discourse should be processed. The meaning of the discourse is not ignored either. Rather, it must be compatible with a particular coherence relation.

The relations are ordered according to four primitives that satisfy the relational criterion. These primitives form independent dimensions along which we can evaluate all coherence relations. To explain the primitives, consider two discourse segments, S1 and S2. These two segments can be bound to corresponding propositions P and Q that can be related in various ways. The first primitive concerns *causality*. A relation is causal if it can be represented as P implies Q; otherwise it is *additive*. However, the logical implication does not suffice to determine the relation, since natural language does not closely correspond to formal logic. Rather, the theory uses the notion of relevant implication. Unfortunately, the work has nothing to say about how to determine relevance.

The second primitive, called *source of coherence*, determines whether a relation involves the propositions expressed by S1 and S2 or illocutions expressed by them (i.e., the intended meanings of the utterances).

The next primitive to consider involves the order of segments S1 and S2. If S1 corresponds to P and S2 corresponds to Q, then the relations are in the basic order; otherwise they are in the inverted order. This does not assume that the old information is presented first, followed by new information. Rather, the order is the property associated with discourse segments, and not with propositional contents.

Finally, a separate primitive describes the polarity of segments. If S1 and S2 cor-

respond to P and Q, then the relation is positive, but if they correspond to negations, then the polarity of the relation is negative. Positive relations often are introduced by such words as *and* or *because*, while negative polarity often occurs with the words *but* and *although*.

Using these primitives in various combinations, we can sort the relations into classes in such a way that all relations in the same class share the values for all the primitives. For example, the relationships claim-argument, consequence-condition, and goal-instrument share the following set of primitives: causal, pragmatic, non-basic order, positive.

Sanders *et al.* claim that, using their primitives, most coherence relations in literature can be assigned to one of twelve such classes. Moreover, they conducted psychological experiments to demonstrate the psychological plausibility of such a classification. There is, however, no discussion on how to realize their theories in a computational model.

This section has shown the variety of approaches for representing, creating, and analyzing coherence. Now, let us turn our attention to a different aspect of text connectedness: cohesion.

2.3 Cohesion in English

In chapter 1, we described cohesion as a mechanism that binds the text to create a unified whole. In a sense, we can consider cohesion as a kind of 'glue' that holds a text together. Cohesion occurs both within a sentence and across sentence boundaries. Usually, all sentences of a text except for the first one display some form of cohesion with a preceding sentence. Typically, the strongest cohesive bond will occur between

consecutive sentences, but different arrangements are not unusual.

Cohesion is not uniform across a text. Some parts, such as whole paragraphs, can be very tightly connected, while others may have many fewer cohesive links, sometimes called ties. We can therefore conclude that cohesion is a matter of degree, with considerable variations even within one text.

The various degrees of cohesion can be put to an interesting use, creating elegant variations in a piece of prose. In fact, many authors use this device to create a rhythm with periodic variations of tight and loose texture [Halliday and Hasan, 1976]. Typically, the switches occur at paragraph boundaries. In fact, the notion of paragraph in written language was introduced to stress the differences of cohesion within a text.

2.3.1 Syntactic devices that promote cohesion

English is rich in constructions that promote cohesion. Since some of these devices are stronger than others, we can order them in a hierarchy according to strength. The strongest such a device is *substitution*, since it relies solely on the text, and not on the outside reference. *Ellipsis*, a form of substitution, does not involve meaning at all—it merely uses relationships between words and structures, without considering semantics. Next in this hierarchy comes *reference*, which relates items through meaning. Therefore its interpretation requires understanding of its linguistic environment. A still weaker cohesive relation is *conjunction*, a form that specifies how the following text constituent is connected to the preceding one. The weakest cohesive relation is *syntactic parallelism*, since it only stylistically unites the text, without specifying any formal cohesive relationship. In this section, we will examine the hierarchy more closely.

Substitution and ellipsis

Substitution refers to a replacement of a particular word, phrase, or clause with a different, usually more generic, one. Ellipsis is the omission of a word, phrase, or clause. Halliday and Hasan [1976] show that we can regard both of these as the same category, since ellipsis can be viewed as zero-substitution. For this reason, we will discuss various types of substitution and ellipsis together. Since substitution, including ellipsis, is purely textual and has no function other than to promote cohesion, it is the most cohesive of all the cohesive devices.

The first type of substitution is nominal. Here, a noun or the whole noun phrase is replaced by the substitute, such as *one* or *same*, as in this example:

(4) These biscuits are stale. Get some fresh ones.

In all cases of nominal substitution, there must be a noun that acts as a head of the nominal group. As Halliday and Hasan put it, "...the noun to fill this slot will be found in the preceding text (occasionally elsewhere)." [Halliday and Hasan, 1976, 92]. In their terminology, the substituting item *presupposes* the substituted one. Usually, only the noun itself, but not its modifiers, is understood by the substitution. The anaphoric pronoun may then itself be modified, as in the above example where the pronoun *one* is modified by *fresh*, while the head noun *biscuit* is modified by the adjective *stale*.

Next we will consider verbal substitution, the construction in which a verb or a verb phrase is replaced with the verbal substitute *do* in the appropriate form and tense, as in the following example:

(5) I don't know the meaning of half of those long words, and, what's more, I don't believe you do either!

Here, the verb substitute *do* replaces the whole verb phrase *know the meaning of half of those long words*. Hence, verbal substitution is exactly like nominal substitution, with the obvious distinction that it applies to verbs. In the above example, the substitution occurs within one sentence, so its cohesive force is not strong. However, substitution across sentence boundaries has a much stronger cohesive effect.

In clausal substitution, not just a phrase but the whole clause is presupposed, as in the following example:

(6) Is there going to be an earthquake? It says so.

In this example, the word *so* presupposes the whole clause *there is going to be an earthquake*. Clausal substitutions can be classified according to the types of clauses they replace, and so we have reported clause substitution, conditional clause substitution, and so on.

We will now turn our attention to ellipsis. As we mentioned above, ellipsis can be viewed as a special form of substitution in which the hearer must supply the omitted information. However, it does have some other characteristics that merit attention. First, the fact that the hearer must supply some additional information in order to understand the sentence does not automatically imply that we are dealing with ellipsis. In fact, in all sentences, vital information is not spelled out and yet we do not feel that we need some previous sentence or clause to understand the text [Halliday and Hasan, 1976]. For an elliptical sentence, there is an empty slot that can be filled with a constituent from elsewhere in the text. As with substitution, therefore, ellipsis presupposes the existence of the referent of the appropriate form. For example, in the following text, the nominal ellipsis presupposes the head noun *verse*:

(7) Would you like to hear one more verse? I know twelve more ∅.

Similarly as for substitution, we can classify ellipsis according to the item it elides. And so we have nominal, verbal, and clausal ellipsis. These behave similarly to their substitution counterparts. However, some types of ellipsis deserve more attention. One form of ellipsis deals with polarity, that is, positive or negative modes of a sentence. The unusual property of this ellipsis is that it allows the whole sentence to be elided, as in this example:

(8) Were you daydreaming?
No.

Another interesting type of ellipsis deals with finiteness and modality. The ellipsis does not require either. As a result, we have all possible combinations of finite and non-finite clauses, such as in the following four examples.

In the first example, a finite clause is presupposed by another finite clause:

(9) The picture wasn't finished. If it had been, I would have bought it.

Next, we show a finite clause presupposed by a non-finite:

(10) He's always been teased about it. I don't think he likes being.

We can also have a non-finite clause presupposed by a finite one:

(11) What was the point of having invited all those people?
I didn't; they just came.

Finally, we present a non-finite clause presupposed by a non-finite:

(12) It was hard work parceling all those books.

I'm sure it was; and I'd much prefer you not to have.

In a similar way, ellipsis does not require the presence of modal verbs (e.g. *would, could, etc.*). As a result, we have all possible modal and non-modal combinations, with the additional possibility of a modal presupposed by a different one. Moreover, ellipsis does not presuppose tense; hence we can have various tense combinations.

There are, however, some characteristics that are always preserved by ellipsis. One such characteristic is voice, active or passive. If the rule of voice preservation is violated, we obtain texts that do make sense, but are considered ungrammatical. This can be illustrated with the following example:

(13) They haven't finished the picture. If it had been, I would have bought it.

However, there is one condition where the rule does not hold—the switch of actor/goal relationships, such as in the following example:

(14) Will you be interviewing today? No; being interviewed.

Reference

References are items that are not semantically interpreted independently. Rather, they reach out to another item in the text for interpretation [Halliday and Hasan, 1976]. Hence the interpretation requires another item to be present. This other item does not necessarily occur in the text itself, but can be a part of the situational context. This kind of reference is called *exophora.* If, however, it is a part of the text, it is called *endophora.* There are two possibilities for endophoric reference. It

can either tie back to the preceding text (*anaphora*), or tie forward to the text that follows (*cataphora*). Of these three reference kinds, anaphora is the most common. The cohesive power of endophoric references lies in their ability to refer forward or backward in the text.

According to Halliday and Hasan [1976], English has three types of referential constructions: personal pronouns, demonstratives, and comparatives. All of them can be of any reference kind and all are usually a part of the nominal group.

Personal reference, as the name implies, usually refers to a person. This type of reference can be further subdivided into personal pronouns (such as *we*, possessive determiners (or possessive adjectives, such as *his*), and possessive pronouns (such as *mine*), all occurring in the grammatical forms appropriate to the given context. However, not all personal references are inherently cohesive. Halliday and Hasan argue that first and second person personal pronouns are not cohesive, because they do not refer to the text at all, but rather are defined by the speech roles of the speaker and hearer, and hence are exophoric. The only exception is the use of first or second person pronouns in direct (quoted) speech. In this case, the pronoun refers to the preceding text and hence does promote cohesion.

The use of personal pronouns is often not an optional feature, but a necessary device. Consider the following example:

(15) John took off John's hat and placed John's hat on the shelf.

In this sentence, it is natural to assume that there are many Johns involved, wearing each other's hats. To avoid this confusion, the use of anaphoric pronouns is necessary.

There are two special cases in which pronouns function somewhat differently than described above. One such case involves a personal pronoun that does not refer to a

person, but to a specific portion of a text, or to a fact. The only pronoun that can function is this role is *it*, as this example demonstrates:

(16) He said he will come. Do you believe it?

The other special case is the zero anaphor, that is, one that does not explicitly occur in the text. Zero anaphor is sometimes a source of disambiguation problems [Givón, 1992]. Consider for example:

(17) Jack thought about Jill
(a) giving birth to their baby.
(b) playing touch football with the gang.
(c) sitting alone on the porch.

There is a strong tendency to interpret the zero anaphor of (a) as co-referential with *Jill*, that of (b) as co-referential with *Jack*, and that of (c) as ambiguous, which results in different cohesive links in each case.

Demonstrative reference is used to express location in terms of proximity and is functionally equivalent to pointing. This type of reference can be divided into two subtypes. The *neutral* subtype is represented by the determiner *the* (which Halliday and Hasan view as a particular type of demonstrative that does not have any content of its own). It serves to differentiate among things that are readily identifiable or mutually known to the speaker and the hearer, and other things. The other subtype, the *selective*, allows the speaker to differentiate among items near and far (*this* versus *that*), among singular and plural participants (*that* versus *those*), between current place and someplace else (*here* versus *there*), and between current time and some other time (*now* versus *then*). Not all demonstrative references are cohesive. For example,

the structural cataphor of the next sentence is not, since it is just a particular type of English syntactic construction.

(18) He who hesitates is lost.

In contrast, the following textual cataphor is very cohesive:

(19) Those were the verses the White Rabbit read: [... *the verses* ...]

Comparative reference is indirect, and is used to establish identity or similarity. English uses two types of comparison. The general, or deictic, comparison establishes identity (*e.g., same, identical*), similarity (*e.g., so, likewise*), or difference (*e.g., other, else*). The likeness is taken as a whole, without considering individual properties of the compared items. The particular, or non-deictic, establishes numerative difference (*e.g., fewer, equally many*), or a difference in a particular property (*e.g., more advanced*).

Conjunction

Conjunctive elements do not reach forward or backward across the text, but rather they bind consecutive textual components together [Halliday and Hasan, 1976]. There are four types of conjunction: additive (in its simplest form represented by *and*), adversative (*e.g., yet*), causal (*e.g., so*), and temporal (*e.g., then*).

The simplest form of conjunction is the *and* relation. Because this relation is really a coordination, it is strongly structural and only mildly cohesive. In its cohesive guise, it is retrospectively additive, that is, its meaning is projected back into the preceding text, as in the series *men, women and children.* Here, the *and* coordinates not just *women* with *children*, but also *men* with *women.*

Other conjunctive elements include *but, yet, so,* and *then.* However, not all of them reach back into the text the way *and* does. The *but* conjunction is contrastive, but it also includes part of the meaning of *and.* Consider this example:

(20) The eldest son worked on the farm, the second son worked in the blacksmith's shop, but the youngest son left home to seek his fortune.

Here, the *but* projects backwards in the same way *and* did above, but the meaning of this projection is to coordinate the constituents not with *but*, but with *and* !

The adversative relation is used to signal a meaning contrary to expectation. It can be expressed by various words and phrases, such as *but, however, on the contrary, instead of, though, at least, yet*, and occasionally even *and.* In the following example, the second preposition is contrasted with the first, giving a strong cohesive effect:

(21) He's not exactly good-looking. But he's got brains.

Causal relations can be signaled by words and phrases such as *so, hence, therefore, in consequence, because of that, etc.* The relations can be chained together to form one long causal relation, as in this example:

(22) The blow to his head caused bleeding. Because of that, he lost a large amount of blood. In consequence, he died.

It is not necessary that the cause textually precedes the effect. The reversed form **B, because A** is equally acceptable as the form **because A**, **B**.

The temporal relation expresses the sequence of described events. It is usually realized using words such as *then, next, after that, subsequently, ten years later, etc.*

It can also describe simultaneous events, using phrases such as *at the same time* or *simultaneously.* It is also possible to narrate the events in the reverse order of occurrence, using constructions such as *previously, before that, etc.*

A separate group of conjunctive items is formed by continuatives. These include miscellaneous items such as *now, of course, anyway*, and *well.* Their role is not uniform. For example, *now* often marks the beginning of a new discourse segment. *Of course* signals that the speaker expects the hearer to already know what is being said. The role of *well* is to acknowledge that a question was heard and the answer will follow, or to give the speaker time to prepare the answer.

Parallel sentence structure

Halliday and Hasan [1976] argue that cohesion does involve meaning. However, syntax can play a strong supporting role in cohesion. For example, similar sentence structure (parallelism) does convey a sense of unity [Kerrigan, 1974].

A parallel sentence construction is one of the simplest forms of promoting cohesion. If used by itself, its effect is quite mild, but used in conjunction with other devices it can be very powerful. To describe it simply, it is the repetitive use of the same sentence structure in two (or more) consecutive sentences. The effect can be strengthened by repeating a particular sentence constituent, always with the same function in both sentences.

2.3.2 Lexical cohesion

The connective power of related words is one of the strongest textual devices for creating cohesion. Using related words results in a text 'talking about' the same

thing, that is, creates a strong sense of continuation. The cohesion of the text is created by semantic relationships among words.

These semantic relationships are described in detail by Halliday and Hasan [1976]. Here, we will summarize their classification. The first class is *reiteration*, which can be further subdivided. The simplest form of reiteration is repetition of the same word, where both occurrences are co-referential, as in the following example:

(23) John bit an apple. However, the apple had an unexpectedly sour taste.

Another form of reiteration does not rely on the occurrences to be coreferential. Rather, one reference is specific, and the other generic, as in this example:

(24) Have some chocolate. I know you love chocolate more than any other candy.

Repetition does not have to be exact. Often, the author chooses a synonym or a near-synonym to increase variety, as here:

(25) Have some chocolate. I know this is your favorite candy.

Another example of this inexact repetition is to use a closely related word that has a different syntactic function. For example, if one occurrence is a verb, the other may be a related noun. Consider the following list of words related in this way: *noun, nominal, nominalize, nominalization.* Using words related in this way in consecutive sentences strongly promotes cohesion.

A reiteration in a broader sense may involve superordination or subordination:

(26) I won't eat broccoli. I hate vegetables.

Broadening the scope again will yield another class of reiteration, this time involving the general word replacing the specific, as in this example:

(27) I got "Voltaire's Bastards" for Christmas. The whole thing is written very well.

Treating the word 'reiteration' very loosely, we have systematically classifiable relations, such as colours. The following sentence is an example of such a relation:

(28) I prefer yellow roses to red roses.

The remaining classes do not involve reiteration, even in a very loose sense. For this reason, they are not easily formalized. Morris and Hirst [1991] call them "not systematically classifiable", while Halliday and Hasan [1976] define them in terms of collocation (i.e., these items occur often together). These classes include antonyms, words related by the part-whole relationship, elements of ordered sets, such as days of the week, common activities, *etc.* In the next example, the words *ill* and *doctor* are related in a non-classifiable way:

(29) You look ill. Go see your doctor right away.

Obviously, lexical cohesion is not limited to pairs of related words. In real texts, one can distinguish whole chains of mutually related words, spanning whole paragraphs, and even crossing paragraph boundaries.

2.3.3 Summary of lexical relations

As we have seen, there are many different types of lexical links, each having different cohesive power. After Halliday and Hasan [1976], we can recognize the following links with diminishing cohesive strength:

- repetition, where both occurrences are co-referential. This is the strongest cohesive lexical link possible.

(30) John bit an apple. However, the apple had an unexpectedly sour taste.

- repetition, where one occurrence is specific and the other generic.

(31) Have some chocolate. I know you love chocolate more than any other candy.

- synonymy, which can be regarded as non-exact repetition. For example, *pretty* and *beautiful.*
- antonymy, which is synonymy with a negative sign, such as *pretty* and *ugly.*
- same word with different syntactic function, such as *nominal, nominalize, nominalization.*
- superordination, such as *broccoli* and *vegetables.*
- systematically classifiable relations, such as colours.
- non-classifiable relations, such as *ill* and *doctor.*

Computational approaches to lexical cohesion

Morris and Hirst [1991] take the first steps towards providing a computational model for analyzing lexical cohesion, with an underlying aim of determining the structure of texts analyzed in this way. They describe a method for identifying lexical chains using thesaural relations. A thesaurus is a reference book of lexical items and specific

relations between them. Morris and Hirst chose Roget's International Thesaurus as the basis for their algorithm.

The motivation behind the use of this particular thesaurus is that it is a large database of purpose-organized words, with over a thousand categories, each with subcategories organized in a hierarchical structure. However, the deciding factor in choosing it was that it groups words by ideas, without naming particular relationships between them, such as the IS-A relationship.

To construct lexical chains, the idea is to scan the text, looking for words that can be related using one of lexical relations in the specified set. These relations are very specific to the thesaurus used. Roget's thesaurus has a tree-like concept hierarchy structure. Morris and Hirst used that structure to determine the lexical links by traversing the hierarchy. And so, the sibling relation was the strongest, followed by the parent relation, the grandparent relation, and the "uncle" relation being the weakest they allowed.

More specifically, a lexical relation is found when one of five kinds of the following relationships are recognized. Words can have a common category, *e.g*, *residentialness* and *apartment*. The second case occurs when the category of one word has an index to the category of the other, *e.g.*, *car* and *driving*. Next, one word may be a label for the other, *e.g.*, *blind* and *see*. Another case involves two words in the same group, *e.g.*, *blindness* and *vision*. Finally, two words may have categories with indices which point to a common category, *e.g.*, *brutal* and *terrified*. The first two categories are used most often.

Their algorithm works as follows. First, the word is checked against a list of ubiquitous words, and if found there, it is discarded. Ubiquitous words, such as *good, do, etc.*, as well as personal pronouns, are not good candidates, since their high

frequency makes them less central to the task of representing text structure. All other words are considered good candidates. Morris and Hirst then evaluate the current word to see if it belongs in one of the current chains. To verify that, they try to find if there is a lexical relation between the word in question and the last lexical item in the chain. If the fit is found, the word is appended at the end of the lexical chain. Otherwise, the next chain is tried. If the word doesn't fit any of the chains, it opens a new chain of length one. In this way, only the physically closest relations are considered.

It is easy to find relationships between pairs of words. However, transitivity poses a problem. Allowing unlimited transitivity may lead to 'chains' that are related only accidentally, as the following sample 'chain': *cow, sheep, wool, scarf, boots, hat, snow.* For this reason, Morris and Hirst allow one step transitivity, that is, one forming a chain of three words. However, this number is arbitrary, and they consider it to be only a guide.

For practical use, the chains are evaluated according to their strength. The strength is determined using the number of reiterations, the chain's density, and the length. The strength of the chain can then be used as one of the clues for determining the structure of a text, since stronger chains tend to be confined to structural units, such as paragraphs. The chains can also serve as indicators of semantic relationships between text units. Another important use of lexical chains is word sense disambiguation. Since the elements of the chain are lexically and semantically related, they supply a context that narrows the set of possible meanings of the ambiguous word.

This view of lexical cohesion is limited, as it considers only lexical relations between words that are physically close together and places restrictions on the form of lexical links allowed.

Another source of problems is the thesaurus used. Since the thesaurus is intended for general use, it is less successful for texts with domains which are specific and narrow. But in an experiment which analyzes sample texts by hand, even for general domains, the method misses about 10% of lexical relations. This limitation is inherent in the method because the thesaurus does not contain all possible relations for all possible word senses.

There are further limitations from the method of lexical chaining. For instance, there is a tendency to join independent chains when an ambiguous word accidentally matches two separate chains. Another source of difficulty is transitivity. Allowing one-step transitivity may eliminate some perfectly valid chains, and may create counterintuitive chains of the length of three words, such as *cow, sheep, scarf*, the one-step transitive subchain of the previously discussed chain.

The work of Morris and Hirst relies on the choice of Roget's thesaurus. Because this thesaurus is not available in electronic form, this work has never been implemented. One variation that has been implemented is the work of Hirst and St-Onge [1995]. Instead of the unavailable Roget's thesaurus, Hirst and St-Onge uses WordNet, which is easily available. The method he used is the same as the original one described by Morris and Hirst, and has inherited the same limitations.

Using lexical cohesion to divide the text into coherent segments

Work on discourse (eg. [Grosz and Sidner, 1986]) often relies on the idea of segmenting the discourse into coherent chunks, but the problem of how to actually perform the segmentations is not addressed in detail. Rather, it is assumed that the segmentation is given.

We have already seen in section 2.2.2 how one can segment the discourse by

searching for certain signals to discourse structure.

A different approach to solve this very problem is addressed by Hearst [1994]. Her work describes TextTiling, a system capable of handling real length texts without preprocessing. Hearst is not interested in how the segments are related, only in the segmentation, and does not compute the hierarchic structure of segments.

Hearst uses lexical cohesion (focusing on the repetition relation) to find the partitions. Her algorithm consists of three parts. First, the text is tokenized into individual units combined into groups. Rather than using actual sentences, Hearst chose to divide the text into pseudo-sentences of fixed length, but keeping track of paragraph boundaries.

Second, the pseudo-sentences are grouped into blocks of fixed size. The algorithm now looks for overall similarity between token sequences in adjacent blocks. Each pair of token groups receives a similarity score based on the number of words they share and on the frequencies of these words.

Finally, the algorithm looks for boundaries between blocks. A boundary occurs when a dramatic decrease in the scores is detected.

The algorithm is implemented. In addition, Hearst has performed a study comparing human judgements of text partitioning and her approach.

A different and rather complex approach to segmentation has been presented by Kozima [1993]. The work uses LDOCE as the underlying dictionary, and attempts to compute similarities between words based on spreading activation on a semantic network that represents the dictionary.

Using the list of similarities, Kozima computes a lexical cohesion profile (LCP) of a text, with text constituents boundaries visible. Again, the method partitions the text into cohesive chunks, but has nothing to say about how the chunks are related.

A valid criticism of Kozima is expressed by Hearst [1997]. She claims that Kozima does more computation than is necessary in order to achieve the same results as her own work.

2.4 Is the distinction necessary?

In the preceding sections we have examined different types of coherence and cohesion. A different approach has been pursued by Zadrozny and Jensen [1991]. They claim that it is not necessary to distinguish among various text-connectedness phenomena, and they do not make a distinction between coherence and cohesion. To establish connectivity, they examine dictionary definitions for all words in the text. If one word occurs in the definition of the other, or if definitions of both words contain a word in common, then this co-occurrence, called *consistency*, is enough to establish connectivity. The most important consequence of this approach is that there no longer is a need to differentiate among types of connectedness. Rather, Zadrozny and Jensen claim that consistency is all that is required.

Clearly, this method is not nearly sufficient. In fact, using this technique we can only establish some aspects of lexical connectivity (for example, antonymy would be difficult to capture). We have already discussed a similar but more systematic method that correctly identifies more links [Morris and Hirst, 1991].

2.5 Cohesion as a clue to understand coherence

Examining the related work described in this chapter leads to an appreciation of the difficulties involved in developing models to analyze cohesion and coherence in texts.

For our research, it was important to be comprehensive, investigating work within computational linguistics, both in natural language analysis and natural language generation, and considering the insights gained by researchers in psycholinguistics as well.

Our ultimate proposal is to take advantage of some of the ideas in both the topic continuity and the discourse structure approaches. Most of the work on topic continuity is not computational, and most of the work on discourse structure is focused on developing a representation of the structure of text, which is different from our aim of specifying criteria for the evaluation of text coherence. Moreover, we found certain fundamental difficulties in simply trying to transfer existing theories of coherence into a procedure for coherence evaluation.

For example, much of the work described in this chapter included aspects of modeling the speaker's and hearer's beliefs (part of the study of natural language pragmatics) or of representing real world knowledge and reasoning with that knowledge. Our decision is to focus on an aspect of natural language processing which was simpler to represent and reason about — lexical cohesion, which may be defined in terms of relations between words stored in thesauri, thus bypassing the need for more complex processing that is not well understood. In the next chapter, we discuss the potential for employing domain-specific thesauri. This strategy helps to factor away some of the difficult problems which arise when trying to process a wide range of general purpose texts. Our intention is to also explore in more detail the relationship between coherence and cohesion, in fact using lexical cohesion as an indicator of text coherence.

In the next chapter, we describe our model for constructing a representation of lexical cohesion and employing it to evaluate text coherence. One contribution is simply to automate a procedure for displaying the lexical cohesion of a text. We

make clear the comparison with the previous efforts on constructing lexical chains, presenting our model as more extensive. We also emphasize the value of our model for the analysis of lexical cohesion, discussing specific proposals for its application to tasks where potential sites of incoherence must be identified.

In chapter 7, within the context of discussing our contributions and possible future work, we re-examine how some of the related work described in this chapter could be employed, together with the model of the thesis, for a more comprehensive coherence evaluation procedure.

Chapter 3

The analysis of lexical cohesion

In the previous chapter we have discussed two competing approaches to text connectedness. We have also seen how the discourse structure approach can be implemented in a computational way, but we said nothing about implementing the other approach. In this chapter, we will show that the topic continuity approach is also a valid computational option. Moreover, we will show how to unify the two competing approaches. We will tackle the problem of text coherence indirectly, using cohesion as a clue and a guidance.

3.1 What can lexical cohesion say about coherence of text?

From our earlier discussion it is clear that designing a computational model to determine text coherence is a difficult problem. Because of this, we decided not to try to solve it directly. Rather, we use another text phenomenon, cohesion, as an indicator of the text coherence.

It is important to note that we don't claim cohesion is a perfect indicator of coherence. We don't believe any partial theory of coherence can fully account for this complex text phenomenon. Still, we will show that lexical cohesion is a useful indicator in spite of its imperfections.

We begin with the observation that lexical choice in a text is not incidental. Quite the contrary, the individual lexical choices are influenced by the topic and the organization of the text. For this reason, lexical cohesion is an excellent candidate for an indicator of coherence if we approach it from the perspective of topic continuity. Our research aims to show how cohesion plays a supporting role to coherence and how a representation of it can be used as a tool for determining potential coherence problems in a text.

As we have seen, cohesion is a text-level phenomenon that is quite separate from coherence. It is a property that binds the text to create a unified whole. In a sense, we can consider cohesion as a kind of 'glue' that holds a text together. It also plays a supporting role to coherence by making connections from one part of a text to another explicit for the reader, which in turn aids the comprehension.

As we discussed in section 2.3, there are two aspects of cohesion: syntactic and lexical. Both are present in any text, and both are useful, but here we concentrate on the lexical aspect only, since the connective power of related words is one of the strongest textual devices for creating cohesion. The lexical cohesion of a text is created by semantic relationships among words. Using related words results in a text 'talking about' the same thing, that is, creates a strong sense of continuation. The possible semantic relationships between words are described in detail by Halliday and Hasan [1976] and summarized in section 2.3.

Obviously, lexical cohesion is not limited to pairs of related words. In texts,

one can distinguish sets of mutually related words, spanning whole paragraphs, and crossing paragraph boundaries.

As mentioned in section 2.3.3, Morris and Hirst [1991] describe a method for building lexical chains using thesaural relations. However, their view of lexical cohesion is somewhat limited, considering only lexical relations between words that are physically close together and placing restrictions on the form of lexical links allowed.

In the remaining sections of this chapter, we will build on the idea of lexical chaining, extending it and applying to our particular task — estimating text coherence. Using a thesaurus, we will collect all the lexical cohesive information present in the text into one structure, the lexical graph. We will then use the graph for determining possible coherence problems present in the text.

3.2 The links and the thesaurus

In section 2.3.3 we have listed the lexical relations available to us. However, many of these relations cannot be computed without performing the full semantic analysis of a text. For this reason, we decided to concentrate on the following relations:

- repetition in both forms, without distinguishing the type.
- same root, different syntactic function.
- synonymy.
- antonymy.
- superordination (is-a link);
- part-of.

Obviously, if we hope to compute lexical cohesion automatically, we need a good, reliable method for calculating individual lexical links. The method we chose was inspired by Morris and Hirst in that we also use a thesaurus, although a different one.

Since we need to collect all the possible lexical cohesive information we can from a text, we require an underlying thesaurus that supports all the lexical links that we need. Because we can only consider the lexical links that can be found using the chosen thesaurus[1], the choice of a particular thesaurus is an important design decision.

We have examined several different thesauri, both general and domain-specific. The general thesauri we looked at include the Webster online dictionary (both the definition and the thesaurus parts), Kipfer's thesaurus[2] [Kipfer, 1995], the online Oxford English Dictionary and Wordnet [Beckwith *et al.*, 1991][3]. In addition, we have constructed our own, domain-specific thesaurus for the domain of financial advice. After analyzing over a hundred texts we have concluded that it is more advantageous to use a domain-specific thesaurus than a general one.

There are several reasons for this preference. First, all the general thesauri we examined were incomplete. For example, the Webster online dictionary did not have crucial links, so it wasn't possible to find lexical relations between obviously related words, such as *cauliflower* and *vegetable*. This in itself wouldn't necessarily be dis-

[1]With the exception of repetition, the computing of which does not require a thesaurus.

[2]This thesaurus is structured around concepts, much like the non-online version of Roget's thesaurus, but it is less elaborate.

[3]Strictly speaking, Wordnet is not a thesaurus, but a hierarchy of nouns connected by various lexical relations. We included it in our considerations because it looked promising due to the quality of lexical relations it supports. In fact Wordnet was used by St Onge [1995] to build lexical chains in the Morris and Hirst style.

astrous if the relationship were simply not supported for all lexical items, but unfortunately Webster's is not consistent and so it was possible to find other links of this type, such as *cabbage* and *vegetable*. These inconsistencies made the thesaurus unreliable. In addition, computing pleonyms, or links between related words of different classes, such as a noun and a verb, turned out to be very difficult with this thesaurus. The third difficulty was that Webster's thesaurus only supports a limited number of link types: synonyms, antonyms, and near-synonyms. Similar problems were found in Oxford English Dictionary.

Another general-purpose thesaurus we considered but eventually did not use was Kipfer's thesaurus [Kipfer, 1995]. The main advantage of this thesaurus is that it has a well-defined structure. At the top of the hierarchy there is *concept*. Each concept has several *entries* associated with it. Each entry has a *definition* and a list of *synonyms*. Using this thesaurus, it is possible to relate words of different classes, such as nouns and verbs. This particular thesaurus, while nicely structured and more consistent than other thesauri, does not support the antonym relation. We feel that since antonymy is such a strong lexical relation, not having it is unacceptable for our purposes.

Wordnet [Beckwith *et al.*, 1991] also did not live up to our expectations. The earlier version of it consisted of separate hierarchies for nouns and verbs, and no obvious way of connecting them. We found that it missed many vital links between related words of different classes, such as a noun and a related verb (e.g. *decision, decide*). For this reason, we ruled out that version. The newer version [WordNet, 1995] solved that problem, but the result is that Wordnet is now much like the Webster online thesaurus, with the same problems. Specifically, we have found that it does not cover many domain-specific concepts.

In addition to all these problems specific to individual thesauri, we found one more problem common to all the general thesauri we considered, and one that likely applies to all general purpose thesauri: in most cases, it was possible to connect massive amounts of lexical items, with most of the links uninformative. In other words, by using a general-purpose thesaurus, one can often connect words by lexical relations that are too general, which results in a massive amount of useless lexical links[4].

For this reason, we opted for the more modest number of lexical links generated by a domain-specific thesaurus, because the links were more specific and hence gave us more information about the lexical cohesion between the items.

To demonstrate how this idea works in practice, we decided to construct our own domain-specific thesaurus geared specifically towards finding the lexical relations in one chosen domain, financial advice texts. Clearly, our choice of domain does not restrict the approach in any way. For other domains, all that is required is a domain-specific thesaurus. If the quality of the thesaurus is acceptable, the method will perform equally well.

Let us now look at the process of constructing the domain-specific thesaurus as we have completed it. We began by selecting the list of lexical links that will hold between the lexical items. Our list includes:

- pleonymy[5],
- synonymy,
- antonymy,

[4]Hearst [1994] makes this same observation when considering the work of Morris and Hirst [1991].

[5]Pleonymy is a relation that holds between words that share the same root but have a different syntactic function. For example, *investor* and *investing*.

- hyponymy (is-a relation),
- meronymy (part-of relation), and
- systematically classifiable relations[6].

We don't include repetition and forms of the same word, because these relations can be computed without a thesaurus.

In order to make sure that our thesaurus had as good a coverage as possible, we collected a corpus of texts in our financial domain[7], ensuring that the texts use as much of the domain vocabulary as possible.

We extracted by hand those words that pertain to our domain, while disregarding the words that are not domain-specific. Once we had a substantial list, we went through it, grouping the items into clusters of related words. Within each cluster, we explicitly recorded the relationship between each pair of words that were lexically related[8].

Because we began with a limited amount of texts, the process of constructing the thesaurus was an incremental one. As new texts were added to our corpus, new words that were missing from the thesaurus were identified and added. As the corpus grew, so did the thesaurus, making the coverage more complete with time.

A more concise description for constructing a domain-specific thesaurus is shown in Figure 3.1.

[6]The kind of systematically classifiable relation which we include is the one that holds between two items that are in an *is-a* relation with a third item.

[7]We used 97 texts found by browsing the World-Wide Web.

[8]To speed up the thesaurus lookup, we recorded each pair twice. See chapter 4 for more detail.

1. decide on the set of lexical links to be used in the thesaurus;
2. collect a corpus of texts in the chosen domain;
3. extract the domain-related words and put them in a list;
4. group the items on the list into related clusters;
5. within each cluster, explicitly record the relationship between each pair of words that are lexically related;
6. if a new word is found after the thesaurus is constructed, add it to the appropriate cluster by specifying all the relevant lexical links.

Figure 3.1: Summary of thesaurus construction.

One interesting avenue of research into thesaurus construction that we considered but did not explore in depth is automatic thesaurus construction [Srinivasdan, 1992]. The idea behind most methods is to compute the frequencies for words as they occur in a large corpus. The words that occur in the corpus most often, and are not "ubiquitous", constitute the basis of the newly created thesaurus. To compute relations, the methods rely on collocation. These words that occur together most often are considered lexically related. The strength of the relationship is computed according to the frequency of co-occurrence. The method offers no way to determine the type of the relationship.

We believe automatic thesaurus building is an interesting research direction. There are many advantages to this approach. The most important ones are:

- It is quick. Constructing a thesaurus by hand is a lengthy and tedious process. If we can speed it up by automating it, it is definitely worth it.
- It increases the likelihood of the resulting thesaurus having a reasonable coverage if the corpus is sufficiently large. Of course in this method too one

must be careful to ensure adequate coverage by choosing a large enough corpus of texts.

- It includes collocation for free. Pairs of items that often occur together will be identified by many of the automatic thesaurus construction methods. It can be argued whether collocation constitutes lexical links, but links based on co-occurrence definitely are informative.

There are also disadvantages to this approach. The most important ones are:

- We already mentioned that it does not give relationship types, only strengths, which is proportional to frequency of co-occurrence. It may not matter in the end, but if we decide to restrict our attention to lexical links only, then we have no way of identifying the links that are not lexical, such as collocation, for example.

- One must be careful about coverage or else the method might miss important thesaural entries. Obviously, the choice of corpus becomes all important.

One interesting possibility that perhaps could offer all the advantages of automatic thesaurus building without any of its pitfalls is to combine these two approaches and make the method semi-automatic. In other words, we could begin by having the automatic thesaurus construction software extract the words from the corpus, and then identify the relations by hand. This would speed up the construction process while avoiding the problems with the fully automatic method. Clearly, more research is needed to determine the usefulness of this approach.

In the light of all these limitations of the method we have considered, we decided that a domain-specific, hand-constructed thesaurus is our best option.

3.3 The data structure — the lexical graph

For the lexical analysis to be indicative of text coherence, we need a way to obtain and store all the lexical items present in the text together with all the lexical links that hold between these lexical items.

In order to represent the complete lexical cohesive structure of a text, we designed a data structure called *lexical graph*. In this section, we define this graph, describe its properties, give the directions to construct it, and show how it can be used for coherence analysis.

3.3.1 The description of a lexical graph

The lexical graph is a data structure that contains all the information about lexical cohesion that can be extracted from a text using the underlying thesaurus. It is a non-directed graph whose nodes consist of all the lexical items present in the text that also occur in the thesaurus.

If an item is repeated, the repetition constitutes a new lexical item that is linked to the first one by the repetition relation. Hence, there are as many nodes of a lexical graph as there are information-carrying lexical items.

We define the information carrying words as these words in our domain that are not overused. We have excluded the non-information carrying, or ubiquitous words. These include all articles, some verbs (*make*, *do*, or domain-specific ubiquitous like *profit*), etc. We have also excluded the pronouns, since the cohesive links they represent are not lexical, and we currently cannot handle anaphora resolution.

The arcs of the lexical graph represent lexical relations as determined by our thesaurus. And so, if two items contained in the nodes of the graph are lexically

related, and we can establish that relation using a thesaurus, there will be an arc connecting these items. This is the reason why choosing a reliable thesaurus is so important.

Since most words, even ones belonging to the same domain, are not lexically related, the lexical graph of a typical text is quite sparse. In fact, usually the graph will consist of several disjoined subgraphs, called components. The components of a lexical graph then are just clusters of lexically related words.

Intuitively, the longer the text, the more and richer components it should have. The reason behind expecting it is rather obvious: longer texts tend to give more detail, and use more varied vocabulary.

The shape and size of a component depends on the role it plays in a text. Some components are very small, consisting only of a few lexical items and spanning only one or two paragraphs. Typically, such local components occur in those portions of a text that describe some detailed concept. Other types of components span considerable chunks of text, often as much as several sections. These large-span components occur when the text describes something in more general terms.

Even though the shapes and sizes of individual components vary from one text to the next, one can distinguish several typical types. The following classification describes the most common types.

Components by type of lexical item

The simplest and perhaps the least interesting component consists of just one node. In well-written texts, particularly very short or very long ones, such *singleton components* practically never occur. The reason for this is that well written short texts are "lexically tight", i.e. contain many lexical links. Since there is little space to

develop any ideas, the space is used for a succinct presentation of related material. Conversely, in long texts, there is plenty of room for developing ideas, and so few lexical items are mentioned in isolation.

The singletons tend to occur sometimes in medium length texts. One possible explanation is that such texts do offer some space for new ideas, but not enough to develop them sufficiently. If this is true, then it might turn out that medium length texts are more difficult to write and require more support.

A slightly more complex component consists of multiple occurrences of a single lexical item connected by the repetition relation. This type of component occurs sometimes in very short texts. It rarely is found in longer texts, because typically the ideas are more developed, using related vocabulary. We call such a component *homogeneous*.

A *heterogeneous* component consists of different words linked by different relations. Some of the words may be repeated.

Components by span

A *local* component spans a short chunk of text, sometimes as short as one paragraph, but more often it touches a few consecutive paragraphs.

In contrast, a *large-span* component touches many paragraphs of the text, and not necessarily consecutive ones.

The largest possible component spans the entire text and touches all paragraphs.

The component that spans the entire text and contains the largest number of nodes is called *the main component* of the lexical graph.[9]

[9]Technically, it is possible for a lexical graph to have more than one main component. However, we have not yet encountered this in our examples. This might be one indication of incoherence.

Components by lexical density

We can compute the lexical density of a component as the number of nodes per paragraph. Texts differ widely in their lexical density, depending on the genre and the intended audience. But even within one text, individual components can vary equally widely.

A *dense component* has many nodes concentrated in one chunk of the text. Local components tend to be dense.

In contrast, a *sparse component* has only one or two nodes per paragraph, and often skips one or several paragraphs altogether. Usually, sparse components tend to be large-span.

3.3.2 What does the lexical graph say about text coherence

By examining both coherent texts and texts with some coherence problems we have discovered that the coherent texts share an important property: most of them contain main components. In contrast, many texts with coherence problems don't. Therefore, we present an important hypothesis:

Hypothesis 1 (Main Component Test) : If a text lacks the main component, this is an indication of a possible coherence problem. Furthermore, the chunks of text not represented in the largest component of the lexical graph are, most likely, the sites of the coherence problems.

The reason for expecting the main component in a well written text is rather obvious when we look at the graph analysis from the topic continuity perspective.

Any coherent text displays unity of topic, and all the words that directly relate to that topic are lexically related.

If the main component is absent in a text, then this is an important clue that the text might be incoherent. We can then point out these chunks of the text that are not lexically related to the rest of the text as potential sites of coherence problems. In other words, we can find chunks of texts, such as paragraphs or sections, that have no lexical links with other parts of the text. If this is the case, then we have a good indication that the particular portion of the text does not really fit in with the remaining text. This suggests that the text is, at least partly, incoherent.

Figure 3.4 shows the lexical graph of a text that lacks the main component. We can see that the first paragraph is not connected to the rest of the text and so we identify it as a site of a possible coherence problem. We will analyze this particular text in more detail later.

3.3.3 The idea of the collapsed graph

Examining the lexical graphs described in the previous section and applying Hypothesis 1 to them predicts coherence reasonably reliably for short texts. However, for longer texts, the results are not satisfactory. The reason for it is that in a longer text the writer has more space to develop more than one simple idea. There is also a lot more flexibility in arranging the contents of a longer text. For this reason, we need a representation more powerful than the lexical graph described in the previous section.

To obtain that new representation, we transform the lexical graph by collapsing it. The collapsed lexical graph consists of paragraphs, rather than individual lexical items, as nodes.

Now, to describe the arcs of the collapsed graph, we need to introduce the idea of a lexical bond. A lexical bond between paragraphs consists of all lexical links that span these two paragraphs. More formally, consider two paragraphs, P1 and P2. Now let W1 be a collection of all lexical items in P1, and W2 in P2. Let w1(i) be the i-th lexical item in W1, and w2 the j-th lexical item in W2. The lexical bond between paragraphs P1 and P2 is a set of all lexical links computed as follows. For any w1(i) in W1 which is lexically related to some w2(j) in W2, the lexical relation is added to the lexical bond between P1 and P2. If there is no pair w1(i) and w2(j) such that there is a lexical link between them, then there is no lexical bond between these paragraphs. In other words, the paragraphs are not lexically related.

The collapsed lexical graph then has lexical bonds as arcs. For example, consider the lexical graph of the text in Figure 3.3 shown in Figure 3.4. We can collapse this graph in the following way.

First, we draw the graph with paragraphs as nodes. In the beginning, we haven't yet computed the lexical bonds and so the graph has no arcs.

Next, we add the arcs one by one, by finding all lexical links between words in each pair of paragraphs. Consider first paragraphs 1 and 2. Since there are no lexical links spanning these two paragraphs, we don't add an arc between them. The same situation occurs for the pair of paragraphs 1 and 3.

Now, consider paragraphs 2 and 3. There are eight lexical links that span these two paragraphs: *investments* — *bonds*, *investments* — *stock*, *investment* — *bonds*, *investment* — *stock*, *stocks* — *stock*, *stocks* — *bonds*, *bonds* — *stock*, and *bonds* — *bonds*. Therefore, the lexical bond between these two paragraphs consists of these eight links. Because there is a lexical bond between these paragraphs, we add an arc in the collapsed graph. The complete collapsed graph for this text is shown in Figure

3.5.

3.3.4 How does the collapsed lexical graph reflect text coherence

If we construct a collapsed lexical graph of a coherent text, we will see that usually the graph is connected. This means that there exists a path between each two nodes. In other words, each pair of paragraphs is lexically related, if only by virtue of transitivity. Moreover, we can identify a paragraph that has the largest number of lexical bonds. We call such a paragraph a *central paragraph* of the text, and consider its presence another important indication of text coherence. Hence, we present another hypothesis about text coherence.

Hypothesis 2 (Central Paragraph Test): If a text lacks the central paragraph, it is an indication of a possible coherence problem. The site of the problem is most likely to be a paragraph for which there is no path to the largest connected subgraph of the collapsed lexical graph.

As was the case with the main component, it is conceivable for a text to have two paragraphs that have the same largest number of other paragraphs directly connected to them. However, this is not a problem from the coherence perspective. In such a text, we needed to break the tie by choosing one of the candidate paragraphs, and so we decided that the central paragraph is the one that occurs physically earlier in the text. The reason behind this choice is an empirical one — the earlier paragraph more often than not is in the introduction, and it might be important for some applications. For example, in information retrieval it would make sense to offer the central paragraph for the user evaluation. The user then would read this paragraph

and decide if the whole text is interesting enough to be retrieved. This could speed up the retrieval process. See section 6.5 for more discussion about this application.

The presence of the central paragraph in a coherent text should not be surprising, since in such texts the paragraphs will be semantically related. In a descriptive text, for example, we will often have the first paragraph describe the general properties of an object. The subsequent paragraphs might describe the more specific properties. These subsequent paragraphs don't necessarily have to be mutually related, but they will be related to the first paragraph. Hence, all paragraphs are connected and the first one is the central paragraph with most connections to other paragraphs.

Consider now a text with some coherence problems. In such a text, there might be a paragraph that doesn't quite fit with the rest of the text. If this is the case, then such a paragraph might consist of vocabulary that is not related to the rest of the text.

Now, if we construct a collapsed lexical graph of such a text, we will find that the graph is not connected with the rest of the collapsed graph. In other words, the paragraph with coherence problems is not lexically connected to the rest of the text.

In this case, the collapsed graph lacks the central paragraph, which after the absence of the main component is our second indication of coherence problems.

3.4 The algorithm

Now, let us turn our attention to the more detailed description of our algorithm for computing lexical cohesion of a text.

The input to our system is a text in ASCII. The idea is to analyze the whole

text.[10]

For each word in the text that is represented as a node in the lexical graph, we store the following information:

- the lexical item as it occurs in the text;
- the root form as it occurs in the thesaurus;
- the sentence number;
- the paragraph number.[11]

The algorithm for computing and analyzing lexical graphs is shown in Figure 3.2. The procedure for building a lexical graph representation is as follows:

1. **Compute the lexical graph**.

   (a) The text is read one word at a time. Each word is examined in turn and checked against the list of ubiquitous words. If the word is ubiquitous, it is discarded and the next word is read. Otherwise, the word is put in lower case and any punctuation is stripped. A rudimentary morphological analysis determines the root form for the word.

   The morphological analysis is necessary since for regular words our thesaurus only stores the root forms. And so, for nouns, we remove the plural markers, for verbs we remove the gerund and third person endings.

[10]However, if a text fragment is chosen for analysis, and that fragment talks about one topic, then such a fragment could possibly be analyzable by the system as well, as it should be reasonably coherent.

[11]Sentences and paragraphs are numbered according to the order in which they occur in the text.

(b) Now, we determine if the root form occurs in the thesaurus. If it does, then the word is stored in the lexical graph together with the root form, and the sentence and the paragraph number in which the word occurred. Otherwise, the word is discarded.

We continue reading and processing words until the end of file is reached.

After all the words have been read, we now have all the nodes of our lexical graph. In the next step, the lexical links are determined and stored.

(c) For each saved word, we calculate the lexical relations with all the words we have saved. We begin with the most cohesive relation, repetition, and try all the relations in turn until either one holds, or we have run out of the possibilities. In this way, we always store the most cohesive relation. If a relation is found, we store it in the graph.

(d) As the links between words are computed, we record in the graph which paragraphs the links span. This is so that we can calculate the number of lexical links between paragraphs.

2. **Find the main component**

(a) At this point, the lexical graph has been constructed. We now have all the information about lexical cohesion of the text and are ready to perform the lexical analysis of the text.

First, we find the largest component of the graph, i.e. one that touches the largest number of paragraphs. It will be the candidate for the main component (to be determined next).

We begin with the first lexical item and assume it is in the main component. We traverse the graph, noting which paragraphs we have visited already. If all paragraphs are accessible from the first word, we have found both the largest and the main component. Otherwise, we store the information about which paragraphs were represented in the component we just traversed and move to the next lexical item, if any. We continue in this way until we either find a component that touches all the paragraphs or run out of lexical items.

(b) **Analyze the lexical graph**

If all paragraphs are represented in the largest component, we conclude that the analysis was successful and that the text is coherent as far as we can tell.

Otherwise, it is possible that the text has some coherence problems and we need to analyze it further.

First, we traverse the largest component and find those paragraphs that are not represented in it. These are the sites of potential coherence problems.

As discussed in section 3.3.3, just because a text lacks the main component does not necessarily mean that it is incoherent. So, we must determine which test applies. In order to do this, we currently just examine the text length in paragraphs. If the text is shorter than 5 paragraphs and it lacks the main component, then we conclude that we likely have found coherence problems.[12] We point out the unlinked paragraphs to the user.

[12]Obviously, any hard threshold is arbitrary. We decided on five paragraphs based on our text corpus, but this is flexible.

For longer texts that lack the main component, we need to perform the following steps, which involve collapsing the graph and finding the central paragraph.

3. **Collapse the graph**

   (a) Next, we construct a graph whose nodes represent paragraphs of the analyzed text. For all pairs of words in the original graph, if they are lexically linked and are in different paragraphs, we add an arc between corresponding paragraph nodes. We also calculate the strength of the arc by counting the lexical links it comprises.

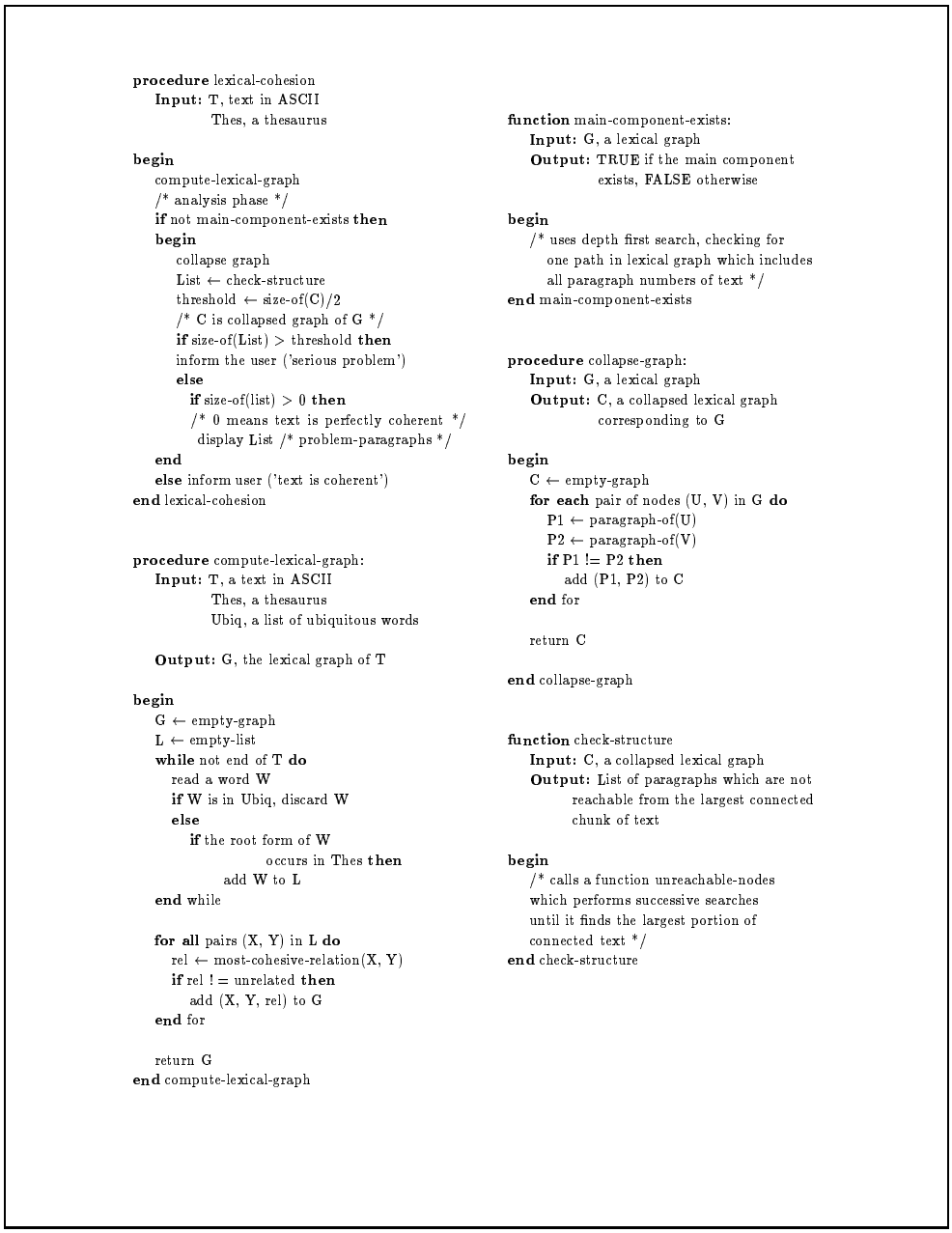
4. **Analyze the collapsed graph and give feedback to the user**

   (a) Starting with paragraph 1, we traverse the graph to identify paragraphs not reachable from the beginning of the text. We display those paragraphs to the user, suggesting that these might be the sites of possible coherence problems.

   If more than half the paragraphs are not reachable from the beginning of the text, it is reasonable to assume that the problem is likely at the beginning of the text. In this case, we try successive paragraphs until either more than half the text is covered or more than half of the paragraphs are examined. In the latter case, we display a message that the text has a poor lexical structure.

   (b) If there are paragraphs that are not connected, we inform the user.

Because we consider all possible pairs of lexical items, the complexity of our algorithm is $O(n^2)$, where n is the number of lexical items in the graph. Since we use



```
procedure lexical-cohesion
   Input: T, text in ASCII
          Thes, a thesaurus

begin
   compute-lexical-graph
   /* analysis phase */
   if not main-component-exists then
   begin
      collapse graph
      List ← check-structure
      threshold ← size-of(C)/2
      /* C is collapsed graph of G */
      if size-of(List) > threshold then
      inform the user ('serious problem')
      else
        if size-of(list) > 0 then
        /* 0 means text is perfectly coherent */
         display List /* problem-paragraphs */
   end
   else inform user ('text is coherent')
end lexical-cohesion


procedure compute-lexical-graph:
   Input: T, a text in ASCII
          Thes, a thesaurus
          Ubiq, a list of ubiquitous words

   Output: G, the lexical graph of T

begin
   G ← empty-graph
   L ← empty-list
   while not end of T do
     read a word W
     if W is in Ubiq, discard W
     else
        if the root form of W
                 occurs in Thes then
           add W to L
   end while

   for all pairs (X, Y) in L do
     rel ← most-cohesive-relation(X, Y)
     if rel != unrelated then
        add (X, Y, rel) to G
   end for

   return G
end compute-lexical-graph


function main-component-exists:
   Input: G, a lexical graph
   Output: TRUE if the main component
           exists, FALSE otherwise

begin
   /* uses depth first search, checking for
      one path in lexical graph which includes
      all paragraph numbers of text */
end main-component-exists


procedure collapse-graph:
   Input: G, a lexical graph
   Output: C, a collapsed lexical graph
           corresponding to G

begin
   C ← empty-graph
   for each pair of nodes (U, V) in G do
     P1 ← paragraph-of(U)
     P2 ← paragraph-of(V)
     if P1 != P2 then
        add (P1, P2) to C
   end for

   return C

end collapse-graph


function check-structure
   Input: C, a collapsed lexical graph
   Output: List of paragraphs which are not
           reachable from the largest connected
           chunk of text

begin
   /* calls a function unreachable-nodes
   which performs successive searches
   until it finds the largest portion of
   connected text */
end check-structure
```

Figure 3.2: The algorithm for computing and analyzing lexical graphs.

a domain-specific thesaurus, n only counts words in the domain, which makes the graph size manageable.[13]

3.4.1 Constructing suggestions for improving the text

As we have already shown, there are two tests we can perform to find some types of coherence problems.

The main component hypothesis is useful for short, tightly structured texts, or short to medium-length texts with large lexical density. (By short we mean no more than five paragraphs long.) The test involves finding the main component of the graph, or if it is not present, identifying those chunks of text that are not lexically linked with the rest of the text, i.e. with the largest connected part of the text, the one spanned by the candidate for the main component.

If we find a paragraph that is not lexically related to the rest of the text according to the main component test, we can suggest how to improve the lexical cohesive structure of such a text in one of two ways. First, we can advise the writer to simply delete the offending paragraph. This sounds like a drastic solution, yet there are cases when such an action would be desirable, for example, when the user accidentally pasted a wrong piece of text while editing, or if he went into a digression.

However, a drastic deletion is often not reasonable. After all, the writer did have a reason to include the isolated paragraph. The problem is that this reason may not be accessible to the reader. In this case, the writer needs to show more explicitly how

[13]In addition, we could improve the algorithm considerably by taking advantage of the fact that repeated lexical items will all have the same lexical relations with other items. Since repetition is a very common lexical relation, this improvement, while not changing the asymptotic complexity, will still result in a faster run time in practice.

the paragraph relates to the rest of the text. The simplest way to do this is to include a transitionary text that links the offending chunk with the rest of the text.

The transition might be a single sentence, or a full paragraph, or something in between. The important function of this added text is to lexically link the formerly lexically unrelated parts. Care must be taken in adding new text to select words which are lexically related to other parts of the text.

For longer texts, as well as for texts with low lexical density, the main component hypothesis is too restrictive, and hence needs to be relaxed. We find the central paragraph hypothesis more useful in such texts.

If the text lacks the central paragraph, this may again be an indication that the coherence connection between the unconnected chunk and the rest of the text may not be obvious to the reader. The remedy is again either to reconsider the need for including the unconnected chunk, or to connect it to the rest of the text via some transition.

When the central paragraph hypothesis is applicable, the transitionary text is more likely to be a paragraph. As a result, not only the shape but even the size of the collapsed graph changes.

Regardless of which hypothesis was used to find the site of the coherence problem, once the text is corrected, it needs to be re-analyzed to make sure that the correction did not introduce any new problems.

3.5 Analyzing texts using lexical cohesion — some examples

Now that we have explained the method in some detail, let us apply it to some examples[14]. First, we will see how the method finds the sites of incoherence problems for texts that are marginally acceptable. Later, we will see the limitations of the method, by examining those examples which were misclassified. Since we have also conducted an experiment with human judges, the examples used in the experiment together with their lexical analyses can be found in Appendix B.

This chapter also contains diagrams of lexical graphs for various texts, so we will explain our notation here. Each node of the lexical graph is represented as an oval that contains the word as it occurs in the text, preceded by two numbers. The first one is the paragraph number in which the word occurred, and the second one is the sentence number. Technically, we should also show the root as it occurs in the thesaurus. For simplicity's sake, we have omitted displaying the root. All the words that occur in the same paragraph are grouped in a rectangle that is labeled with the paragraph number. The lexical links are represented as lines linking the ovals.

For the collapsed lexical graphs, we use rectangles as nodes. Each rectangle is labeled with the paragraph number. The lexical bonds are shown as lines connecting the rectangles.

[14]The reader may find it valuable to consult our thesaurus at this point (Appendix C).

3.5.1 Applying the method to a text with one small coherence problem

As our first example, we will analyze the text shown in Figure 3.3. In this short text, we find a small coherence problem: it is not clear how the first paragraph fits with the rest of the text. The problem is not a very serious one, since it is quite possible for a reader to come up with a semantic interpretation in which the link between paragraph 1 and the rest of the text is understood. However, this requires the reader to make some assumptions that may not be warranted. One such assumption might be that the concern for the financial well-being of Canadians prompts them to manage their investments for profit. While not unreasonable, it is not present in a text, and it is impossible to tell if the assumption is valid. Furthermore, it is not clear if that is the connection the writer had in mind. Clearly, the text would benefit from making the connection between the two loosely connected parts more explicit so that the assumptions are not necessary.

Figure 3.4 shows the lexical graph of the text in Figure 3.3. As we can see, the first paragraph is not linked to the remaining paragraphs in the lexical graph. Thus, this text has no main component. Hence, we have correctly identified paragraph 1 as the site of incoherence problems.

Since the text is only three paragraphs long, the lack of the main component is a good indicator of text incoherence. Still, we might decide to analyze the text further, by constructing its collapsed lexical graph. The resulting graph, shown in Figure 3.5 indicates that the text lacks the central paragraph, which is a further clue that confirms our earlier diagnosis, and paragraph 1 is not connected with the rest of the text, as expected.

Next, we will look at the text in Figure 3.6. It is a short simple text, and again

As governments take an increasingly hard look at universal social programs and growing deficits, Canadians are becoming more and more concerned about their ability to meet their individual income needs during retirement.

Rebalancing your portfolio is a powerful disciplinary tool that allows you to manage your investments for profit, by selling high and buying low. For example, you may have decided originally that an investment mix of 60% in stocks and 40% in bonds met your objectives.

However, due to a prosperous period in the economy, you find that the stock component of your portfolio grows to 75% while bonds now represent only 25% by value. In this case, we would advise you to take profits and rebalance back to your original mix so that when the economic cycle reverses — as it always does — you will be positioned to take advantage of the change.

Figure 3.3: A sample text with some coherence problems

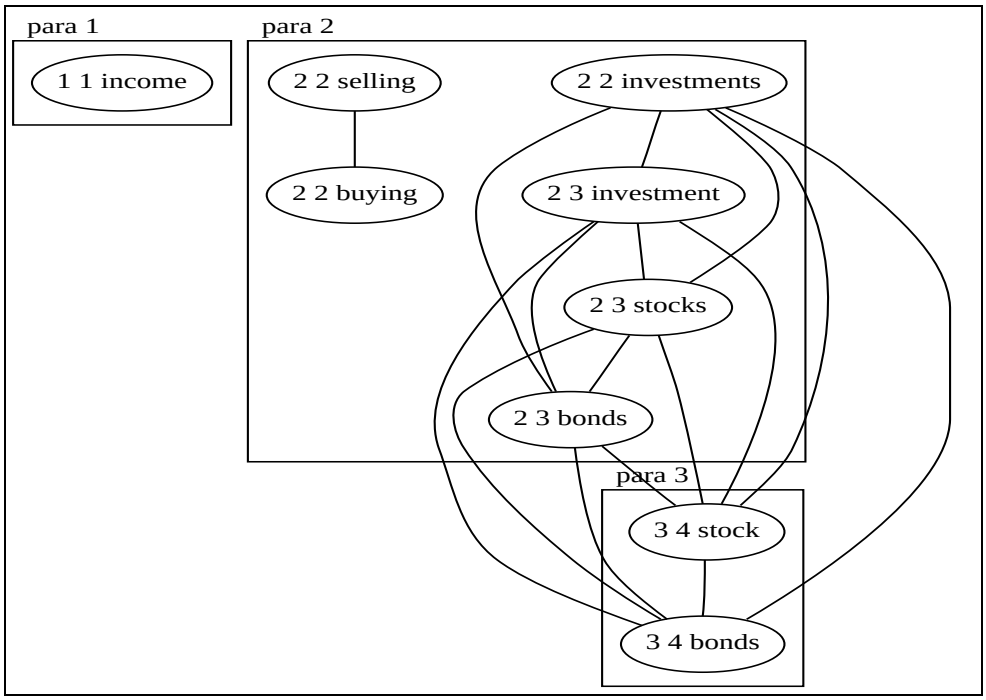


Figure 3.4: The lexical graph for the text in Figure 3.3. The first paragraph is not represented in the largest component of the graph.

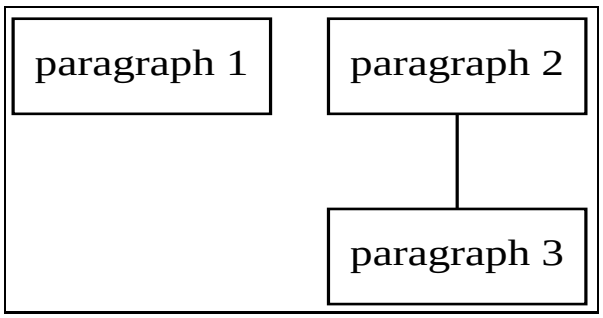


Figure 3.5: The collapsed lexical graph for the text in Figure 3.3. The first paragraph is not linked to other paragraphs.

it is not completely incoherent. However, the impression the reader gets is that the writer was trying to say too much in the small space, and so the text is crammed with information that is difficult for the reader to connect. Specifically, it is not obvious why insurance is important to financial well-being. A knowledgeable reader might be able to figure it out, but as it is, the text would benefit from some improvements.

Let us now analyze this text. Its lexical graph is shown in Figure 3.7. As we can see, the graph consists of several components, but no one single component touches all three paragraphs of the text. In other words, the main component is missing.

There are two components of the graph that touch two paragraphs each. The first one consists of the words *invest, invest, investment*, and touches paragraphs 1 and 2. This component leaves out paragraph 3, suggesting it as the site of the coherence problem.

However, another component that touches two paragraphs also exists. It consists of the words *insurance, insurance, whole-life, whole-life, insurance, term, term, insurance*. It touches paragraphs 1 and 3, leaving paragraph 2 out and suggesting it as the possible site of coherence problems. Hence, a viable alternative suggestion for

the coherence problem is paragraph 2.

Let us examine the text again. It is not clear what the focus of the text is, investing or insurance. Or possibly the writer intended to tell us about both — we cannot infer the writer's intention from lexical analysis alone. So it seems logical to suggest several possible ways the text could be improved, so that either of the two candidate components become main. One fix might involve changing paragraph 3 so that it explains how insurance relates to investing. A sample fix is shown in example 32.

(32) Before you build a nest egg so that your investments alone can cover your emergency needs, you will need adequate insurance. There are two types, term and whole-life. Typically, whole-life insurance is to be avoided. Choose the cheaper term insurance that will cover all your needs.

In this way, we have added a sentence explaining how insurance is related to investing, which makes the text more coherent. We have also linked the third paragraph to the component that originally touched only paragraphs 1 and 2. Thus, the new text has the main component and hence would be judged coherent by our method.

Note that the original text does have the central paragraph, paragraph 1 (see Figure 3.8). This indicates that the coherence problem of this text is less severe and a lot more subtle than the abrupt shift we have seen in text 3.3.

Consider now the text in Figure 3.9. It consists of 5 paragraphs, four of which are tightly lexically connected. However, paragraph 4 has no lexical links to any other paragraph. At the same time, it is not exactly clear how this paragraph relates to the rest of the text.

Now, let us turn our attention to the lexical graph of this text, shown in Figure 3.10. In this case, the text fails our coherence test, having no main component.

One of the most important decisions you can make regarding your future is to create a financial plan. It consists of two parts. One is to save and invest money. The other, to have adequate insurance. This will allow you to achieve financial independence, and a lot sooner than you think.

Generally, it is recommended that you invest about 10% of your pretax income. If you arrange for automatic withdrawal of investment money immediately after your paycheck is deposited to your account, you will find that the savings accommodate painlessly.

Many people don't have adequate insurance. There are two types, term and whole life. Typically, whole life insurance is to be avoided. Choose the cheaper term insurance that will cover all your needs.

Figure 3.6: A sample text with some coherence problems

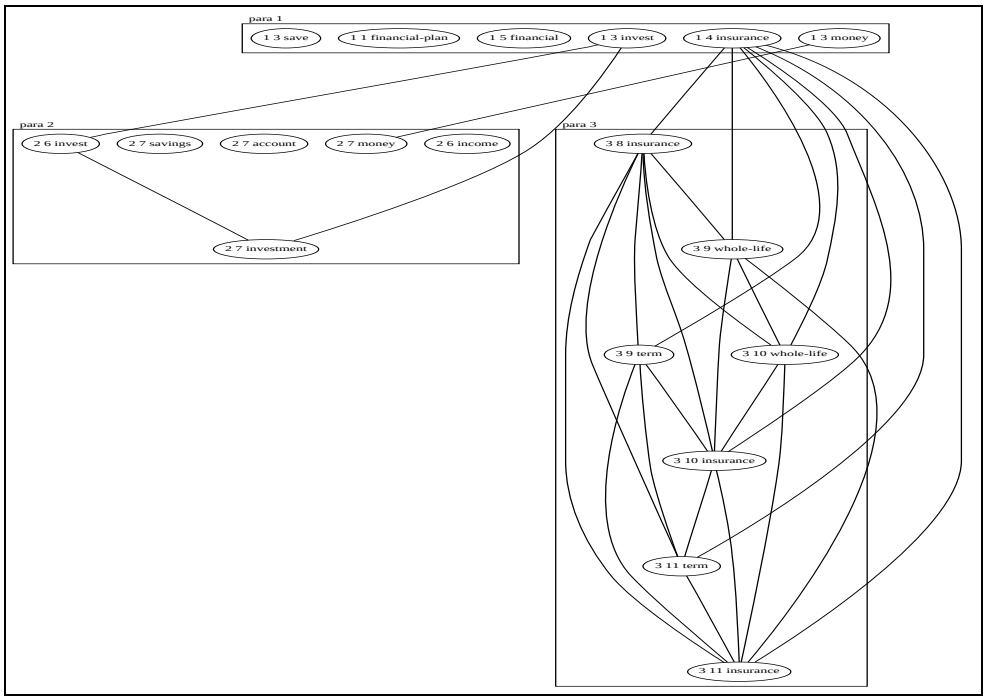


Figure 3.7: The lexical graph for the text in Figure 3.6. The graph lacks the main component.

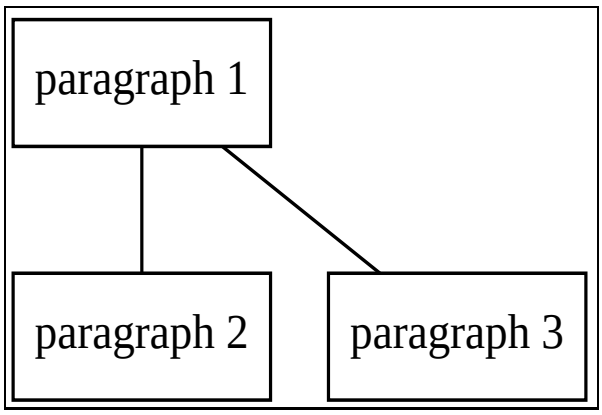


Figure 3.8: The collapsed lexical graph for the text in Figure 3.6.

The problem is a subtle one, but it is important to recognize, as our analysis has done. In constructing our corpus, we found many texts that exhibit the same local coherence lapse, which suggests that the problem is common. In fact, the problem is common enough that having a mechanism for detecting it could be worthwhile.

3.5.2 Applying the method to texts with no problems

Consider now a text created by correcting the small coherence problem in the text shown in Figure 3.9. We have improved the coherence of the text by removing the paragraph that seemed irrelevant to the rest of the text. In addition, we have improved an awkward transition between paragraphs 2 and 3. The new version reads more smoothly. The modified text is shown in Figure 3.11.

Let us now briefly analyze this new version. The lexical graph of our new text is shown in Figure 3.12. Since we have removed the paragraph that was disconnected, the largest component of the original text now touches all the paragraphs. Note that the text now has the main component, which indicates that the original low coherence

When developing your financial plan, you first need to consider whether you're an investor or a saver.

Investors look to invest some money for the longer haul. They want capital growth over time, some income, and/or tax-free income. Investors have adequate reserves to meet their current needs and can, therefore, stay invested in the market. They are also willing to ride out any short-term market fluctuations and invest instead for long-term growth potential to outpace inflation over time.

Savers need to focus on current or short-term needs. Their primary concern is preservation of their capital. Savers also seek liquidity for ready access to cash when necessary.

Your decisions will be affected by your time frame, your objectives, and your tolerance for risk.

You and your financial adviser can determine whether you are a saver or an investor, by looking at your needs and goals. Once this determination is made, your advisor can help you build an investment portfolio that's right for you.

Figure 3.9: A sample text with some coherence problems

was improved.

The texts that we have examined so far were intentionally kept short, to make the lexical analysis process clear. Now, consider a somewhat longer sample text shown in Figure 3.13. This example illustrates the case of texts which are long enough to be analyzed in terms of the collapsed graph alone.

This text is five paragraphs long, and has a logical structure. The first paragraph is an introduction, followed by three paragraphs, each describing a different precious metal in the context of investing, and the final paragraph contains conclusions. Note that there is a rather rich set of lexical relations between the domain-specific words in the text.

The lexical graph for this text is shown in Figure 3.14 and the collapsed graph is in Figure 3.15.

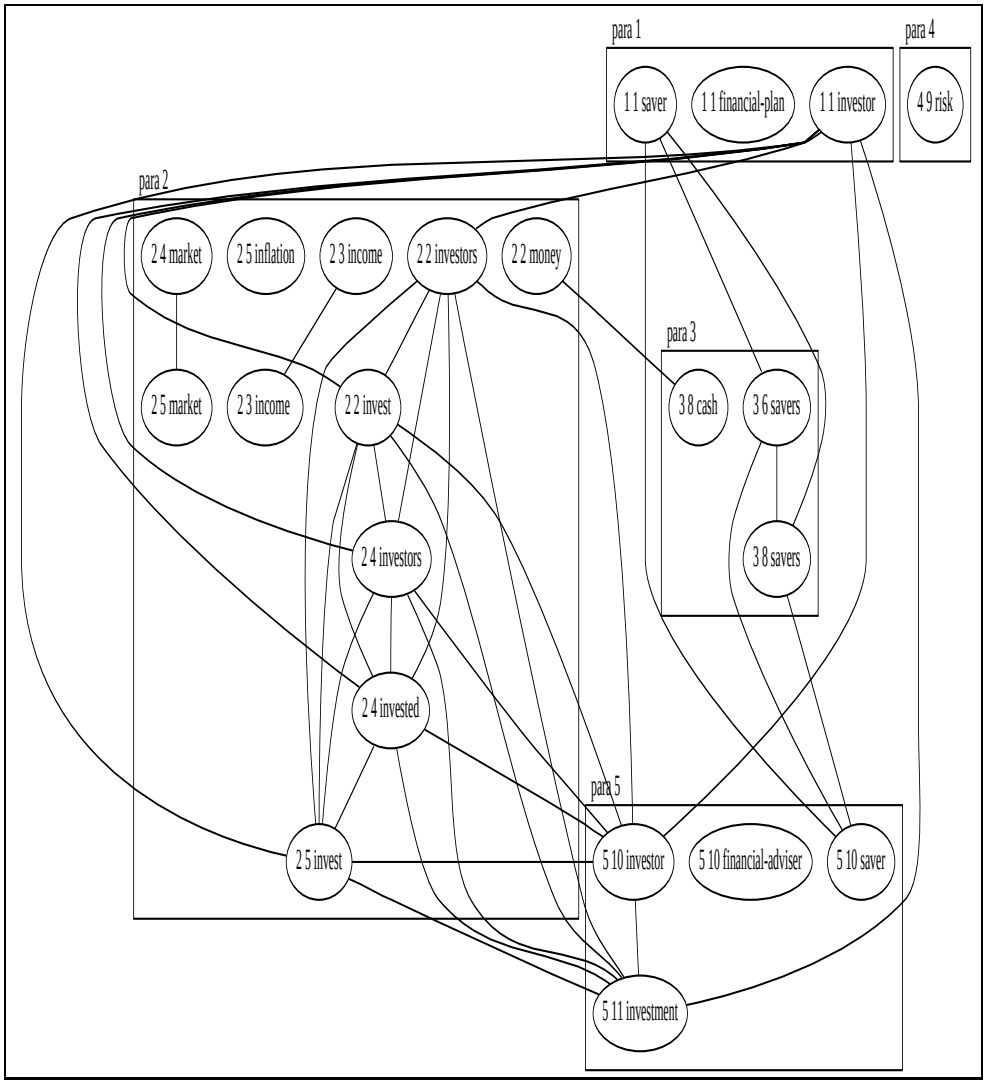


Figure 3.10: The lexical graph for the text in Figure 3.9. The graph lacks the main component.

When developing your financial plan, you first need to consider whether you're an investor or a saver.

Investors look to invest some money for the longer haul. They want capital growth over time, some income, and/or tax-free income. Investors have adequate reserves to meet their current needs and can, therefore, stay invested in the market. They are also willing to ride out any short-term market fluctuations and invest instead for long-term growth potential to outpace inflation over time.

Unlike investors, savers need to focus on current or short-term needs. Their primary concern is preservation of their capital. Savers also seek liquidity for ready access to cash when necessary.

You and your financial adviser can determine whether you are a saver or an investor, by looking at your needs and goals. Once this determination is made, your advisor can help you build an investment portfolio that's right for you.

Figure 3.11: A corrected version of the text in Figure 3.9.

Let us now examine the collapsed graph more closely. The graph is fully connected, mostly by means other than repetition or pleonymy. In fact, the only lexical bond that involves repetition is the bond between paragraphs 1 and 5. It contains of the repetition of the lexical item *precious metal*.

The bond between paragraphs 1 and 2 (and also between 2 and 5) is based on the *is-a* link between *precious metals* and *gold*. Similarly, the bond between paragraphs 1 and 3 (and between 3 and 5) is based on the *is-a* link between *precious metals* and *silver*, and the bond between paragraphs 1 and 4 (and 4 and 5) on the *is-a* link between *precious metals* and *platinum*.

Paragraphs 2, 3 and 4 are bonded based on the systematically classifiable relations among *gold*, *silver*, and *platinum*.

Many more examples, coherent ones processed correctly, and ones with coherence problems which are detected, can be found in appendix A.

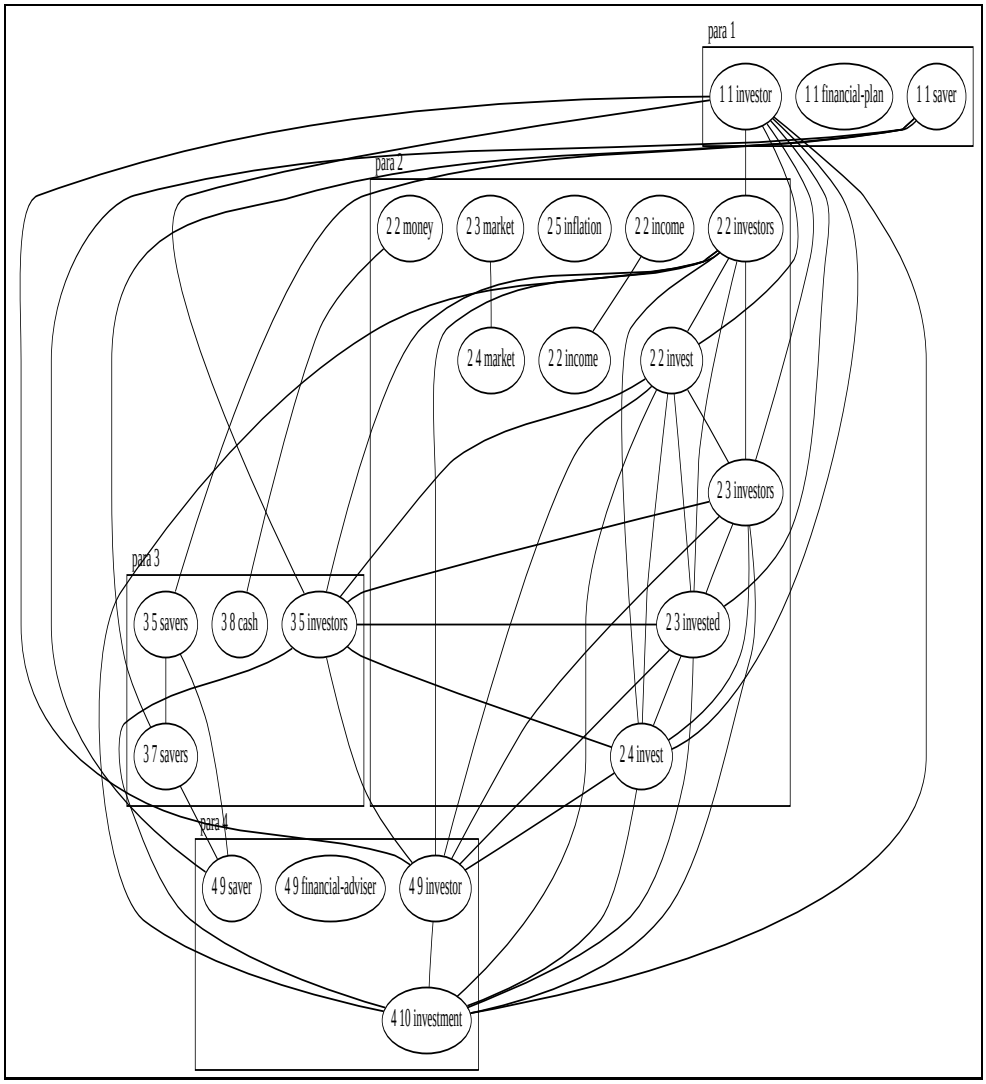


Figure 3.12: The lexical graph for the text in Figure 3.11.

With the recent bull market, precious metals have fallen out of favour. Investors prefer to put their money into the stock market, ignoring the possibility of market collapse. A contrarian therefore should take a close look at the neglected precious metals market.

Gold has always been held in high regard for its beauty and often used for ornaments. It has also been used as a security of choice to shelter the investor from high inflation. The inflation will return sooner or later, and then gold will again be the security of choice.

Less glamorous but perhaps more important is silver. It has also been used for jewelry, but its primary use has been in industry. For this reason, silver will always hold its value well.

Finally, you should consider platinum. The most expensive of the three, it has unfortunately not performed well recently. Still, the price might be bottoming out, and the time will be soon to add platinum to your portfolio.

With the prices for precious metals falling, you might be understandably nervous to buy now. But that's exactly what successful contrarians do — buy when everyone else is afraid to.

Figure 3.13: A longer coherent text.

3.6 Limitations

As we have seen already, the lexical analysis method is a useful indicator of potential incoherence in text that often can stand alone. For some texts, however, the lexical analysis alone cannot determine incoherence. One reason for this is that there are many other aspects to cohesion, not only lexical. Moreover, text coherence is influenced by other factors in addition to cohesion. For example, text constituents can be related semantically ([Hobbs, 1976], [Mann and Thompson, 1983]) which is in a different dimension than lexical cohesion. In some situations, we would like several different modules to work together to evaluate coherence from different points of view. For example, we can envisage one module which would perform the analysis of the underlying semantic relations which contribute to coherence. Another module might keep track of the clue words used in the text as an indicator of text coherence. In

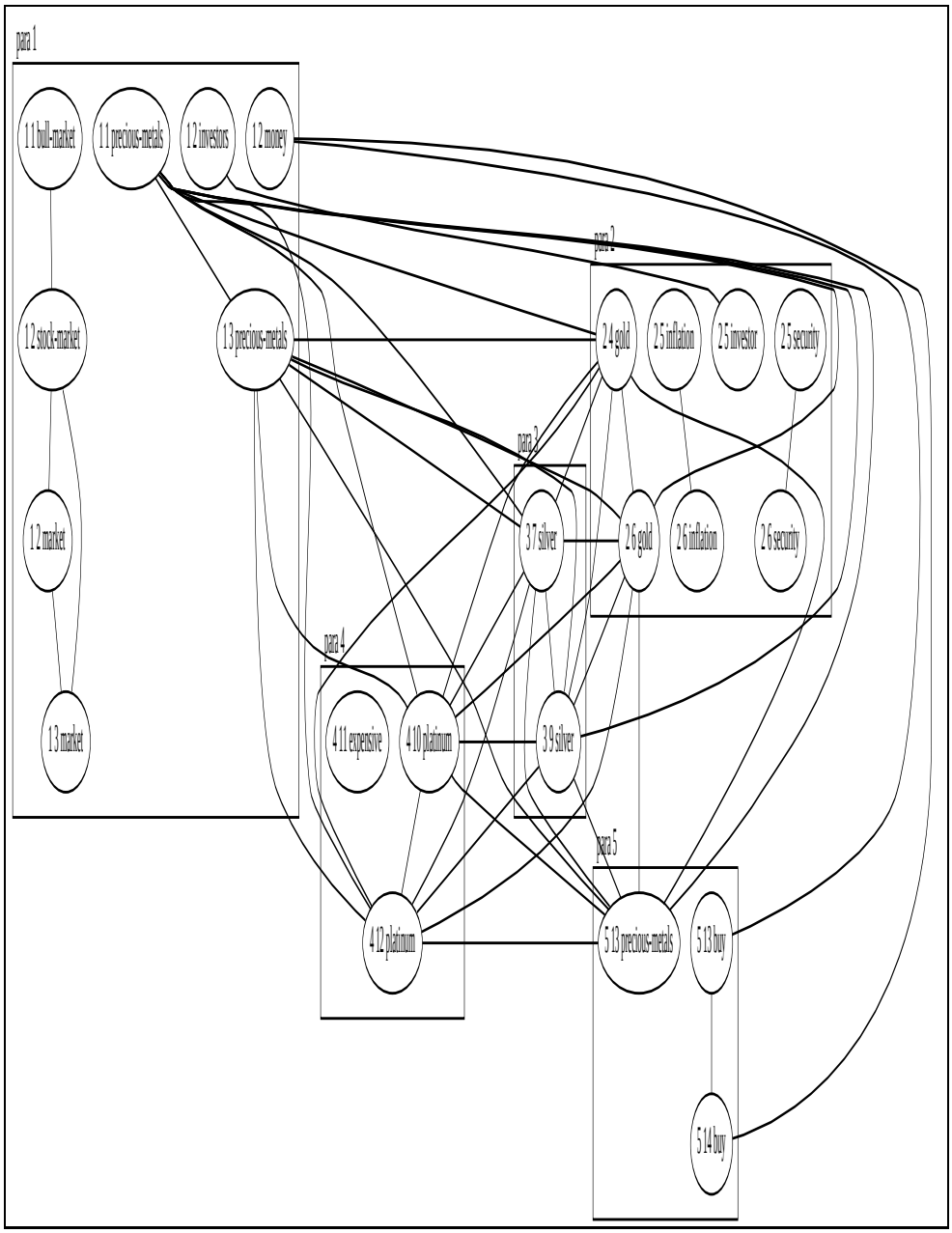


Figure 3.14: The lexical graph for the text shown in Figure 3.13.

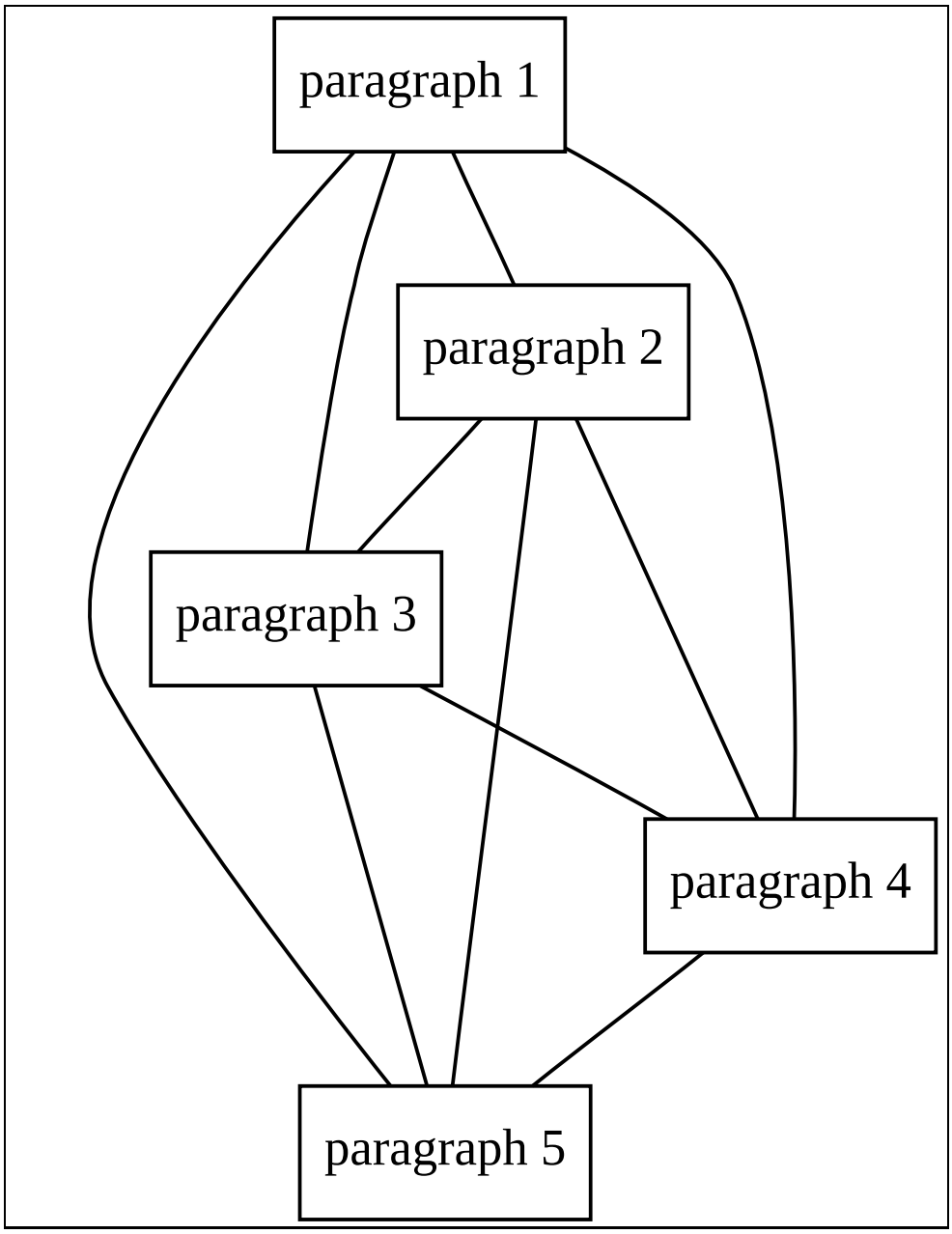


Figure 3.15: The collapsed lexical graph for the text shown in Figure 3.13.

section 7.1 we will revisit this idea and give some more detail. In this section, we will see some examples that would benefit from more extensive analysis.

3.6.1 A coherent text for which the lexical cohesion analysis is incomplete

In some situations, the analyzed text is quite coherent, but the lexical analysis does not recognize this for various reasons.

Consider the text in Figure 3.16. In some sense, this text is not very well written, since the last paragraph constitutes a clumsy jump from the rest of the text, making the reader stop and figure out how the last paragraph relates to the text read so far. Still, even with that choppiness, the text is readable, and makes sense. In other words, this text is borderline coherent. The problem with it is not with coherence, but rather with cohesion.

In spite of this, the lexical analysis labels the text as having coherence problems. Since the last paragraph is not represented in the largest component of the lexical graph (see Figure 3.17), that last paragraph is identified as the potential source of incoherence. This is further confirmed by collapsing the graph. As we can see, the last paragraph is not lexically connected to the rest of the text, indicating some coherence problems.

Even though the method misclassifies this and similarly structured texts, it can still offer some constructive suggestions to the user. More precisely, the text, although coherent, is not very cohesive. The logical link between paragraphs 2 and 3 is not immediately obvious. Strengthening the cohesion of the text then would make the link explicit, improving its quality.

The brokerage houses strive to create a perception that stocks and bonds will provide security for you. The perception is that they are trustworthy, financial professionals and all you have to do is take their expert advice and give them your money and they will make money for you. The reality is that these members of the financial community wish to use your money to make money for themselves. There is nothing wrong with the financial community making money off of your money if full disclosure on their part is given, and they fully explain to you the potential of profit against your potential loss. Full disclosure (total and accurate information) is very hard to come by, even when you know the questions to ask.

And make no mistake about this. A great many members of the respected brokerage firms deal in half-truths and embrace an unwritten law which says, Pass the losers on to the unsuspecting public.

How can you protect yourself from these predators? There is no simple answer, but the closer you can be to the principals of a company and the company itself, the better you can evaluate it. Relying on others for accurate and timely information is risky.

Figure 3.16: A coherent text for which the analysis is incomplete. The graph of this text is shown in Figure 3.17.

One way to improve this text is to change the last sentence so that paragraph 3 is connected to the preceding paragraph. The revised last paragraph is shown in Figure 3.19. The resulting text is not only easier to read and more coherent, it is also more cohesive. In fact, it now passes our test, since the revised text now has the main component (see Figure 3.20).

3.6.2 A text with coherence problems that are not detected

As we have seen, the lexical analysis method works well in many cases. Moreover, as we have seen in the previous section, there are cases when the analysis itself is weak but the results may still be useful.

In this section, we will see an example of a text that has a serious coherence

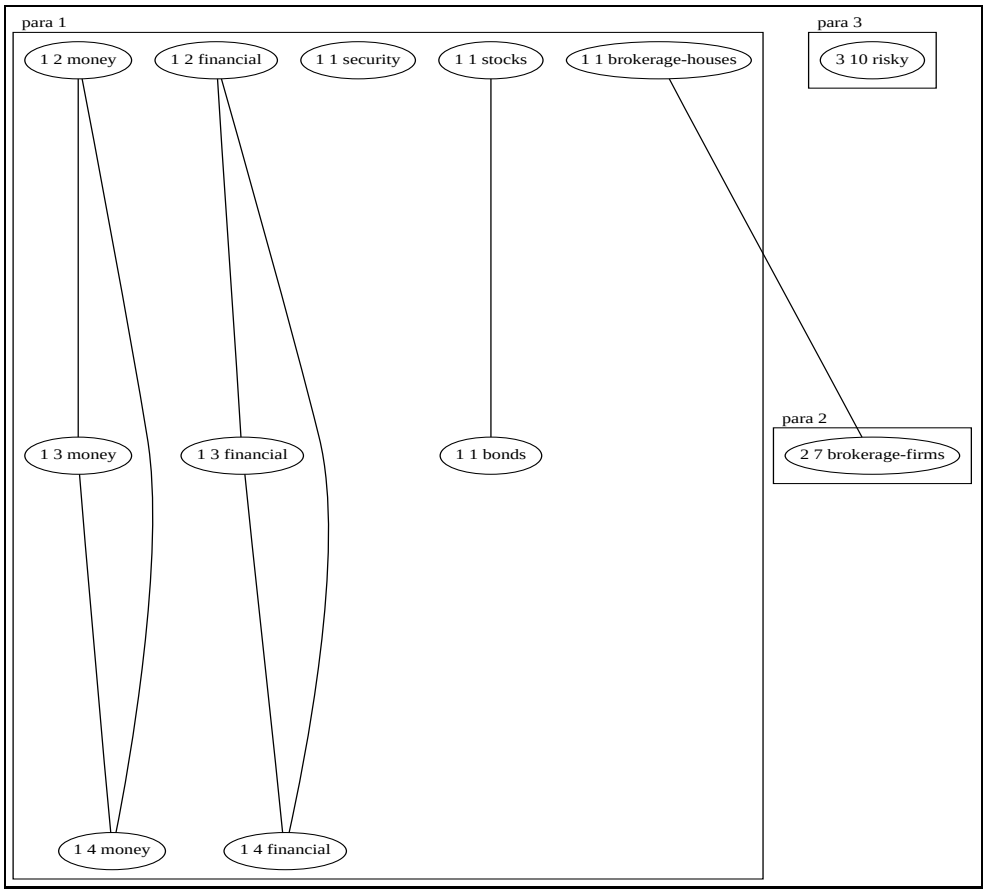


Figure 3.17: The lexical graph for the text in Figure 3.16.

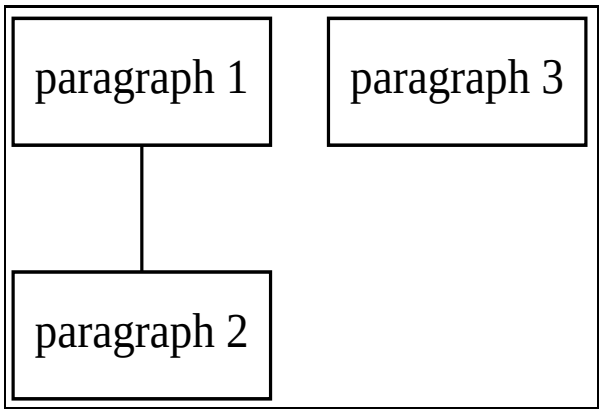


Figure 3.18: The collapsed graph for the text in Figure 3.16.

How can you protect yourself from these predators? There is no simple answer, but the closer you can be to the principals of a company and the company itself, the better you can evaluate it. Relying on brokerage firms for accurate and timely information is risky.

Figure 3.19: A revised paragraph 3 for the text shown in Figure 3.16.

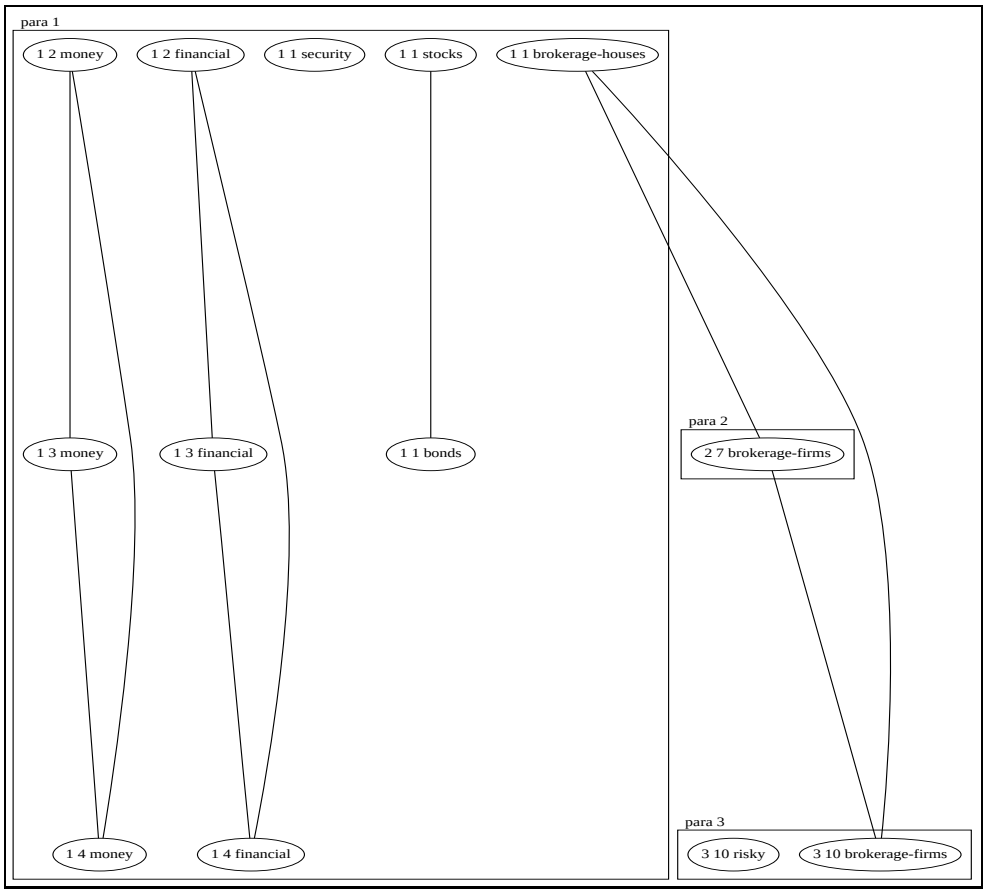


Figure 3.20: The lexical graph for the text in Figure 3.16 with the last paragraph replaced by the one shown in Figure 3.19.

problem that is not detectable by the lexical analysis alone.

Consider the text in Figure 3.21. This text was artificially constructed by placing an unrelated paragraph at the end of a section of a coherent text. Clearly, the last paragraph is not linked to the rest of the text. It is difficult to find an interpretation in which the text would make sense as it is written. However, there are some lexical items in the original text that are accidentally linked to the items in the last paragraph. As a result, the lexical analysis cannot find the coherence problem, although it clearly is present in the text.

The collapsed lexical graph for this text (Figure 3.23) clearly shows the presence of the central paragraph, paragraph 1.

Figure 3.22 shows the lexical graph of this strange text. Not only does it contain the central paragraph, but worse, it also has the main component.

These kinds of coherence problems that are difficult to detect do occur in real texts in less severe form, and it is difficult to guard against them in practice. This particular text was constructed by hand. However, most of the texts shown in this thesis were drawn from real texts, where the author created the text in order to communicate with an audience. One can speculate that in such texts, the problems illustrated in Figure 3.16 would arise infrequently.

We acknowledge that the lexical analysis method is not always reliable on its own, giving a passing grade to some clearly badly constructed texts. We will discuss how to handle such cases in section 7.1.8.

Let's use an example to demonstrate the types of investments. For instance, pretend you are going to start a lemonade stand that you call Lemo. You need some money to get your stand started. You ask your grandmother to lend you $100 and write this down on a piece of paper: "I owe you (IOU) $100, and I will pay you back in a year plus 5% interest." Your grandmother just bought a bond (IOU) by lending money to your company named Lemo. To get more money, you sell half of your company for $50 to your brother Tom. You put this transaction in writing: "Lemo will issue 100 shares of stock. Tom will buy 50 shares for $50." Tom has just bought 50% of the shares of stock from Lemo.

You sell $500 worth of lemonade. Business is good. Your costs for setting up the stand are $150, plus you pay yourself $100 for the hours you work. The company makes profits of $250.

After one year, from the $250 profits, you pay back your grandmother $100 plus $5 interest. You pay $20 to Tom and yourself, shareholders (a fancy name for owner). In business, the $20 paid to the owners is called a dividend. You decide to put the dividend money in the bank. Banking the money is a short-term investment.

If you are investing a small amount of money monthly with a long term in mind, you might want to choose a balanced fund. As your investment grows, you could move a portion to some other fund in the fund family. If you have a larger lump sum to put in, you might split it between two or three different funds - perhaps a bond fund, a growth equity fund, and an international growth fund.

Figure 3.21: An incoherent text for which the analysis does not find the problem. The graph of this text is shown in Figure 3.22.

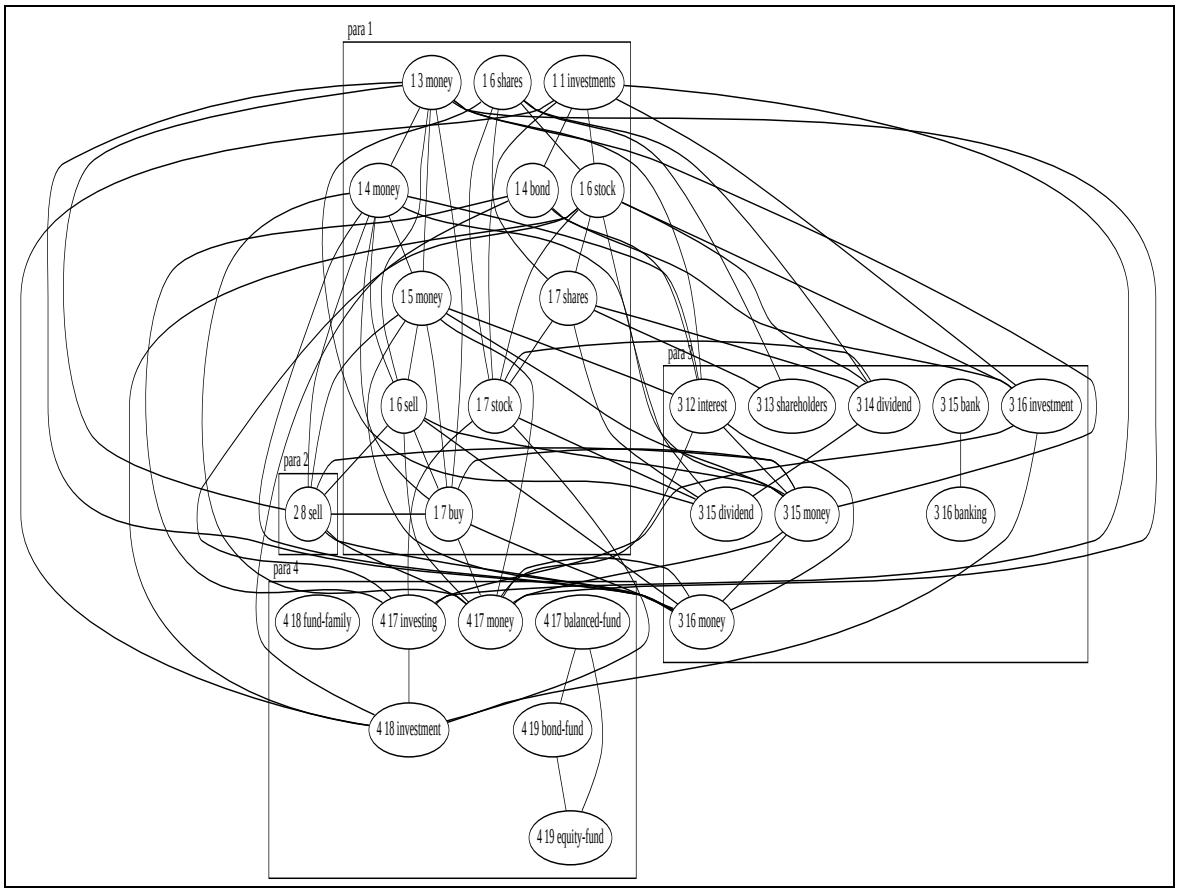


Figure 3.22: The lexical graph for the text in Figure 3.21.

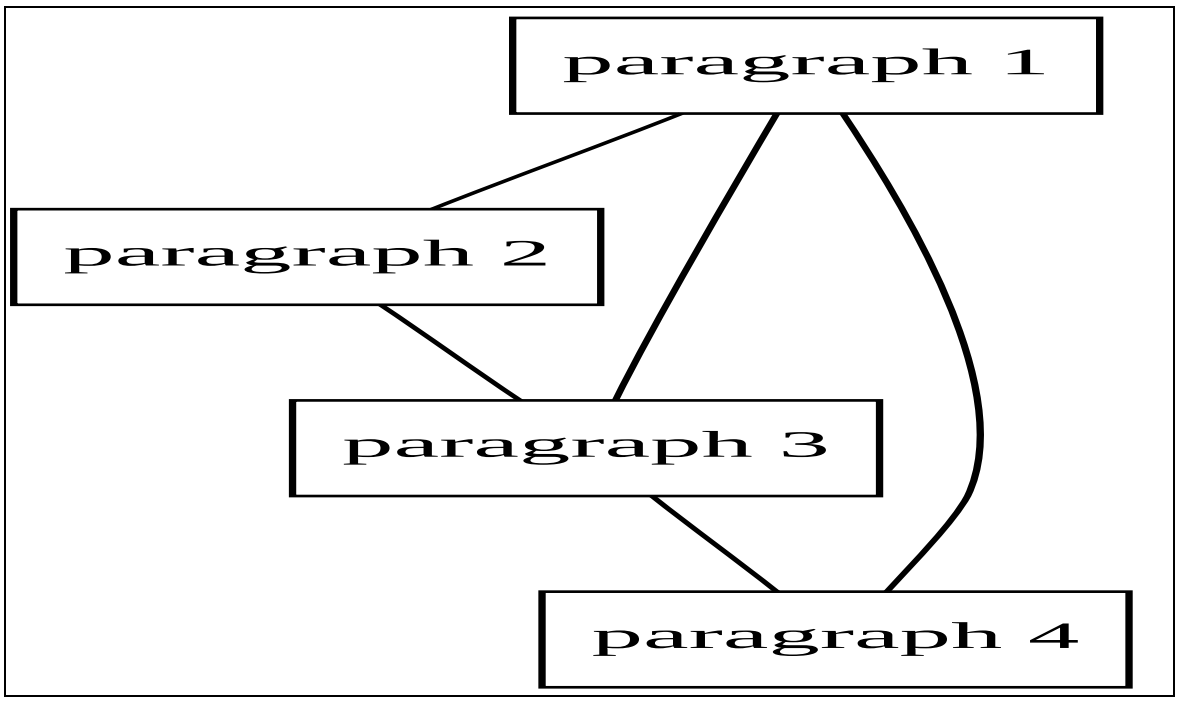


Figure 3.23: The collapsed lexical graph for the text in Figure 3.21.

3.7 Comparative advantages

3.7.1 Very long texts

One advantage of our model for lexical cohesion analysis is its applicability to texts of any length, in part due to the use of the collapsed graph to examine sites of possible incoherence. In this section, we discuss an example which is considerably longer than several paragraphs.

The text we have chosen is over 40 pages long, so it would be impractical to show in a thesis.[15] Therefore, we will describe the full lexical analysis but only show fragments of the text.

The text describes a program for managing personal finances. The program consists of 13 steps. Because the text is so long, and because it also contains graphics, for technical reasons it was divided into separate pages, one step per page, plus one page for the introduction. However, the text is clearly intended as one coherent whole, and so we treat it as such.

Since the text is long, we constructed a collapsed lexical graph and applied Hypothesis 2 to it, in order to examine the text's coherence. According to this hypothesis, the text might have some coherence problems if there is no central paragraph. With this example, we indeed found no central paragraph.

According to Hypothesis 2, the sites of potential incoherence problems are identified at those text fragments which are isolated from the main body of the text. For this particular example, we found a few such sites.

[15] An interested reader may find it in the electronic version of the thesis available via ftp from University of Waterloo archives, or by visiting the World-Wide Web site that hosts this text at www.fool.com/13Steps.htm.

> Here's our solution to baseline accountability: Any money that you have to invest for three years or longer should NOT underperform the Standard & Poor's 500 (S&P 500) over that 3-year period. If it does, you've blundered, because you can get average market performance out of an index fund without doing ANY research and without taking on significant risk.
>
> Stick close to those expectations; prepare and aim to beat them; know why you have or haven't; and laugh at the business pages of our national newspapers and magazines, which give plenty of room for "professional" predictions but don't typically allow even a day each year for reviews of bottom-line performance—including the deduction of all trading costs. Not a chance.

Figure 3.24: A marginally coherent paragraph that is not lexically related to the rest of the text.

For example, Figure 3.24 shows a fragment of the text where the second paragraph is not lexically connected with the rest of the text. When we read this paragraph in its intended context, we can see that it is borderline coherent. In other words, it does make some sense, but it is a little confusing. We don't know exactly what expectations the text refers to, although we can guess it has to do with the performance of the Standard & Poor index. So, we have found a site that could and should be improved.

We have also found another site of potential coherence problems at the beginning of the last step in managing personal finances: keeping informed. This step is nothing but an advertisement for the hosting site, and so it is connected with the personal finances only in that this is what the site does. But some paragraphs, such as the paragraph shown in Figure 3.25, are intended to entice the user to visit the site often, and contain no financial advice at all.

We intended for our algorithms to work on texts of any length. As shown in the above long example, our investigation of longer texts indeed shows some value to applying the proposed analyses to these kinds of texts.

> Like many media mavens and education experts, we believe that while a picture may be worth a thousand words, the right thousand words can change the world. You want that in technical terms? Text-based learning works better than image-based learning. Our online materials – the stuff that can help you make money – it's words, sentences and paragraphs, not pictures, not videos. And we're *not* sorry, because we think that words work better.

Figure 3.25: The lexically unconnected paragraph from the section that advertises the Motley Fool investment site.

3.7.2 Alternate thesauri

In section 3.2 we discussed why we decided to use a domain specific thesaurus in our system. We show the comparative advantage of this kind of thesaurus by demonstrating what the algorithm would produce with a different kind of thesaurus.

For example, consider the text shown in Figure 3.26. This text, written by a second language learner, was analyzed using the Kipfer thesaurus described in section 3.2. At 728 words and 9 paragraphs, the text is not particularly long. In addition, many words are ubiquitous and so were discarded. This left us with 437 words that participated in the analysis.

Using the general-purpose thesaurus with the algorithm described above, our method has found over a thousand lexical links[16]. The number itself is surprising, making the method unusably slow. But there is a more fundamental problem: many of these links do not carry any lexical cohesive information. In fact, many pairs of words (e.g., *clock* and *day*) that the thesaurus claimed to be related via one of our chosen lexical relations shouldn't be related at all. But the method found them re-

[16]Because of the massive amount of lexical links, we cannot show the actual lexical graph here — the result is a black page.

After a good night of sleep my alarm clock will wake me up, or if not my mother. Barely walking I will crawl out of my room which is usually a mess to my bath room and turn on the water to wash my face and teeth. Although my room is a mass I still like it its very worm and cool looking. Meanwhile, mom will prepare me a tasty breakfast and a healthy lunch to school. My breakfast varies from morning to morning. Sometimes she will prepare me scramble eggs with toast and some other day there will be a bowl on the table. Which is also fine.

Then its off to school, I kiss my mom goodbye and jump to my brand new 95 Honda Civic, which I really love. I just bought that car so I have to take good care of it. My mom was good enough to me to help me pay the bills because I only work part time and the money that I earn is never enough. Even though I take the express way to get to N.E.I.U. it takes me about 40 minutes. I take the Stevenson and Kennedy express but it is obvious that at that time there is a traffic jam.

My first class is an English class which I am writing this paper for. So far I enjoy it is very helpful because I am learning to operate a computer and how to write papers. Writing is important not only in my education, but in everyone's too. Next is music and I play percussion for three years now and things are getting better every time i sit down to practice. After music I go to math class. I've always hated math and I am pretty sure that I will never like it. This hour is the versed time of my entire day. I could never understand the math problems or its concepts, therefore I am doing very poorly in that class.

Finally I have a break, I eat my lunch prepared by my mother it is usually a turkey sandwich, by the way my favorite, then I do my homework from my favorite class math, NOT.

In the afternoon I go to my last class of the day , Freshman Seminar. In that class they teach us how to survive your first year in college.

Since I live far away from school right after classes I go to my girlfriend's house it is more convenient for me because at 6:00 PM I have to go to work which is located closer to her house. My girlfriend Marta which I love very much, usually helps me with the homework. She is very bright and smart girl, knowing three languages Polish, English and Spanish. In her high school she was a honor student and a great athlete she was a captain of cross country team. Her mother is a great cook so there is always a tasty dinner. Her family treats me really good they take care of me like I would be one of the family.

After some rest I go to work. I am a karate instructor. I teach kids, teenagers, seniors how to defend themselves. I really enjoy being with people and teach them what I've learned from my instructor, Sensei Samitowski. I been doing Karate for almost eight years now. I have a first degree brown belt, which means I need one more promotion to black. Class are very aggressive and physical sometimes people do get hurt if they are not careful.

Usually I get home after 9:30 at night always tired and exhausted, who wouldn't after a long and busy day. I go downstairs to my sister apartment to see what's going on. I'll play a little with my niece Anita who is two and a half years old she is so adorable is hard to resist her. Later I take a shower eat something light and relax in front of T.V. My favorite show is the Late Show With David Letterman. That man is very hilarious and funny looking. He just make me laugh, which helps me relax.

I would spend the whole night watching T.V. but I know that there will be a next busy day to go bed. So I just go to bad and fall a sleep like a baby.

Figure 3.26: A text written by a second language learner and analyzed using the Kipfer general-purpose thesausus.

lated because of the presence of words such as *time*, which are related to a great number of words and which in turn caused loosely related words to show as related.

It is worth noting that removing such words from a thesaurus will not solve the problem. First, there are many such words, and second, these are legitimate words that aren't exactly ubiquitous, since they do carry information.

3.8 Chapter summary

As we have seen, the discourse structure approach for processing text coherence concentrates mostly on how the text is structured, paying less attention to the area of continuity. Conversely, the topic continuity approach concentrates on the aspects of continuity, while ignoring the structure of the text. However, it is advantageous to unify these two approaches, taking the best characteristics of each.

We see our work as a step towards that unification process. While the lexical cohesion analysis relies heavily on topic continuity for building the lexical graph, the conclusions we draw from it pertain to text structure.

At the same time, the process of collapsing lexical graphs relies somewhat on the text structure because it accepts paragraphs as units. The analysis of the collapsed graph yields information about problems with both topic continuity and discourse structure.

Chapter 4

Implementation

Now that we have seen how the model works, let us turn our attention to actual implementation issues. The main goal of the implementation was to demonstrate the practical application of our model.

The model was implemented in two parts. The modules that compute the lexical links and collapse the graph were written in C++ [Davis, 1994]. The user interfaces part was implemented using Tcl/Tk [Sastry and Sastry, 1997]. For drawing the graphs, we use the *dot* utility, and the resulting PostScript file is displayed using *ghostview*.

4.1 Input and output

The input to the system is the text file in ASCII. To make sure that the paragraph boundaries are intended, and not just accidental carriage return characters, each paragraph break is marked by the end of paragraph marker, $<p>$.

The output of the system consists of several pieces of information. First, the system produces a graphical form of the lexical graph. The graph is shown to the user, but may also be saved and printed out if desired. Next, the system performs the computation to find the main component if it is present. If it is absent, the user is notified via a message in the message window.

Third, the user is shown a picture of the collapsed lexical graph, with lexical bonds of different strengths displayed in various thicknesses. If the collapsed graph lacks the central paragraph, the user is also notified by a message in the message window. The unconnected paragraphs are easily identifiable from the collapsed graph picture in the graph window.

4.2 The user interface

The interface was implemented using Tcl/Tk. The sample screen is shown in Figure 4.1.

The interface window consists of the upper command bar and three display areas. The buttons on the command bar help the user choose the desired operations for the system to do.

Clicking the leftmost button, *File*, causes the drop-down menu to appear, which makes it possible to perform operations associated with files, such as creating a new file, opening an existing file, saving the file, and quitting. The file must be opened before it can be analyzed.

The *Test* button is intended for use in demos. However, this button does not currently have any functionality. The next button, *Thesaurus* allows the user to specify the thesaurus to be used for computing the lexical graph. Currently, we only

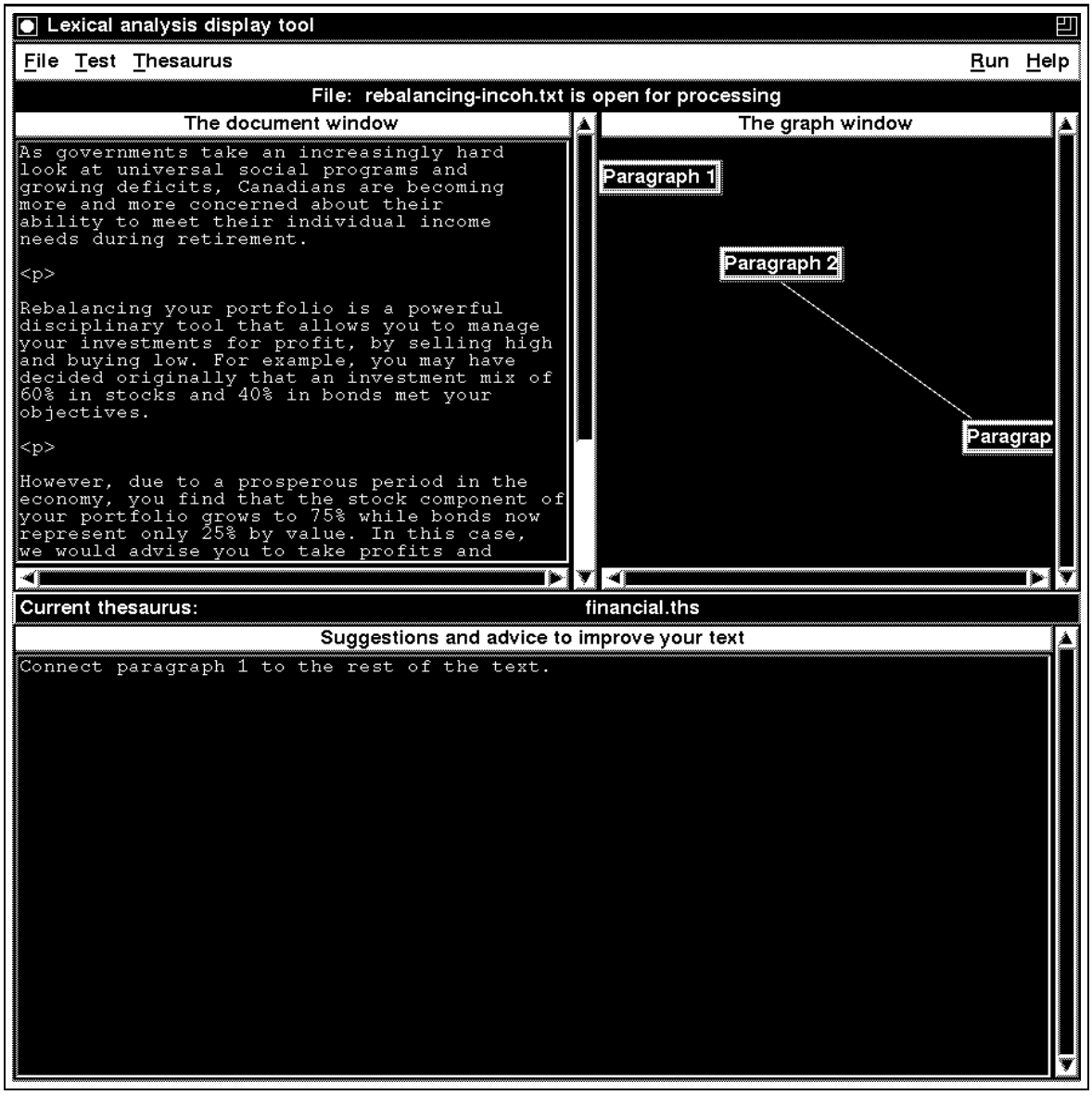


Figure 4.1: The interface screen for the lexical analysis software. The figure shows the output for the text shown in Figure 3.3.

use the financial thesaurus, but adding more thesauri is not a difficult task. In fact, we also have available the module that implements the Kipfer thesaurus.

The second rightmost button causes the system to perform the actual lexical analysis of the text. The text to be analyzed is taken from the text area on the left hand side of the window. For this reason, the system requires the text file to be open before it can be analyzed.

Clicking the final button, labeled *Help*, activates the display of help information.

The line just below the command bar displays the name of the ASCII file that is open for the lexical analysis. Below this line there are two areas placed side by side. The area on the left-hand side contains the edit window for the text. Here the user can make changes to the text. Once all the changes are made, the user can see how each change is reflected in the lexical graph by hitting the *Run* button.

The collapsed graph itself is displayed in the area on the right-hand side. We currently do not display the whole lexical graph, because the window is too small for a typical graph. Rather, we save the graph in the PostScript file so that the user can view it using another utility, such as *ghostview*.

Below these two areas there is a band that displays the name of the current thesaurus. Directly below it, there is the area for suggestions and advice to the user. It is here that the user can read about the results of the analysis, such as presence or absence of the main component or the central paragraph.

4.3 The modules

In section 3.4 we have already seen the algorithm describing the logic of the implementation, so we will not repeat it here. Our implementation followed the algorithm

very closely.

The program is divided into modules. The most important modules are: the thesaurus module, the word extraction module, the lexical graph module, and the user interfaces module. In addition, we also have a rudimentary morphological analysis module. The reason for having it here is to obtain root forms for words to be looked up in our thesaurus. The module receives words as they occur in the text, and removes the endings for the thesaurus lookup. All the processing is done by simple string manipulation.

One of the obvious advantages to have a thesaurus as a separate module is that we are then able to exchange one thesaurus for another without affecting the other parts of the code. In fact, this is exactly what we did when we were testing various thesauri.

The thesaurus module is responsible for the word lookup (see also 3.2). Because our thesaurus is rather small, and because we need the lookup to be fast, we store the thesaurus entries in an array, sorted alphabetically. For each word pair, the array contains the root form of the word, the index in the array to the word that forms the other element of the pair, and the code for the lexical relation. This way, the thesaurus contains some redundancies, but the advantage is that the lookup is fast. Consult appendix C to see the complete thesaurus.

The word extraction module is responsible for reading individual words from the text file. Each word is placed in lower case, and stripped of any endings to obtain the root form.

The lexical graph module performs all the functions that deal with graph building, collapsing, and analyzing. It is also responsible for interfacing with the drawing software.

As the words are extracted from the text, each word is looked up in the thesaurus. If its root form is found there, we add the word together with other information, into the lexical graph. And so, for each word we store the following:

- the word as it occurs in the text.
- the root form. This is so that we can speed up the graph linking later.
- the paragraph in which the word occurred, for collapsing the graph.
- the sentence in which the word occurred, so that we can tell each occurrence apart.
- the list of lexical links for that word. In the beginning, this list is empty. The actual list is computed after the words have been collected.

We construct the lexical graph only after all the words have been collected. In order to do so, we again follow the algorithm presented in section 3.4. For each pair of words in our for now unlinked graph, we check if they are lexically related.

To check for repetition, we simply compare the words as they occurred in the text. To find out if they are different forms of the same word, we compare roots. In both these cases, there is no need to consult the thesaurus.

If both previous tests fail, we look up the words in the thesaurus. All we need to do is call the appropriate routine with both words as arguments, and the thesaurus module will take care of the actual lookup. In this way, we can switch thesauri according to the domain.

If the words turn out to be related, we store this information in the lexical graph as follows. Since each word has a list of lexical links it is a part of, we add the link

to the lists of both related words. The link structure contains only the link type, and the pointers to both words.

Also at this time, we create a .gr file that will be the input of the graph drawing package. We do it right now rather than later because there is a line in the .gr file for each lexical link, so we output those lines as we compute each link.

A sample file is shown in Figure 4.2. Each paragraph is treated as a separate cluster, so that it will have a box drawn around it. The weights control the layout of the nodes and links. The lexical items that are not connected to other items also need to be specified so that they are drawn in the final picture as singleton components. The .gr file prepared in this way is later used by the *dot* package to create the graph in the PostScript format. This graph will be then displayed to the user after the analysis has been completed.

Once all the lexical items have been linked within the lexical graph, we are ready to search for the main component. This is done by a simple depth-first search, while also remembering all the visited paragraphs. We keep track of the largest component even if it is not the main one, so that if the text fails the analysis, we can identify the disconnected parts.

The next step is to collapse the graph. This is done by first creating a disconnected collapsed graph, one that has all the nodes and no links. We chose to implement the collapsed graph the same way as we did the full lexical graph, with nodes in a linked list. Each node contains the paragraph number, and the pointer to the list of lexical bonds. Each record for the bond, in turn, contains an integer to represent the strength of the bond, a pointer to each paragraph that takes part in the bond, and a pointer to the next bond for that paragraph.

After the graph is collapsed, we traverse it again, looking for the central paragraph.

```
graph g {
node [shape=ellipse];

subgraph cluster0
"1 1 income";
label = "para 1";

subgraph cluster1
"2 2 investments";
"2 2 selling";
"2 2 buying";
"2 3 investment";
"2 3 stocks";
"2 3 bonds";
label = "para 2";

subgraph cluster2
"3 4 stock";
"3 4 bonds";
label = "para 3";

"2 2 investments" -- "2 3 investment" [weight = 7] ;
"2 2 investments" -- "2 3 stocks" [weight = 4] ;
"2 2 investments" -- "2 3 bonds" [weight = 4] ;
"2 2 investments" -- "3 4 stock" [weight = 4] ;
"2 2 investments" -- "3 4 bonds" [weight = 4] ;
"2 2 selling" -- "2 2 buying" [weight = 4] ;
"2 3 investment" -- "2 3 stocks" [weight = 4] ;
"2 3 investment" -- "2 3 bonds" [weight = 4] ;
"2 3 investment" -- "3 4 stock" [weight = 4] ;
"2 3 investment" -- "3 4 bonds" [weight = 4] ;
"2 3 stocks" -- "2 3 bonds" [weight = 4] ;
"2 3 stocks" -- "3 4 stock" [weight = 7] ;
"2 3 stocks" -- "3 4 bonds" [weight = 4] ;
"2 3 bonds" -- "3 4 bonds" [weight = 8] ;
"2 3 bonds" -- "3 4 stock" [weight = 4] ;
"3 4 stock" -- "3 4 bonds" [weight = 4] ; }
```

Figure 4.2: The sample input file to the graph drawing utility *dot*. This output was created by analyzing the text shown in Figure 3.3.

We do it by first identifying the paragraph with the largest number of lexical bonds. Obviously, this is our candidate for the central paragraph. If there are two or more paragraphs with the same number of lexical bonds, we choose the first one as our candidate.

Now all we need to do is find out if there is a path from our candidate paragraph to all the other paragraphs. Again, we do it by depth-first search.

All the figures in this thesis were produced using our implementation as a part of our demonstration of the usefulness of lexical analysis.

Chapter 5

Support for the work

5.1 The experiment

The analyses described in chapter 3 are based on specific hypotheses about text connectedness. To compare how these hypotheses about text connectedness relate to how people perceive texts, we designed an experiment.

This chapter consists of two parts. First, we will describe the experiment in detail. In the second part, we will discuss the psycholinguistic foundation of why the experiment worked, by comparing it with a psycholinguistic theory of text connectedness described by Hoey [1991], which corroborates our work [1].

5.1.1 The goals of the experiment

By now, we have seen the lexical analyses performed using our model. Intuitively, the results make sense, since they are based on both the topic continuity and the

[1]Our work was developed independently of Hoey's, but we find his work very encouraging.

discourse structure approach to text coherence.

We had two goals in the experiment. First, we wanted to compare the results of human and computer judges and to see if they can arrive at similar results. If we can show that the results are comparable, we can then show that the results are useful for text processing, where humans would be provided with feedback about sites of potential incoherence.

Second, we had a more ambitious goal: we wanted to show that cohesion, although separate from coherence, nevertheless serves a useful supporting role in text coherence, and can be used as an aid in automatic text processing. In addition, if we can find evidence that people use cohesion as an indicator of coherence, then this provides further justification for the value of our model.

5.1.2 The design of the experiment

We began by choosing the subjects of our experiment. Our pool consisted of thirty subjects, all of them native speakers of English. To ensure a high level of literacy, our subjects were either university graduates or undergrad computer science students. None of them were expert in our domain.

Next, we collected a corpus of six texts, all from the same genre of financial advice. All texts were found by searching various web pages, and all were fresh texts, i.e., they did not participate in the thesaurus construction. We collected texts of various lengths, ranging from three to twenty paragraphs.

For each text we created a complementary text in the following way. If the original text had what our program would judge to be a coherence problem, we fixed that problem so that the modified text was judged coherent according to our model. On the other hand, if the text was judged coherent, we introduced some type of incoherence

into the complementary text. The introduction of incoherence involved either deleting a chunk of the original text, or adding some snippet of another text in the same domain.

In this way, we ended up with six pairs of text. One element of each pair was judged as perfectly coherent by the standards of our model. We used these texts for control. The other element of each pair had some coherence problems according to our model, and we hoped that our subjects would find the site of the problem. For the complete texts used in the experiment, see Appendix A.

Each subject was presented with a set of six texts, of which two were coherent and four had problems. Each coherent text was therefore viewed by 10 subjects and each one with coherence problems was viewed by 20 subjects. However, no subject saw both texts from any text pair.

The subjects were asked to read the texts in the same order they appear in appendix A, and to evaluate each paragraph of each text by assigning it a score between 1 and 5 according to how well they perceived the paragraph to fit with the rest of the text. The score of 1 means that the paragraph is seriously out of place, while five is the perfect score indicating no coherence problems.

We have decided to use the scale rather than ask for a binary judgement because we felt that it was important not to limit our subjects too much. In other words, we recognize that coherence is not a binary property of text. If we asked for a binary judgement, then paragraphs that were not perfectly coherent but also not badly incoherent could have been misjudged.

5.1.3 The results

Overview

After distributing the texts and collecting all the data from our subjects, the scores were tabulated and analyzed for each text. We had to show two things: first, that the human subjects did indeed perceive the incoherence problems with the texts we found incoherent, and second, that the problems were located at the same sites that our model identified.

Towards this end, we have used the one way ANOVA method of statistical evaluation [Shavelson, 1988]. This method allows us to compare two populations for statistically significant differences. We had our two groups of subjects as our populations — i.e. one group of ten people who saw the coherent version of a text, and one group of twenty people who saw the incoherent version. We wanted to learn whether judgements about coherence of presented paragraphs within the tested texts would differ significantly between the populations. We used the average scores of the subjects in each population to compare in the ANOVA test.

ANOVA works as follows. First, one forms the null hypothesis, i.e. that there is no statistically significant difference between the populations. By performing statistical analysis, we can disprove the null hypothesis. ANOVA compares the standard F distribution function with the results of the experiment. For each population, we need to compute the sum of squares of the results and the degree of freedom, both within each group and between groups. Mean squares are computed by dividing the sums of squares by the degrees of freedom. From these, we compute the value of the observed F distribution. The observed F-value is compared with a critical F-value in order to determine whether there is a statistically significant difference between the groups.

In our experiment, for each pair of paragraphs, one in the incoherent text and the corresponding one in the coherent version of the same text, we have calculated the $F_{observed}$ value. We chose to use the standard value of .05 that the null hypothesis holds (i.e. that there is a statistically significant difference between the groups with the probability of .95). In our experiments, there is one degree of freedom between groups and 28 degrees of freedom within groups (i.e. 20 - 1 + 10 - 1 = 28). Looking up the value of $F_{critical(.05,1,28)}$ we found it to be 4.20.

Therefore, we were able to compare the scores for each paragraph in the coherent and incoherent version. If the scores were significantly different, as indicated by the $F_{observed}$ value being larger than the $F_{critical}$ value, the subjects differ in their coherence judgements for this paragraph. Moreover, a comparison with the scores obtained from lexical analysis shows that the subjects are correctly identifying the paragraphs which are sites of incoherence. In fact, for all the texts that our method labeled as possibly incoherent, the human subjects, on average, also found coherence problems. In other words, whenever our method labeled a paragraph as incoherent, so did the human subjects, as we expected.

What we did not expect is that in human judgement the incoherence is not pinpointed with great precision. In other words, if a paragraph was not lexically connected, the paragraph immediately before it and the paragraph immediately after it also received lower scores. The statistical significance of this kind of scoring was confirmed by the ANOVA analysis of the results for these paragraphs. This indicates to us that perhaps untrained humans, such as our subjects, do perceive coherence problems but have trouble telling exactly what is wrong. This may indicate the usefulness of our approach as an assistance for human judgement.

For the paragraphs that were further away from the incoherence site, the results were not significantly different. This further indicates that we have found the correct

location of the problem.

In just one case, text 5, the method was unable to find the problem even though it clearly was perceived by our subjects. This illustrates the problem with detecting incoherence without performing full semantic analysis. Hence, our method does not detect coherence problems in all possible texts. Still, we were able to find all the other coherence problems in the texts we have tested, a fact we find encouraging.

Detailed description of results

In this section, we present detailed results of our experiment. For each version of each text, we first present the table containing the average score amongst subjects for each paragraph, as obtained from the experiment. Since the scores ranged from 1 to 5, the average falls between these numbers.

We also include the result of our lexical analysis for each paragraph. This is a score of either 1 (not connected and therefore potentially incoherent) or 5 (connected, therefore no problem detected).

We compared the scores for each paragraph using ANOVA. The results of these comparisons are included only if we found them statistically significant.

For the actual texts used in the experiment, please consult appendix A.

Text 1

This is a medium length text from a web page advertizing managed futures of a particular company. The coherence problem of this text occurs in paragraph 4, where the reader does not know how managed futures fit into the reviewing process.

paragraph number	experimental average score	computed score
1	4.8	5
2	4.2	5
3	3.35	5
4	2.5	1
5	3.8	5
6	4.4	5
7	4.6	5
8	4.7	5
9	4.6	5

Table 5.1: Results for the incoherent version of text 1.

In the corrected version, we explicitly show how in the review process the reader might add managed futures to the portfolio.

The average scores for the incoherent and the coherent versions of this text, as well as the results computed using our method, are shown in Figure 5.1 and 5.2 respectively.

Table 5.3 shows the ANOVA results for paragraph 4 (where df is an abbreviation for *degrees of freedom*). Clearly, the resulting differences are statistically significant, as expected. In fact, we can be very confident that the subjects did perceive incoherence at paragraph 4.

The results for the surrounding paragraphs are equally interesting. In the coherent version, paragraphs 3 and 5 are perceived as more coherent than the same paragraphs

paragraph number	experimental average score	computed score
1	4.8	5
2	4.6	5
3	4.3	5
4	4.3	5
5	4.4	5
6	4.7	5
7	4.5	5
8	4.9	5
9	4.8	5

Table 5.2: Results for the corrected version of text 1.

source of variation	sum of squares	df	mean square	F
between	33.8	2 - 1 = 1	33.8	49.54974
within	19.1	30 -2 = 28	0.682142857	
total	52.8	29		

Table 5.3: ANOVA results for paragraph 4 of text 1.

source of variation	sum of squares	df	mean square	F
between	6.02	2 - 1 = 1	6.02	6.321451
within	26.7	30 -2 = 28	0.951786	
total	32.7	29		

Table 5.4: ANOVA results for paragraph 3 of text 1.

source of variation	sum of squares	df	mean square	F
between	2.82	2 - 1 = 1	2.82	4.345271
within	18.2	30 -2 = 28	0.648214	
total	21	29		

Table 5.5: ANOVA results for paragraph 5 of text 1.

in the incoherent version of text 1. The differences are statistically significant, as shown in Tables 5.4 and 5.5.

For the remaining paragraphs, the ones that are further away from the incoherence site, the differences are not statistically significant.

Text 2

This short text contains last paragraph that is not related to the rest of the text. In the corrected version, we have deleted this paragraph. Another possibility for correcting this text was to explain how bonds with their lower appreciation do not show compounding as dramatically as stocks do.

Since we have deleted the paragraph that was the site of the coherence problems,

paragraph number	experimental average score	computed score
1	4.6	5
2	3.85	5
3	1.92	1

Table 5.6: Results for the incoherent version of text 2.

paragraph number	experimental average score	computed score
1	5	5
2	4.5	5

Table 5.7: Results for the corrected version of text 2.

source of variation	sum of squares	df	mean square	F
between	2.82	2 - 1 = 1	2.82	4.345271
within	18.2	30 -2 = 28	0.648214	
total	21	29		

Table 5.8: ANOVA results for paragraph 2 of text 2.

we did not have the coherence scores for this paragraph in the corrected version. For this reason, it was not possible to do the ANOVA test on this paragraph. Moreover, because the deleted paragraph was the last one in the text, we could only analyze the results for the paragraph immediately preceding the incoherent one. The results are shown in Table 5.8. Again, the subjects rated the paragraph preceding the offending one as less coherent than the same paragraph in the corrected version.

Again, there was no statistical difference for paragraph 1. The scores for both versions of the text are displayed in Tables 5.6 and 5.7.

Text 3

This is another short text in which the writer tried to say more than the space allowed. The resulting text consisted of two unrelated parts: paragraph 1 describing the Canadian political climate, and the rest of the text, addressing the need to keep the investment portfolio in a good shape.

The correction consisted of deleting the first paragraph — its relation to the remaining text was unclear.

If the writer had more space, an alternative suggestion would have been to show the relationship between the parts in an explicit way.

paragraph number	experimental average score	computed score
1	2.95	1
2	2.9	5
3	3.6	5

Table 5.9: Results for the version of text 3 with coherence problems.

paragraph number	experimental average score	computed score
1	4.4	5
2	4.7	5

Table 5.10: Results for the corrected version of text 3.

source of variation	sum of squares	df	mean square	F
between	14	2 - 1 = 1	14	11.76812
within	33.4	30 -2 = 28	1.191071	
total	47.4	29		

Table 5.11: ANOVA results for paragraph 1 of text 3.

source of variation	sum of squares	df	mean square	F
between	21.6	2 - 1 = 1	21.6	33.78771
within	17.9	30 -2 = 28	0.639286	
total	39.5	29		

Table 5.12: ANOVA results for paragraph 2 of text 3.

Because of this split, and because the text was so short, the results for this text has fallen into two categories. Some subjects judged all the paragraphs incoherent, others decided that paragraph 1 was fine and the remaining paragraphs were the problem, and the final group saw what we saw, i.e. they viewed paragraph 1 as the problem. In any case, the incoherent version was certainly judged as such (cf. Table 5.9), although human judges were not certain what exactly the problem was.

All the problems disappeared for the corrected version, as demonstrated in Table 5.10. Here, the text was judged as coherent, just as we expected.

In tables 5.11 and 5.12 we show the results of the ANOVA analysis for paragraphs 1 and 2 of this text respectively.

paragraph number	experimental average score	computed score
1	4.47	5
2	4.32	5
3	4	5
4	2.5	1
5	3.5	5

Table 5.13: Results for the incoherent version of text 4.

Text 4

In this text, paragraph 4 stands out as lexically unrelated to the rest of the text. It is also somewhat incoherent. In principle, it could be simply taken out and the result would be an improvement in the overall structure of the text.

On the other hand, the offending paragraph does contain some information which could be viewed as relevant given enough context. Our correction therefore includes the needed context. The resulting text reads more smoothly, and also is ranked higher on the coherence score.

Tables 5.13, 5.14, 5.15, 5.16, and 5.17, show the analysis for text 4.

Text 5

In this text, we have removed the second paragraph as it can sometimes happen by hasty copying. The resulting text is not quite coherent — the old paragraph 3 became paragraph 2, and the clue phrase *on the other hand* is now incoherent. This example illustrates the difficulty in evaluating text coherence without processing the text for

paragraph number	experimental average score	computed score
1	5	5
2	4.8	5
3	4.6	5
4	4.4	5
5	4.4	5

Table 5.14: Results for the corrected version of text 4.

source of variation	sum of squares	df	mean square	F
between	24.1	2 - 1 = 1	24.06667	26.53018
within	25.4	30 -2 = 28	0.907143	
total	49.5	29		

Table 5.15: ANOVA results for paragraph 4 of text 4.

source of variation	sum of squares	df	mean square	F
between	2.4	2 - 1 = 1	2.4	5.419355
within	12.4	30 -2 = 28	0.442857	
total	14.8	29		

Table 5.16: ANOVA results for paragraph 3 of text 4.

source of variation	sum of squares	df	mean square	F
between	5.4	2 - 1 = 1	5.4	8.689655
within	17.4	30 -2 = 28	0.621429	
total	22.8	29		

Table 5.17: ANOVA results for paragraph 5 of text 4.

meaning. The clue phrase, normally a coherence marker, is also out of place. However, neither the lexical analysis nor the clue phrase analysis can find the problem.

The coherent version is the original one, with the second paragraph intact.

Again, because we have deleted a portion of the text, it is impossible to apply ANOVA directly. Instead, we have compared paragraphs 1 and 3 of both versions. The results are shown in tables 5.20 and 5.21.

After removing paragraph 2, paragraph 3 is judged overwhelmingly as incoherent. However, because the links to other paragraphs still are intact, our method misses the problem.

Tables 5.18 and 5.19 show the scores for both versions of text 5.

Text 6

This text illustrates another problem that can occur when people edit texts by the cut and paste method. The last paragraph of the text is unrelated to the rest of the text. Our lexical analysis module correctly identifies the last paragraph as the site of coherence problems.[2]

[2]For our purposes, the comparison table does not count as a paragraph.

paragraph number	experimental average score	computed score
1	3.6	5
2	deleted	no score
3	1.95	5
4	4.8	5
5	3.5	5
6	4.9	5
7	4.9	5
8	4.9	5

Table 5.18: Results for the incoherent version of text 5.

paragraph number	experimental average score	computed score
1	4.4	5
2	4.5	5
3	4.7	5
4	3.5	5
5	4.9	5
6	4.9	5
7	4.9	5
8	4.9	5

Table 5.19: Results for the coherent version of text 5.

source of variation	sum of squares	df	mean square	F
between	4.27	2 - 1 = 1	4.26667	6.222
within	19.2	30 -2 = 28	0.685714	
total	23.5	29		

Table 5.20: ANOVA results for paragraph 1 of text 5.

source of variation	sum of squares	df	mean square	F
between	50.4	2 - 1 = 1	50.4	127.7526
within	11.1	30 -2 = 28	0.394643	
total	61.5	29		

Table 5.21: ANOVA results for paragraph 3 of text 5.

As previously, there are two ways to improve this text. The simpler one is to delete the offending paragraph and this is exactly what we have done. The alternative solution would be to show how this paragraph relates to the rest of the text. One way to do this is to discuss the fact that compounding works faster when the interest or growth rate is larger, and that there is a risk/reward ratio associated with each method of investing. We decided against including this elaboration, because this would make our text considerably longer, making the texts more difficult to compare.

Again, the human subjects have found paragraph 7 less coherent if it was followed by a weakly related paragraph 8 than if it was the last paragraph of the text. The scores for paragraph 8 were low, which is a further indication of coherence problems in that site.

Tables 5.22, 5.23, and 5.24 show the analysis for text 6.

paragraph number	experimental average score	computed score
1	5	5
2	4	5
3	4.8	5
4	3.5	5
5	4.9	5
6	4.9	5
7	3.35	5
8	1.1	1

Table 5.22: Results for the incoherent version of text 6.

paragraph number	experimental average score	computed score
1	5	5
2	4	5
3	4.8	5
4	3.5	5
5	4.9	5
6	4.9	5
7	4.3	5

Table 5.23: Results for the coherent version of text 6.

source of variation	sum of squares	df	mean square	F
between	6.02	2 - 1 = 1	6.02	8.158192
within	20.7	30 -2 = 28	0.7375	
total	26.7	29		

Table 5.24: ANOVA results for paragraph 7 of text 6.

5.2 Linguistic support (Hoey)

One linguistic theory that deals with text connectedness is presented by the linguist Michael Hoey [1991]. Although his theory as presented is not computational, and although it differs significantly from our model discussed here, it contains many aspects that support our view.

Hoey's aim is to show how "cohesive features combine to organize text." ([Hoey, 1991] p. 3) In other words, his claim is stronger than ours — he claims that cohesion is not just an indication of coherence, but an integral part of it.

Let us now examine Hoey's work as it relates to ours. Traditionally, one describes a text in terms of sentences. In other words, metaphorically speaking, a text is one huge sentence that is divided for the convenience of the reader. Hoey proposes a different metaphor, one more useful for his purpose — a text is a collection of interrelated packages of information. Each package contains one sentence. If we now find links that connect those packages, then we can determine the ones that are central to the text, i.e. they have the most links, and the ones that are peripheral to the text, i.e. they don't have that many links.[3]

[3]Peripheral to the text doesn't mean unimportant in general, it merely means not prominent in that text.

Hoey does not distinguish between lexical cohesive and other types of links. He states that repetition, in a very broad sense, is the device that shows the relatedness of sentences, but the repetition is treated very broadly and apparently arbitrarily. In some cases, it is ignored, and Hoey claims those cases are context-dependent. The individual links contribute to the general sense of connectedness, and so if an occasional link is lost due to an arbitrary decision then it does not greatly affect the task of determining the coherence structure of the text.

The following list describes the links used by Hoey:

- simple lexical repetition (or just simple repetition)—like us, Hoey does not worry about co-reference. Only open-set lexical items are considered, and closed set items, such as determiners and prepositions are excluded.

- complex lexical repetition—occurs when lexical items share the same morpheme but are not identical. For example, *drugs* and *drugging*. The distinction between the two becomes important if you consider collocation. Hoey doesn't, so it is unclear why he distinguishes these two repetition types.

  Hoey distinguishes between text-forming and chance repetition. The former is the essential property of the text, the latter happens "when the only common ground is the choice of the same lexical item" ([Hoey, 1991] p 56).

  In order to identify the text-forming repetitions Hoey checks the following. First, both items have to have a common referent. For example, the phrases *a three-wheeled pickup truck* and *a three-wheeler* that refer to the same physical truck have the referent, the physical truck, in common. Otherwise, perhaps they share context. If not, then maybe the contexts are parallel. Unfortunately, the question of context is a question of judgement.

Hoey himself admits there could be plenty of possible disagreements about it. So far, it seems there is no computational solution to this problem.

- simple paraphrase—this is essentially synonymy.

- complex paraphrase—this is a very broad category. It includes anything from antonymy to non-systematically classifiable relations (cf. 2.3). It also includes transitivity. For example, if we establish a relation between *writer* and *writing*, and independently between *writer* and *author*, then there is also a relation between *writing* and *author*. The extent to which Hoey intends to use transitivity in general as a useful tool for calculating his relations is unclear however.

- superordinate, hyponymic, and co-reference repetition—these are the same as Halliday and Hassan's definitions. The interesting twist is that the order of lexical items matters. If the more specific item is mentioned first, then the relation holds. If the order is reversed, then it doesn't, unless one can establish a common referent.

- other ways of repeating (not quite lexical)—include personal pronouns, demonstrative links and modifiers ellipses, and items such as *one* and *do*.

After deciding on the units and links for his analysis, Hoey then collected a set of sample texts and analyzed them. To represent his data, Hoey used a repetition matrix. It is a lower triangular matrix, where the rows and columns are indexed by sentence numbers. Each entry of the matrix contains all lexical links between items of the two sentences. This way, it is easy to determine the number and types of links for each sentence pair. It is also possible to compute lexical chains similar to ones used

by Morris and Hirst [1991] from this representation, which, having collected more information, in some sense supercedes their data structure.

The repetition matrix is then compressed to count all the links. In a compressed matrix, the entries are numbers of links between sentences.

The uncompressed and compressed lexical matrices give the appearance of computability. Indeed, the compressed matrix can be automatically derived from the full one. Unfortunately, Hoey gives no algorithm for computing the uncompressed matrix to start the whole process.

Hoey defines a bond as a relation between two sentences that share three or more lexical links. Three is an arbitrary cutoff point[4] that was established experimentally, and holds most of the time for the kind of texts Hoey examined. In general, the cutoff point depends on the length of the text, and on lexical density. Some texts in the legal genre have such a huge density that the cutoff had to be fixed at eleven or even twelve links.

Hoey suggests there is a correlation between genre and the level at which lexical bonds are formed, but he doesn't pursue the idea any further. We will revisit the idea of how genre influences the rhetorical structure of the text in section 7.1.6.

Having used the repetition matrix for calculating the bonds, Hoey next creates a *repetition net* where the nodes are sentences, and arcs are lexical bonds. With such a diagram, the text structure is easily determined. For example, the topic sentence is easy to find, since it is the one most connected with the rest of the sentences. Also, the subtopics are easy to identify, since they have many links to other sentences that are mostly interconnected but not well connected with the rest of the text.

[4]A cutoff point of x means that if there are fewer than x links between sentence A and sentence B then sentence A and sentence B are not related.

The topic nets have two important properties. First, they remove all contextual information, leaving only the coherence and cohesion information to examine. And second, using the nets one can identify which sentences are central and which are marginal.

The central sentences have more links to other sentences, while the marginal sentences have fewer links. This is somewhat similar to our central paragraph, but there are important differences. The idea behind the central sentences is to find those sentences in a text which best describe what the text is about. Often, short texts will have one such "topical" sentence. However, for long texts, often there is no one such single sentence. Rather, several sentences, grouped into a paragraph, describe the problem. Because unlike Hoey we process longer texts, the idea of the central paragraph is more appropriate.

The nets seem like a useful data structure. However, the longer the text, the more involved the net is. In fact, for texts longer than a few paragraphs, the net becomes too complex. This is intuitively expected, since long texts have several orders of magnitude more connections than short ones.

Hoey's theory offers some help with reading a text for various needs. In many situations, it is not necessary to understand, or even to read, the whole text. Often it is enough to read only mutually relevant sentences, as long as these sentences are bonded with each other.

5.2.1 Comparison of our approach and Hoey's

The aim of Hoey's work is similar to our own: to determine coherence structure from cohesion information. However, Hoey assumes that the text is coherent to begin with, and his task is to discover its structure. In contrast, we do not assume at the outset

that the text is structurally sound. Rather, we attempt to decide this very question.

In contrast to Morris and Hirst, and in similarity to our approach, Hoey attempts to gather all the lexical-cohesive information present in the text, rather than limiting the collection to the links between physically closest items. The information collecting phase occurs before any analysis takes place.

Both Hoey's approach and ours rely on the use of predefined types of links and both aim to preserve all the information. This means that transitivity is preserved. The links that Hoey chose to use are, unlike our links, not readily computable. In particular, complex paraphrase is not even defined rigorously enough to attempt an implementation.

Like us, Hoey has a cohesive unit. However, the size of the unit is different. For Hoey, the unit is a sentence; for us — a paragraph. From our perspective, a sentence is too small a unit. Particularly for longer texts, sometimes several pages long, using a sentence as a unit of cohesion is just not practical. This is not an issue for Hoey since his texts are never longer than a few short paragraphs.

Finally, the important difference is that Hoey uses all lexical items present in a text, and not only domain-related ones. Again, this has more to do with the fact that his texts are short, and so it is possible for him to use general vocabulary. But since his work is not computational, there is no need for any particular thesaurus. He simply uses his own vocabulary for deciding whether there is a link between two items.

In summary, Hoey's work is valuable and seems particularly well suited for short texts. In contrast, our approach is capable of handling texts of any length. One interesting direction for future work might be to combine these two approaches, having both a sentence-level and a paragraph-level analysis. We will revisit this idea in

section 7.1.

5.2.2 Hoey and the experiment

The work of Hoey, a linguistic theory with psycholinguistic roots, leads us to consider the possibility that people process text coherence by looking at the individual links between various components of a text and trying to arrive at some representation based on these links. Therefore, it makes sense to see the results of our experiment in which people indeed looked for such connections, and in their absence rated the poorly connected part lower on coherence than the other parts. In this sense, the results of our experiment were not in the least surprising.

In the same way, our own model is based on the very same connections between text constituents. Although we do not use cohesive information other than lexical, the strength of lexical cohesion is sufficient to determine text coherence in many cases.

However, analyzing one's own writing is different from analyzing texts written by a third party. Hoey offers some clues about this difference when he allows complex paraphrases as links. Processing such a link requires knowledge that sometimes goes beyond the information that is already in a text. We claim that such information may not be accessible to the reader, thus making a text difficult to process because of coherence problems.

This seemed to be the case with our experiment. In several cases, one could create an interpretation in which a text with some coherence problem would make perfect sense. The fact that our subjects still perceived such texts as not quite coherent supports our view.

For example, consider the text in Figure 3.3[5]. As we have seen in chapter 3, the

[5]This example is also a part of our study. See text 3 in appendix A.

first paragraph of this text is not lexically related to the rest of the text and for this reason we concluded that the text has a coherence problem at the beginning. However, one can offer an interpretation for which the text is acceptably coherent. For example, one can argue that investing is necessary in order to meet one's financial needs at retirement, and that rebalancing a portfolio is one money management technique suitable for retirement investing. Viewed in this context, the text is more acceptable because there is a logical link between the first paragraph and the rest of the text. Yet, our diagnosis of a coherence problem was confirmed by our experiment in which a majority of subjects gave a low score to the first paragraph.

This and other similar results of our experiment, and the work of Hoey, help to corroborate our model as a valid approach to detecting text coherence.

Chapter 6

Applications

In previous chapters, we have examined the lexical analysis process[1]. Now, we will consider its application to several areas of computational linguistics. These areas include text critiquing, second language learning, natural language generation, machine translation, and evaluating coherence of conversations, among others.

6.1 Text critiquing

There are many text critiquing systems available to users. Most of these are commercial systems designed for either business or college use (e.g. RightWriter [RightWriter, 1991], StyleWriter [StyleWriter, 1996], WinProof [WinProof, 1997]). In addition, many word processing packages offer some text critiquing capabilities.

In addition to the commercial developments, there has been some interest in the computational linguistics community, which resulted in some large-scale research projects such as Epistle [Heidorn *et al.*, 1982] its successor Critique [Richardson and

[1]By *lexical analysis* we mean the analysis of the lexical graphs.

Braden-Harder, 1988], and currently Gramcheck [GramCheck, 1997]. In fact, some commercial products use some ideas developed by these projects.

There are two approaches that system designers take. The easier method is to make the user write in a form that a system knows how to handle. One way to do this is to control the linguistic choices that the user makes, and to reject those choices that don't fit the system's expectations. In its pure form, such systems[2] make sure that the user follows the prescribed writing methods to the letter. In other words, such systems limit a user's writing choices. This is usually implemented by building a parser and a lexicon into the editor. As the user types, each sentence is processed to see if it conforms with the specifications. If it does, it is accepted. Otherwise, it is rejected. If the sentence parses, but the lexical choice is outside of the lexicon, sometimes the user has an option of adding the unknown word to the lexicon. Clearly, this approach to text processing severely restricts the user's freedom of choice.

A more flexible way of handling the text is to allow the user to write freely, and to later judge how closely the user's text matches the system's expectations. This is what most commercial systems do.

In spite of the limitations, many users find commercial packages helpful. Therefore, there are many style checker packages that users can buy. Most of these are called style and grammar checkers. They differ little in what they have to offer. All of them have a spelling checker and some sort of parser that covers English reasonably well. One of the oldest system, RightWriter [RightWriter, 1991], is a good example of such style checkers. Other interesting examples include StyleWriter [StyleWriter, 1996], FogFinder, PC-Proof and WinProof, GrammarPlus, WStyle, StrongWriter, and many others.

[2]One example is the IBM manuals intended as an input for machine translation (R. Boyd, personal communication).

StyleWriter, advertised as "Plain English at your fingertips!", is a typical example of a commercial style checker. The vendors claim that the system "acts as a professional editor for anyone in business or government who needs to write reports, proposals or press releases." They further claim to teach users better writing style, by concentrating on problem areas such as overuse of jargon or of the passive voice.

Naturally, some users might indeed have trouble with some of the problem areas. However, the approach taken by the checkers for addressing these problems is simplistic (e.g., flagging all occurrences). Some of these so called "errors" have their place in good writing, while many true errors are not caught.

6.1.1 What a typical style checker does

Style checkers operate on several levels: looking at words, phrases and sentences to check for spelling, grammatical and style errors. Many style checkers also analyze the text to compute a readability score. This can be done in several ways.

Number of words per sentence

This is the oldest and the simplest measure of sentence complexity. The assumption is that shorter sentences are easier to process. Clearly, this is a simplistic view that doesn't take into consideration the lexical choice and the sentence structure.

Flesch readability score

Just because a sentence is short doesn't necessarily mean it is easy to process. It is important to also consider the lexical items, since short sentences made up of short

words are easier to read than ones containing long words. There is also an assumption that long words are likely to be less familiar than the short ones.

One of the best known readability scores is Flesch's formula [Flesch, 1948]. The basic idea is to assign the highest scores to texts containing short sentences with few-syllable words. The formula is:

Reading ease score = 206.835 - (0.846 * SYLLS/100W) - (1.015 * WDS/SEN)

where SYLLS/100W = syllables per 100 words and
WDS/SEN = average number of words per sentence.

Other readability scores

The two methods described above are not very reliable. One of the main problems is that they don't consider sentence structure. While the length of a sentence tends to correlate with its complexity, it is not always so. Therefore, some readability scores also attempt to include sentence structure in the evaluation. However, most of the available systems don't fully include the parsing information. Instead, they count the percentage of sentences that can pose difficulties to the reader, such as passive sentences.

Problems with readability scores

The readability scores currently in use are not reliable for several reasons. First, they don't always consider the whole text, only a random sample of sentences, particularly for longer texts. It is therefore possible that they miss the problem areas.

Second, they don't consider the intended audience of the text. The level of education needed to understand a text is supposed to estimate the audience's level of

understanding, but that is not good enough. Consider a medical text, for example. Many people, even very well educated ones, but not medical professionals, might have trouble understanding such a text. Yet, a medical student or a nurse with less education but with some knowledge of the area might find the text easy to understand. Hence, what the user knows (the area of their education), and not just the level of their education, matters. Varying text structure to accommodate a user's background knowledge has been studied by various researchers (e.g. [Paris, 1988]).

Third, and perhaps most important, readability scores ignore the fact that the text structure plays a major role in text understanding. If the structure is problematic, then even the simplest sentences containing the basic vocabulary will be confusing for the reader. As far as we know, no method of evaluating readability takes the text structure under consideration.

Finally, the cohesion of the text aids understanding by making the text structure more obvious to the reader. Hence it stands to reason that cohesion should be incorporated into a readability score. Again, we know of no system that does so.

6.1.2 Incorporating lexical analysis into style checkers

All the systems available so far are only of limited use. They do offer some advantage to the writer, by pointing out some problem areas. And particularly for the second language learner, they do teach some good writing habits.

Unfortunately, they are only of limited use since they only analyze sentences and paragraphs in isolation. Yet, augmented by the lexical analysis module, a style checker would be able to perform a useful analysis of the whole text. Using the results of the analysis, the checker could then help the user identify potential coherence problems of the text, and offer practical advice about how to correct them.

For the financial domain, our software is already usable, but it is not integrated with any word processor. Fortunately, the combination of our system with a style checker would be relatively straightforward. The user would simply type in the text as he always does. Once he is finished typing, he can start the checker in the usual way. The first step would be to check the spelling and next the grammar, as it is now. As the last step, our lexical analysis system would build the lexical graph and analyze it, looking for indicators of incoherence. These would be displayed to the user in the same way as syntax mistakes are now — using a separate dialog box explaining where the mistake is and suggesting how to correct it.

We could apply one or both our hypotheses to texts in order to find those chunks of text that are not lexically related to the rest of the text. Choosing the appropriate hypothesis to apply depends on the length of the text (see section 3.4.1), and perhaps also on genre and the user's level of writing proficiency (this is further discussed in chapter 7.1.6). We can easily find the length of the text. However, the user would be responsible for setting the other parameters. The genre setting is a familiar one; many systems, such as Word 97, already allow the user to select one from an available list. The level of writing proficiency could be set up in a similar way.

If, using the appropriate hypothesis, the lexical module finds a chunk of text that is disconnected from the rest of the text, it can not only point out that chunk to the user, but also suggest that a bridging paragraph be added, or that the chunk be deleted. At the very least, the user will be made aware that there is a potential coherence problem. She can then choose to take action to correct the problem, and have the coherence of the corrected text evaluated again. This interactive process can continue until the user is satisfied with the evaluation.

To see how this approach would work in practice, we could apply it to one of the texts we have analyzed in chapter 3. Let us see what a typical style checker would

do to the text in Figure 3.3.

First, the spelling is checked and no problems are found. Second, the sentences are analyzed for possible problem areas. The first sentence is flagged as too long, with no specific suggestion. The phrase *hard look* and *more and more* might trigger the list of cliches, and be brought to the user's attention.

What happens next depends on whether the user specified the style. In formal writing, some style checkers, such as Word 97, do not allow the pronoun *you* as too informal. Hence, each occurrence will be flagged.

Some checkers won't be able to parse the last sentence. Thus, it may be flagged as wrong (it is in Word97), even though it is grammatically acceptable.

This is almost all the user finds from a typical style checker. There might be some additional information, such as a readability score, but actually many users don't know how to interpret this, so this information is often ignored.

The coherence problem is never found by the checker. It could easily be, though, if after the typical processing the system used the lexical analysis, as we did in chapter 3.

An interesting direction to explore in the context of text critiquing is to evaluate the shapes of collapsed lexical graphs. Many graphs that we have seen can be classified according to the number and placement of lexical bonds (somewhat similar to the approach proposed by Skorochod'ko [1972]).

Perhaps in some situations, such as working with beginning writers, it would be instructive for them to aim for texts that have particular shapes. For example, a narration tends to have a linear shape (i.e. constituents typically relate back to either the previous or the subsequent constituent) and so we can enforce that shape on the beginning writer. More advanced writers have more experience and hence should be

able to handle texts that have more unusual structure.

6.2 Critiquing texts of second language learners

Second language learners constitute a special group of users of style checkers. In some respects, they are simply writers, and so they often have the same problems as native speakers. Hence, everything we said in the previous section applies to the second language learners as well. But they do have some additional special considerations that a system designer should address to help the learners write better texts easier.

Second language learners often make errors that fall into one of two classes[3]. A transfer error occurs when the learner tries to apply the structures and constructions of his native language to the second language. An overgeneralization error occurs when the learner attempts to adapt the rule of the second language and applies it inappropriately. The former is more important from the text structure point of view, since different languages tend to require texts to be structured differently. If such a structure is then incorrectly applied in another language, it may result in an incoherence problem. In addition, by applying the text structures appropriate to the second language, the students learn to write texts that are more natural sounding and hence more appropriate.

One way to help the second language students to learn the acceptable structures is to teach them to adhere to the structure chosen in advance. For example, if the learner is to describe a room, one reasonable structure would be a spatial order, starting at the door and moving to the right. This ordering is the preferred one for English, but not necessarily for other languages. For example, a perfectly reasonable and

[3]I learned this in my English-as-a-second-language class.

more natural structure for Japanese would involve describing the most important items first. Hence, the ordering of the items would be from the most to the least significant, which for an English reader might give the impression of lack of order.[4]

Since the acceptable structures are known to the system, the user would have a warning any time the text deviates from this known, selected structure. Lexical cohesion analysis could be used to provide feedback to the students on whether the current shape of their text conforms to the required structure, as a kind of "structure drill". Furthermore, once all the errors are corrected, the student would know that the resulting text is structured appropriately.

Obviously, this rigid approach restricts the user's freedom of choice of text structure, but this is advantageous in case of second language learners who need stricter guidance than the native speakers.

6.3 Generation

Natural Language Generation (NLG) is the area of computational linguistics that is dedicated to computers producing their output in natural language. One of the difficulties in NLG is to figure out what to say and in what order. Active research attempted to answer this question by the use of data structures such as the focus stack [Grosz and Sidner, 1986] or the focus tree [McCoy and Cheng, 1988]. Another direction is to use a frame-based representation, and to traverse this representation, choosing the items to be talked about [Paris, 1988]. While these approaches help control the structure of the text, sometimes they fail to produce natural coherent structure because they do not link the text constituents the same way people do.

[4]For interesting examples of various kinds of transfer errors, see [Sperling, 1997].

Another challenge facing the generation systems is the need to adjust the text to the user, both in terms of contents and the form of the text. Some important work to address this challenge was done by Hovy [1988] and by Paris [1988] [1993] and others. This latter approach is frame-based, with frames that describe objects organized in a hierarchy. The hierarchy supports various relations, such as *instance-of* or *part-of*. In addition, other relations contain information about spacial or functional links between objects in the hierarchy.

Paris discovered that when deciding what to say next, different types of links should be followed, depending on the intended reader. For knowledgeable readers, traversing the hierarchy worked well. In contrast, for more naive users, following the cause-effect links worked better.

We believe that the lexical analyzer can be a useful tool that helps generate a more appropriate text. This can be done in several ways.

First, the lexical analyzer could be of use to an NLG system as an automated check of the quality of a generated text. Typically, human intervention is used to help produce coherent text, but we would apply the lexical analyzer before human intervention. Now, if a particular output fails the coherence check, then the text is likely to be misunderstood due to coherence problems. The NLG system can be asked to re-generate such a text, paying more attention to the particular reasons it failed.

For example, if the lexical analysis detects the absence of the main component of the lexical graph, the chunks of text not linked to the rest of the text are likely where a coherence problem occurs. Those chunks should then be re-generated, paying closer attention to the lexical links. Alternatively, a linking sentence or two could be added that makes the connection with the rest of the text explicit and hence brings the disconnected chunk back into the lexical graph. This approach could be especially

useful for very long texts. In this case, it would be more difficult for a human style checkers to perform well. Since the lexical analyzer described here is designed to operate on texts of any size, it would be appropriate to use.

6.3.1 Using the lexical analysis in a sample system

Let us now consider a sample system in more detail. Moore and Paris [1989] describe a text planner for advisory dialogues. They show that a generator needs to know not only the general purpose of the text, but also the intended effects of individual parts of text.

Therefore, their text planner pays attention to the intentional and the attentional structure (along the lines of [Grosz and Sidner, 1986]), and also the rhetorical structure (RST-style). In addition, although a full user model is not constructed, information about the user is tracked.

The text is planned top-down, and all the decisions affecting the intentional, attentional, and rhetorical structures are recorded.

Each plan operator therefore contains specific information geared towards these structures. And so, the operator has knowledge about its own effect (e.g. persuade, motivate). It also has a constraint list that contains conditions which must be true before the operator can be applied. A nucleus of an operator requires a speech act (e.g., inform, recommend, ask) that needs to be expanded, or a goal. Optional satellites are subgoals associated with the goal of the operator.

Moore and Paris distinguish between plans for achieving intentional and rhetorical goals. These goals are kept separate, because there isn't a one-to-one correspondence between them. This is so because there are many rhetorical strategies to achieve the

same intentional goal (e.g., one can describe the same concept in several different ways).

The text is planned top-down as follows. First, the goal is set. Next, all operators are examined to select those for which the effect field matches the discourse goal to be achieved. The selected operators are then examined further to find those with constraints that are satisfied. These operators with satisfied constraints become candidates for achieving the goal. Only one of the candidate operators is chosen for the generation. Its nucleus and satellites (if any) are then expanded in much the same way.

This leads us to the following considerations. What if, rather than choosing one operator, several operators were retained? This may not make sense for the very top operator, since coherent discourse often has one goal. But when expanding a nucleus or a satellite, it may happen that several operators could be combined to strengthen the effect. For the kinds of text Moore and Paris describe, this is not an issue because the texts are short. But for longer texts, this could be more common. When this happens, we need to order the chosen operators in a reasonable way, and make sure that the resulting text is coherent. In other words, we need to test the text plan to make sure we did not lose coherence.

Assume then, for simplicity, that we order the text so that the nucleus is presented first, followed by all its satellites. Then we could test if all satellites are lexically connected to their nuclei, and all nuclei connected to the goal. If there is some gap, we may specify this to the text planner with the suggestion to make the connection more explicit.

6.3.2 Lexical selection

Our approach could also be useful for guiding lexical choice during generation. When a text is generated, it makes sense to adjust the style of the generated text so that it reflects the way the users normally speak (for one example of a system that incorporates this idea, see [Bateman and Paris, 1989]). Naturally, this adjustment should include not only grammar, but also vocabulary. At the same time, we want the text to be cohesive; this may therefore direct the generator to use only a subset of a typical user's vocabulary.

One idea to promote cohesion of a generated text is to build the lexical graph of the text as it is being generated. In this way, we can have the ready analysis of the text so far. When the lexical item is to be chosen, we might try to fit the candidate items into the lexical graph, and choose the one with most lexical connections.

Other approaches to natural language generation have been proposed. For example, recent work by Marcu [1997] proposes a bottom-up construction of text plans, based on constraints of acceptable RST sequences. It is unclear how this work would extend to generating large-scale texts. In any case, the research does not yet address the question of "how to say" — the topic of lexical selection. The suggestions described above for employing our model towards the lexical selection process would be applicable in this context as well, and could therefore serve to complement the existing algorithm.

6.4 Machine translation

The discussion of text connectedness presented in this thesis pertains to English only. This does not mean that other languages do not have the same constructions to

express coherence and cohesion. They might have some or even all of the devices that English has, in addition to some that are absent in English. It does mean, however, that the same devices may be used in different contexts and with different cohesive effects.

However, few translation theories have anything direct to say about text connectedness. The reason for this omission is not lack of interest. Quite the opposite, human translators do recognize the value of connectedness in translation (see [Malone, 1988], [Stoddard, 1991], [Rickheit, 1991], and various others). Rather, the problem is that these theories address human translation. Humans are capable of intuition, hence the 'feeling' for language that humans learn includes textuality problems. Consequently, most translation textbooks discuss text connectedness in very general terms. A translation student reads such advice as "make the target text sound natural." There is no attempt at formalizing exactly what this naturalness is.

To compound the difficulty, most translation systems do not take the whole text under consideration. Rather, sentences are translated as they are encountered, with little context information having any bearing on the outcome. Hence, many heuristics are employed, but few usable theories exist. This is unfortunate, since clearly the context plays an important role in translation. Ideally, we would like a text-planning based MT system, but that is a matter for the future.

One of the few theories of translation that discusses connectedness in more specific terms is the theory of trajections [Malone, 1988]. The author formalizes a mechanism to describe translations as a set of transformations, called *trajections*. These trajections match a source-text component with corresponding target-text components.

Trajections occur simultaneously at many levels. For example, on the lexical level, one chooses a word that may influence the syntactic structure of a sentence. For

example, this happens when one chooses between divergent verbs, one that takes an object and one that does not. Hence, we sometimes must sacrifice a certain freedom of choice on one level to satisfy the choices we make on another. All these choices influence text connectedness in many different ways simultaneously.

Since the lexical graph of a translated text will differ from the graph of the source text, we can view the changes as trajections. In other words, we can classify the changes, and hopefully we can form theories that can predict them.

The predictions will be useful in two ways. One way is to use them to guide the translation process, choosing the lexical items that result in construction of the expected lexical graph. The other way is to use the graph comparison as an indication of translation quality. If the graphs are related by the acceptable trajections, this can be one indication of faithfulness of translation. If, on the other hand, the resulting graph is not the one we expected according the theory of trajections, then this may indicate some problems with the quality of translation.

An interesting observation has been made by Bencze [1985]. He claims that the author may include some extra information in the text that is meant to preempt any misunderstanding from occurring. This extra information, however, may be placed not at the place of potential misunderstanding, but later. This delayed placement creates translation problems for several reasons. First, its unusual placement reduces the cohesion of the original text. It is clearly sub-optimally placed, but the author must have considered it important enough to override this concern. The role of the translator is to decide if the peculiar placement should be preserved. Another source of problems is the fact that the extra information may in fact obscure rather than clarify. Consider the following example, translated from the Greek original:

(33) Then, taking hold of her hand, he said to her, "Talitha cum", which means

> "Get up, my child." Immediately the girl got up and walked about—[namely/for] she was twelve years old.[5]

Here, the additional information, the girl's age, is added because the original word, *talitha*, is ambiguous. Not only is it in Aramaic while the rest of the citation is in Greek, but it also has several meanings. It can mean either a newborn of either sex, or a little girl aged from birth to about fourteen, or, affectionately, any woman. Hence, the comment precisely describing the girl's age is intended to communicate the fact that there was nothing extraordinary in her walking—she was old enough for that. The author, obviously knowing both languages and well aware of lexical ambiguity created by the use of the word *talitha*, adds the extra information so as not to distort the truth. For the reason of faithfulness, the translator has included the extra information. Unfortunately, unless the reader knows Aramaic, the extra information does not make much sense—it is incoherent.

Situations similar to the one described above are quite common. The translator is therefore expected to decide whether to include the extra information, and if so, where to place it. Moreover, in many situations an explanation will involve cultural differences and hence be a paragraph long, or longer.

Consider the following example of a text in Polish[6] (Figure 6.1) and the first attempt at its English translation shown in Figure 6.2.

To a reader unfamiliar with Polish rural culture, this text sounds very strange. It clearly describes an unsuccessful courtship between two people, but the reason for the failure seems trivial at best. Although not stated in the text explicitly, the reader might attempt to deduce that the reason for Adam's breaking up with Janina

[5]Mark, 5: 40-42, after Bencze [1985].

[6]Text and both translations were constructed for the purpose of illustration.

Adam odwiedzał Janine co wieczor. Czasem przynosił drobne prezenty, jej ulubione czekoladki albo kwiaty. Az ktoregos dnia Janina poczestowała Adama czernina, wiec przestał przychodzic.

Figure 6.1: A sample Polish text.

Adam used to visit Janina every evening. Sometimes he would bring a small gift, a box of her favourite chocolates, or flowers. But one day Janina served Adam czernina soup, so his visits stopped.

Figure 6.2: The first attempt at translating the text in Figure 6.1.

is that he did not like czernina soup. But to Polish readers the meaning of this text is completely different. In order to preserve the original meaning then, let us add the information that the Polish readers know and that the readers of the translation need in order to interpret this text correctly. The corrected translation is shown in Figure 6.3.

Now it is clear that it was not Adam but Janina who broke the courtship. The meaning, very different now from the one inferable from the first attempt, is preserved

Adam used to visit Janina every evening. Sometimes he would bring a small gift: a box of her favourite chocolates or flowers. But one day Janina served Adam czernina soup.

This soup is considered a delicacy in Poland and is eaten quite often. By the tradition that was quite popular in the old times, and still observed in some parts of rural Poland, czernina soup has special meaning in the courtship ritual. If a woman serves it to a man, it means that she is not interested in him. She does not need to say this explicitly, thus sparing both parties an unpleasant conversation. Therefore, when Janina served the soup to Adam, his visits stopped.

Figure 6.3: The correct translation of the text in Figure 6.1.

by adding a paragraph explaining the role of the soup in Polish rural culture. The translator's decision to insert the whole paragraph rather than just one sentence is motivated by the fact that the explanation is quite long and elaborate.

In automated systems, there must be a formalism that will make such decisions. This is precisely where lexical analysis helps, since if the extra information is placed to preserve the connected property of the lexical graph, it has a greater chance of being perceived coherent. And in some cases, perhaps it is better to exclude the information that clarifies the meaning in the original text, but obscures the meaning in the translation. One way to determine if the extra information is coherent is to analyze the lexical graph of the translated text. If the addition is not lexically linked with the surrounding text constituents, it is likely better to omit it.

We have a clearer understanding of how cohesion behaves under translation than of how coherence does, particularly in the context of computational language processing. At the very least, we know that cohesion is language-dependent because there are different cohesive devices for different languages. For example, some types of clausal ellipsis common in English are impossible in Polish. Hence, the distinction between these two sentences

(34) You can borrow my llama if you want to.

and

(35) You can borrow my llama if you want.[7]

is impossible to preserve in translation into Polish. Rather, both sentences will be translated as:

[7]This example is from [Green, 1992].

(36) Mozesz pozyczyc moja lame jesli chcesz.

The fact that coherence is also language-dependent is less widely known. (Strictly speaking, coherence depends on culture, not just the language.) Moreover, while studies on cohesion are language-specific as many authors carefully point out (for example [Halliday and Hasan, 1976]), many studies of coherence do not address linguistic differences and it is not clear how the results transfer from one language to another. There are some indications that they do not translate well (cf. the discussion on psycholinguistic research on text connectedness and culture in chapter 2.1). In this case, a high-quality translation must change the structure of a text in such a way that the new structure is not only easier for target readers to understand, but is also more natural. This change will depend on both source and target languages, and will most probably be less extensive for pairs of languages within the (more or less) same culture than for unrelated languages.

For example, for Indo-European languages, restructuring of the whole text may not be necessary; perhaps minor adjustments are all that is required. In contrast, for cultures that differ significantly, a serious restructuring of the text may be the only possibility if we want the text to be understood at all. Alternatively, we might add some information, a guide for the reader, unaccustomed to the different way of thinking, through a maze of unfamiliar text construction. Inuit myths are sometimes presented in this way.

It may not be immediately obvious why accounting for coherence is necessary. But it becomes clear when one looks at bad translations, where there is no obvious connection between various parts of the text. Such texts are not really texts, since they lack unity. For this reason, they are extremely difficult to process. Clearly, evaluating coherence of such translated texts before they reach the readers will improve

the chance of the translators noticing and correcting at least some of these problems.

Because the one-to-one correspondence between linguistic constructions is rare (*cf.* Figures 6.1 and 6.3), the lexical graph of the translated text will be different from the graph of the source text. It is so not only because the words used in the target language may interact differently, but also because the target text will often be structured differently. Quite often, it will not even have the same number of paragraphs.

However, clearly a translation of a coherent text should be judged coherent. Hence, the use of the lexical analysis might prove useful. In other words, if the lexical analysis detects a possible coherence problem with the translation and not with the source, this is a reason for concern.

Another useful area for lexical analysis is to address the inclusion of the extra information, as we discussed earlier. In example 33, the lexical links generated by the word *talitha* will clearly be different in the original version and in the translation. Hence, a weak connection between chunks of text might indicate these areas where the extra information is not necessary and even not desirable.

The lexical analysis might prove useful to solve another problem common in machine translation, word sense disambiguation. When a lexical item is found to be ambiguous, there are several methods that can be applied to disambiguate it. One way is to simply ask a human. This is clearly undesirable, and in some cases impossible, if a bilingual human is not available. Another way is to attempt a disambiguation based on the information present in a text.

The lexical graph contains a lot of information that can help disambiguate words. If an ambiguous word is found, when it is added to the lexical graph, typically only one sense will fit. Most likely this sense will be the one intended. The reason for

this is the topic continuity — when a text talks about a particular topic, it tends to elaborate on that topic, rather than switch to some other, unrelated one (*cf.* our discussion of topic shifts in chapter 2). Hence, a lexically related word is more likely than the one that is unrelated.

When a domain-specific thesaurus is used, an interesting twist to the method arises: if a word doesn't fit into the graph, it may be the case that the intended meaning is not domain-specific. Morris and Hirst were the first to point out the value of lexical analysis for word disambiguation. However, by considering more lexical cohesive information, our approach gives more context in order to establish the meaning.

6.5 Information retrieval

Another area of application is information retrieval. With a large number of documents that need to be searched, the user needs as much support as possible to retrieve the required texts.

There are many methods that do a good job of retrieving the relevant texts. However, the number of texts they retrieve might still be too much for a user who must read each text to assess how useful it is. Some systems try to avoid this problem by presenting the user with the first paragraph of the text, with the understanding that the user can decide, based on this first paragraph, if he wants the whole text retrieved. This often works in some genres such as newspaper articles for example, because the first paragraph tends to contain the introduction and is thus related to all the remaining paragraphs.

In other words, the first paragraph of many newspaper articles is the central

paragraph. This is the reason why retrieving it is a good strategy. But in some texts, the central paragraph is not necessarily physically first. Therefore, for such texts, retrieving the first paragraph is not the best strategy.

We propose to retrieve the central paragraph instead. To speed up the retrieval process, we would pre-process all the texts so that we know the central paragraph of each text in advance and can fetch it and display on the user's screen without much real-time processing. Since the central paragraph is the one that is best connected with the rest of the text, it tends to contain lexical items related to all the ideas presented in the text. For this reason, it might be a good strategy to retrieve the central paragraph, and let the user decide based on it whether the whole text is worth reading. Clearly, this method would work better if the texts used more words from one domain.

6.6 Monitoring television programming

With more and more television stations offering more news each day, it is difficult to keep track of all the information. Users who suffer from information overload are turning to products that monitor and record television news programming for later viewing. One company that offers such a product is Televitesse[8].

The package monitors chosen television stations by keeping track not of the spoken words, but of close captioning, looking for designated keywords. A small section of the broadcast, typically 30 seconds, is kept in the buffer. When a designated keyword is found, the system records a portion of the broadcast, starting with the chunk that was stored in the buffer, and stopping after a fixed, user-specified amount of time.

[8] www.televitesse.com

Clearly, this product offers some advantage to its users. But there are some problems. First, it uses a specific list of keywords that trigger the recording. But if a synonym or some other related word occurs within the broadcast, the system will not save it even though it is very likely that the user would find it interesting. Thus, having not just a simple list of keywords but a thesaurus would help keep track of all the desirable broadcasts, not only the ones that happen to contain the keywords.

The other problem is that the recording time interval is fixed. Often, this may create situations where the recording will continue needlessly, or worse, stop abruptly at a wrong time.

We suggest an alternative that helps avoid this problem. We would monitor the broadcast as before, consulting the appropriate domain-specific thesaurus so that related words are found as they occur. If the match is found, we use a buffer to access the last 30 seconds of broadcasting. Out of this recording, we construct the lexical graph of the broadcast. In this way, we do not need to store a fixed amount of recording, but rather we can pinpoint with more precision where we should start the recording — we simply discard any unconnected parts of the lexical graph that occur at the beginning, since most likely these belong to the previous news clip.

Similarly, using the lexical graph we can determine when to stop recording, since the incoming vocabulary will no longer fit within the lexical graph we have constructed so far.

6.7 Partitioning web pages

Retrieving information from the world wide web can sometimes be a study in patience. One would expect that once the document is retrieved, the user's troubles are over.

Unfortunately, this is not always the case.

Long documents seem to be most difficult to handle. Since the time to load a document is proportional to the document's length, many web page designers choose to split one logically coherent and cohesive page into several smaller, more manageable chunks. From the technical perspective, this seems to solve the problem of a long loading time. However, the success of this method depends on how the original page was partitioned.

If the partitions are placed inappropriately, then rather than improving the situation, they made it worse. A user loads a page fragment, reads it, and moves to the next chunk. But if the information in both chunks is closely related, the user might need to flip back and forth between the chunks. The flipping is frustrating, time-consuming, and costly. It can also be preventable.

One way that perhaps could remedy this problem is to first build a lexical graph for the page to be partitioned. Now, the result of each partition can be seen clearly. We hypothesize that partitions placed so that they cut across the smallest number of lexical bonds will tend to create the least amount of flipping.

6.8 Evaluating coherence of a dialogue

The work on lexical cohesion as it pertains to coherence can be extended to evaluate coherence of conversations. This can be useful in situations in which an interlocutor utters a sentence that does not obviously fit into the current model of the conversation (for example, this occurs when a person suddenly remembers something important to the other participant, but not related to the current topic of the conversation). When the utterance does not quite fit the current conversation model, we have two

problems: one is that we might be required to commit considerable resources to process the utterance, and the other is the possibility of misunderstanding. If that is the case, then instead of exhaustive processing with doubtful outcome, it might be better to ask the interlocutor to rephrase the incoherent utterance. In other words, it makes more sense to reject those utterances that are not sufficiently coherent, even if sometimes we risk rejecting perfectly coherent utterances. Those rejections are a small price to pay for quicker, more reliable systems.

The incoherence detection mentioned above is purely local. We could also evaluate the coherence of the whole conversation, by determining where in the stream of utterances the current one fits. Of course we realize that there are important differences between spoken and written language, such as lexical density, that need to be accounted for [Donaldson *et al.*, 1996]. There are, as well, inherent differences between dialogue as a dynamic exchange between speaker and hearer, and texts, which are written for some audience. For future work, we can study how to modify our approach to be useful for determining incoherent utterances, as they occur, in conversation.

6.9 Summary of applications

As we have seen, lexical analysis offers some useful insights into several important areas of computational linguistics. Much work still needs to be done on text coherence, but in the meantime we have one useful indicator that is immediately applicable.

Chapter 7

Conclusions

7.1 Future work

In this thesis, we have made a step towards recognizing potential coherence problems in texts. Clearly, this is only the beginning and many problems still need to be investigated.

7.1.1 Extending the analysis of lexical cohesion

We have demonstrated how lexical cohesion analysis can find some types of coherence problems in text. In this section, we discuss several different ways the analysis can be extended.

Differentiating among lexical links

In this work, all lexical links were considered equally important. However, it is not necessarily so. Some links, such as exact repetition, are more cohesive and therefore

are potentially more indicative than other, less cohesive links. It is then reasonable to allow stronger links to carry more weight than other, less strongly cohesive links. For this reason, when we construct the lexical graph, we might take the type of link under consideration. Adding this extra information to our representation is a straightforward process.

This modification seems to offer some important advantages. In particular, for the main component hypothesis, it may turn out that if the main component touches a particular paragraph with one word only, and that word is connected by a sole weak lexical link, this might prove to be an insufficient connection for coherence purposes. For this reason, this refinement might make our method more precise.

The central paragraph hypothesis might benefit from this approach even more. Intuitively, a bond consisting of one weak link is less significant than one consisting of several strong ones. The next step then is to discover the threshold of significance.

We could also establish the strength of the lexical bond between paragraphs based not only on the number of lexical links that hold between individual words in those two paragraphs, but also on their types. This might prove useful in identifying and eliminating bonds that result from accidental links, such as the ones we have seen in Figure 3.21. Clearly, further study is needed to determine how useful these modifications can be.

Taking advantage of structure markers

Our method takes some advantage of the text structure on a very basic level. However, we make no attempt to use other information about the structure that the writer himself created for the text. Structural units such as subsections, sections, and chapters, exist in texts for reasons of coherence and cohesion, and hence might

supply important clues about incoherence. One approach perhaps worth trying is to collapse further the collapsed lexical graph, making these larger units the nodes of such a super-collapsed structure, and creating super-bonds, most likely using some sort of threshold mechanism. We hope this approach will yield further clues that will help detect more types of coherence problems in very long texts, such as full-length books.

7.1.2 Adding the syntax information

When a text is being created, the writer must make many choices before achieving the final, desired text. Some of these choices involve syntax. The writer often will give more prominence to some ideas than to others. This can and often is reflected in syntactic choices. In other words, the lexical items related to more prominent ideas will themselves be placed in more prominent positions in sentences. These patterns can be analyzed and might prove useful for our lexical analysis of text coherence.

For example, in a complex sentence, the main idea will tend to be expressed in the main clause, and the subordinate clauses will tend to contain elaborations on and clarifications to the idea expressed in the main clause.

If so, then the lexical links between items placed in the main clauses might prove more significant than the links between obscure items buried deeply within subordinate clauses.

7.1.3 Creating domain-specific thesauri

One source of difficulty for us was the lack of an appropriate thesaurus. The thesaurus we decided to build and use covers one domain only, which is enough to demonstrate

the usefulness of our approach, but not immediately applicable to a wide range of texts.

Note first that since we built our thesaurus by hand, we had to be extremely careful to specify all lexical links that apply. Quite obviously, this is an important issue. Fortunately, our method uses whole paragraphs as units, with many lexical items in them, and so it seems quite robust. In other words, even if we missed a link, this is not necessarily fatal. After all, typically there are many inter-paragraph lexical links, so even if we occasionally miss one our method still works. This does not mean that the thesaurus construction is not important — it is, and we checked the links very carefully.

Automatic thesaurus building is still in its infancy. However, as we mentioned, the idea seems promising. In fact, it might prove to be necessary to use some sort of automation process in constructing domain-specific thesauri to speed up the slow and tedious process and to avoid human mistakes.

The biggest advantage of this approach is that we will have no problem finding collocation links. In fact, this is one of the best methods to account for collocation.

Unfortunately, automatically built thesauri will not contain any information about link types, only about strengths (as discussed in section 3.2). For this reason, this extension is incompatible with some of the other extensions described above.

7.1.4 Creating user-specific thesauri

At present, we have one, generic thesaurus that is designed to serve the needs of an average user. However, this may not be sufficient in some situations. For example, cohesion and coherence are in the eye of the beholder, i.e., they depend on some aspects of the user knowledge, including the user vocabulary. Hence, the analysis

should be different for different users, reflecting their knowledge of the domain. In addition, the domain knowledge can change over time, and therefore it would be helpful to allow the thesaurus to adjust as the user learns new words.

Similarly, we may want to adjust the thesaurus to reflect the knowledge of the intended readers of the text. It may therefore be ideal to have some kind of "audience modeling", or a way to assess how the text will be perceived by the audience before actually presenting it to them.

For example, a user model which represents what a user knows may include what the user perceives to be systematically classifiable relations between concepts. If this is checked against the systematically classifiable relations in the existing thesaurus, it may point out parts of the text which the user will have trouble connecting.

There are also two possibilities for using user-specific thesauri in a kind of testing phase. One possibility is to run algorithms with a thesaurus which represents what the audience really knows (for instance, a user may not know that bonds earn interest, and therefore the PART-OF link between these nodes would not be present). Then, the writer could examine sites of possible incoherence to see where additional bridging and background for the user may be necessary.

Another possibility is that a user who constructs the thesaurus for the system may look at sites of possible incoherence from the analysis of sample texts to find words which should have been related in the thesaurus, perhaps resulting in revisions to that thesaurus.

7.1.5 Combining several existing thesauri

Another interesting idea is to use several different existing thesauri at once. This would make it possible to compensate for the shortcomings of any particular the-

saurus, while taking advantage of each of their strengths.

For example, while Wordnet doesn't have a nice way of computing relations between words of different categories, it can find antonyms. In contrast, Kipfer's thesaurus doesn't have a way of computing antonymy, but allows readily to compute relations between words in different categories. So, perhaps it would be useful to combine them in one package, using one as a primary thesaurus and the other when the first one can't find the relation.

Including general-purpose vocabulary

In addition to using the domain-specific thesaurus we might include some limited lexical relations from a general-purpose vocabulary. Of course care must be taken so that we do not reintroduce the problems we tried to avoid by using the domain-specific thesaurus in the first place, namely, explosion of the number of links in the graph. One perhaps productive approach is to use the repetition relation only on the lexical items not present in the thesaurus and not ubiquitous. Hearst has claimed some success with tracking of repetition alone for her segmentation experiments [1997].

7.1.6 Analyzing the collapsed graph

Analyzing the shape of the graph

As we have seen previously, the collapsed graph is a good indication of coherence problems in longer texts. However, we have only begun analyzing the information contained in it. It is perhaps possible to find more specific indicators by analyzing the shape of the collapsed graph.

Intuitively, we expect some shapes will be more typical than others, depending on the domain and genre. For example, a descriptive text will typically have the central paragraph at the beginning, and not somewhere in the middle, or worse, at the end. In contrast, a text that presents an argument may well have the central paragraph at the end. Hence, it might be possible to determine the structure and then present it to the user, indicating the type of text the structure matches. If the user is not satisfied, the system might help re-structure the text.

Clearly, much work needs to be done to determine what is the typical structure for a given genre. In addition, care must be taken, for a non-typical structure is not necessarily wrong. Still, for writers whose skills are not strong, such an extension might be helpful.

This type of analysis might also be useful for writing texts aimed at an intended audience whose level of knowledge about the topic is known. One might choose to structure a text differently for a knowledgeable audience than for a naive one.

The work of Skorochod'ko [1972] might be particularly suitable for using text shape information to guide natural language generation. He proposed to divide the texts according to how much overlap there is between sentences. These overlaps then would be represented graphically, with sentences as nodes and the overlap indicated by an arc. The resulting texts form patterns that can be classified and analyzed.

We found that working on the sentence level is not practical for longer texts. Still, the patterns of our collapsed graphs seem classifiable and we see this as useful. The reason is that the shape of a collapsed graph might offer important clues about the appropriateness of a particular text structure.

For example, we might want to decide in advance, before the generation begins, what shape the text will be. The choice will depend on the intended audience, and on

the contents of the text. And so, in the descriptive texts aimed at general audience we might choose to make the first paragraph of the text the central paragraph, while the remaining paragraphs might or might not be closely interconnected. We would also make sure that the last paragraph has many lexical bonds to other paragraphs, as is the case with summary paragraphs. For a generation system based on RST, this would roughly correspond to always placing the nucleus before the satellites, as we described in section 6.3.

In contrast, for a well-educated and sophisticated audience, and a text containing an argument, we might choose the last paragraph to be central. The somewhat unusual construction of a text will be appreciated by this kind of audience, and the argument can be laid out carefully without making conclusions before they are due. In practical terms, this effect can be achieved by placing satellites first, and the nucleus second. This construction is more unusual, and therefore it might be more difficult for the reader to process. For this reason, we will want the generated text to show all the relevant connections. One way to determine if the connections are there is to use the lexical graph.

Even keeping the genre constant can allow for interesting variations of text construction depending on the audience. Consider a narrative, for example. In its most basic form, a typical narrative has a rather linear structure, with obligatory bonds occurring between consecutive paragraphs. Other bonds are more likely to occur between paragraphs that are physically close than between distant ones. A text structured in this way is familiar, easy to understand, and therefore appropriate for audiences that are not familiar with the topic and not particularly well educated in general.

Now, consider a narrative aimed at more educated people. It can have a richer structure, with asides, and bonds between paragraphs that are not physically close.

Of course there should be some limit to the freedom of form, otherwise chaos would result. One way to control the chaos is to impose a minimum strength on the lexical bond between distant paragraphs.

Of course, much research will need to be done to determine which texts structures are appropriate for which topics and for which audiences. This is an interesting problem that perhaps will interest the user modeling community.

Combining several approaches

Following Halliday and Hasan, we have chosen a paragraph as a unit of cohesion. However, as we have seen in section 5.2, other arrangements can sometimes be advantageous. It might be valuable to combine several approaches. One interesting idea is to use different types of analysis at different levels of abstraction. For example, one could partition a text, using a method similar to the one described by Hearst [1994]. This would result in segments that would become units of cohesion for further analysis. Next, it will be necessary to determine if the individual units are coherent on the local level. A sentence by sentence analysis inspired by [Hoey, 1991] could determine this. Finally, a lexical graph of the segments as units and lexical bonds between them would establish the coherence of the whole text.

Improving suggestions for the users

Another area of research involves suggestions on how to improve texts. Right now, all we are able to do is tell the user where the potential problem is located, and offer very general suggestions. In order for the suggestions to be more specific, we need to identify those lexical items that need to be included to make the graph appropriately connected. Currently, one promising idea is to look at the lexical graph and analyze

the component that is not connected to the largest component. In it, we can find the item that has most lexical connections to other items within that isolated component. This item, called the *anchor node*, is the likely candidate for linking with some item in the largest component.

7.1.7 Incorporating syntactic cohesion

Lexical cohesion is only one type of cohesion. As we discussed in chapter 2, syntactic cohesion also has strong unifying effects and hence contributes to overall sense of unity in the text. For this reason, it can be potentially applied to our problem of detecting incoherence.

At this point it is not clear if syntactic cohesion analysis should be incorporated into the lexical analysis module, or if it should be treated separately. Clearly, if we were to include it in the same module, our hypotheses about text coherence will change. In particular, syntactic cohesion does not contribute to the lexical graph, which is confined to lexical relations alone.

The collapsed lexical graph, however, can be extended to accommodate syntactic cohesive links. Any such link would contribute to the strength of an inter-paragraph bond, thus making the central paragraph hypothesis even stronger.

It is worth noting that in order to include syntactic cohesion in our model, one needs to perform semantic analysis of a text. For example, anaphora resolution would require this in order to locate the proper referent. This is a difficult problem, and no satisfactory solution exists yet.

7.1.8 Broader view of evaluating text coherence

Lexical cohesion analysis presented in chapter 3 was described as one source of information about text coherence. We also mentioned that there could be other modules that contribute to the overall evaluation of text coherence. In order to gain understanding of what the analysis of lexical cohesion might contribute within a broader context of evaluating text coherence, we will sketch some of these other modules and discuss how they might work together with the lexical cohesion analysis module.

One interesting idea is to combine the modules into a *coherence filter* [Donaldson *et al.*, 1996]. The filter is an integrated framework which combines the results of coherence analyses from independent modules.

Each module analyzes a text and returns a numerical score indicating how coherent it believes the text to be. The scores are then combined to arrive at the overall coherence measure (where zero means the text is completely incoherent, and 1 — the text has no coherence problems).

Semantic relations analysis

Several researchers examining coherence have attempted to characterize the possible semantic relations between sentences (e.g. [Hobbs, 1976], [Mann and Thompson, 1983]). This research has focused primarily on representing the meaning of a text in terms of the underlying intersentential relations. We are interested here in determining conditions of possible incoherence.

This is a difficult problem because, as we have seen in chapter 2, uncovering semantic relations requires real world knowledge. Moreover, a deeper analysis of the underlying intentions of the speaker might also be required.

As an example of how semantic relations analysis could be used for evaluating coherence, consider the work of Marcu [1996] which proposes a method for text analysis using the Rhetorical Structure Theory (RST) ([Mann and Thompson, 1983]). The input to Marcu's algorithm is a set of sentences T, and the set R of all RST relations that hold between pairs of sentences in T. From this input, Marcu can generate an RST-tree that is the most coherent according to locality constraints, and Mann and Thompson's canonical nucleus/satellite ordering constraints.

Unfortunately, for the general coherence problem, we are given T but not R. If Marcu's algorithm were to be used to evaluate coherence, it would be necessary to automatically derive or approximate this set of relations. Marcu does not consider how R might be generated, and indeed, it appears to be a very hard problem in general. In particular this requires determining which relationship best fits particular text chunk. This is largely a matter of taste, since the RST relations are not formalized fully. Marcu's current solution relies on outside help, an oracle, to find the relationship that best fits the given pair of text constituents.

A simplified idea of how to evaluate coherence using an RST-based semantic relations analyzer, explored with examples in [Donaldson *et al.*, 1996], would be to decide that if a text has an RST analysis then it is coherent. In contrast, a text that has no RST analysis is considered incoherent.

Clue words

Clue words [Reichman, 1978, Cohen, 1987, Hirschberg and Litman, 1993] are words such as *now* and *well* that sometimes act as explicit markers of discourse structure.

Some clue word patterns guide the understanding of the text organization and hence make the text more coherent. For example, *first, second, third, ...* are parallel

structures [Cohen, 1987] that identify text constituents that together form a coherent unit and therefore should be in the correct order.

One idea from Donaldson [Donaldson *et al.*, 1996] is to use the clue phrases signature of a text, defined as an ordered list of clue phrases present in a text with a * to indicate the part of the text between clue phrases. These signatures are analyzed for patterns that suggest incoherence. For example:

- Some clue phrases seem unlikely to begin a text: clue phrases such as *otherwise*, *however*, *on the other hand*, *but*, *and*, *second/third/fourth*, *for example*, *finally*, ..., etc., all require previous text with which the forthcoming text will somehow be related.
- Similarly, some clue phrases seem unlikely to start a clause near the end of a text: *once upon a time*, *first*, *originally*, etc.;
- We would usually expect ordinal clue phrases, such as *first/second/third* ..., *initially/finally*, *one/two/three* etc., to appear in order in a paragraph, and without gaps.

Additional rules may be suggested by gathering statistics on the likelihood of various clue phrase patterns in a corpus of texts.

Note that this simple method of determining clue phrase signatures will sometimes be wrong, as clue phrases can have both sentential and discourse meanings [Hirschberg and Litman, 1993]. But, since this is only one of a number of methods to be combined in the coherence filter (described in a later section), complete accuracy for any one method is not essential.

A numeric coherence score based on clue phrases can be calculated as follows. The rules suggested above can be treated as preferences, and a penalty assigned every time

a preference is broken. Thus a coherence score between 0 and 1 for a text can be calculated. If the total possible penalty is known, then the score can be calculated using a formula such as

$$1 - \frac{\sum \text{accrued penalties}}{\text{max penalty sum}}.$$

An alternative, useful when the maximum penalty is unknown, would be a formula such as

$$\frac{1}{1+(\sum \text{accrued penalties})}.$$

Coherence filter score from the lexical analysis

For the lexical analysis module, we focus on determining text connectedness, using the paragraph as a unit of cohesion. Therefore, we now only consider paragraphs as nodes, without looking at the cohesion inside them. One proposal for deriving a coherence score is as follows. We traverse the lexical graph, and if there is a paragraph that cannot be reached from another node (representing another paragraph), then that unreachable paragraph is identified as a possible site of coherence problems. For the actual score, we find the largest chunk of connected paragraphs. Expressed as a percentage of the total number of paragraphs in the text, this is our coherence score.

Calculating the overall coherence score

Suppose we have a text T and n independent coherence judges $C_1 \ldots C_n$. The coherence evaluation by judge C_i is a real number $C_i(T)$ between 0 (total incoherence) and 1 (total coherence) that represents how coherent C_i finds T. For each C_i, we associate a weight w_i between 0 and 1 that represents how much we "trust" C_i's evaluation.

The coherence value of the whole system is just the weighted average of the individual coherence values: $C(T) = \sum_{i=1}^{n} w_i C_i(T)$, $\sum_{i=1}^{n} w_i = 1$. Each module independently makes a judgement as to the literal coherence of the text: the RST analyzer looks for a coherent RST tree, the clue phrase analyzer looks for reasonable clue words, and the lexical graph analyzer examines lexical graphs. We take $C(T)$ to be the overall coherence measure.

For this integrated approach to be successful, we need both good individual coherence modules, and a method for determining the individual trustworthiness weights w_i for each module. We expect that different w_i's may be appropriate for different styles of text, and so machine learning methods may be applicable here. For some examples of text analyses that show the interplay between the modules, see [Donaldson *et al.*, 1996].

We believe a lexical analysis module will have a prominent place within the coherence filter. The reason for it is that there are potentially many texts for which other modules will detect no errors while the lexical analysis module will be able to identify the coherence problem correctly.

Consider for example the text shown in Figure 7.1. While this text does have some coherence problems, it should probably not be classified as fully incoherent. In particular, the presence of clue words suggests the presence of some sort of relation between paragraphs two and three. In addition, this relation is probably co-ordinate rather than subordinate. However, it is unclear from this text what the relationship actually is. For this reason, the text should not be classified as perfectly coherent. Fortunately the lexical analysis module is able to find the small coherence problem in this text (which would not be detected by the clue words module).

Figure 7.2 shows the lexical graph of this text. The graph lacks the main compo-

> Are you thinking of investing your hard-earned money? There are many opportunities that can help you reach your financial goals. In order to choose the right investment strategy, you need to know about the products available to you.
>
> First, most people who watch the stock market think investing in it is something for the rich and the smart. Yet, you don't have to be an expert to benefit from the stock market. You can have someone else's expertise to work for you if you buy one of the well established mutual funds.
>
> Second, don't forget insurance. It is important for your peace of mind, and to protect the ones you love in an event of your accident.

Figure 7.1: The text with some coherence problems and clue words that are intended to overcome them.

nent. Moreover, the sole lexical item in paragraph 3 is not connected to any other lexical items in other paragraphs, which means the text lacks the central paragraph — a further indication of coherence problems.

Other issues

The idea of the coherence filter is intriguing and promising. Still, there are many details to research further. The most pressing issues include the method of combining the coherence scores from different modules, and the interplay of modules within a text.

We have sketched three modules that would form our coherence filter: the lexical analysis module, the semantic relations analysis module, and the clue words module. There are, however, other aspects of coherence that so far have not been represented in our discussion.

One such aspect is readability. Clearly, a text that is easy to read will be accessible

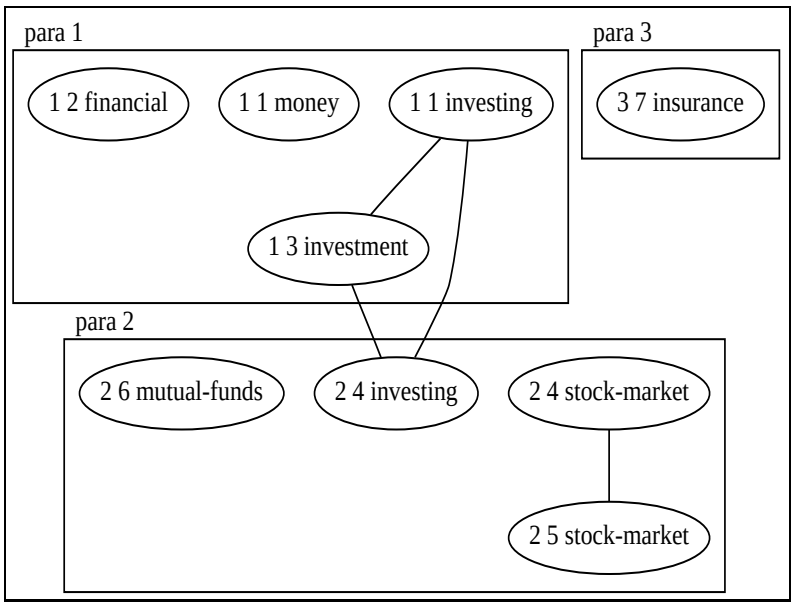


Figure 7.2: The lexical graph for the text in Figure 7.1.

to more readers. We have mentioned already that coherence depends on the reader in that a text on an unfamiliar subject is sometimes perceived incoherent simply because the reader lacks the knowledge to recognize connections between text constituents. One way to deal with this problem is to include a readability score as one module in our coherence filter. More precisely, a highly readable text would receive a higher coherence score than a more difficult one.

Another aspect of coherence is reflected in focus shifts. Clearly, inappropriate shifts of focus reflect poor text structure. For this reason, we could include a focus shifts analysis as another module in our filter.

There are other modules that we might choose to include in the filter. Since bad spelling and poor grammar make a text difficult to process, we might include a spell-checker and a syntax analyzer. And of course, if we choose to implement syntactic

cohesion as a separate module, that would also be a good candidate to include in our filter.

For future work, it might be worthwhile to see how both our method and the rest of the filter perform in the applications we have sketched out in this thesis.

7.2 Contributions

In this work, we had two goals. One was to investigate the problem of lexical cohesion as an indicator of text coherence. The other was to increase our knowledge about cohesion in general, which is an interesting research area in its own right.

7.2.1 Our starting point

We started with a specific point of view, considering cohesion and coherence as separate but related phenomena. Coherence is concerned with what makes sense in an utterance. Therefore, the semantics of discourse is the most important aspect of text from the coherence point of view.

Because processing texts for coherence involves fully understanding the contents of the text, it is a difficult problem. Therefore, in spite of the fact that coherence has been an object of intense scientific scrutiny for the last several years, we still don't understand it very well.

Cohesion, on the other hand, is concerned mainly with how various parts of the text fit together, independent of semantics or discourse. In other words, the contents are less important than the links between text constituents.

Although the last word on cohesion has not yet been said, many of the cohesive links are not only very well understood, but also possible to process computationally.

Cohesion plays an important supporting role for coherence. When two text constituents make sense together, that is, when they are coherent, they often have cohesive ties in common. One such a tie is lexical. Hence, lexical cohesion usually appears when coherence is present.

The converse also occurs — the text constituents that do not have any cohesive ties in common often are not semantically related.

One way to look at the relationship between coherence and cohesion is to just say that cohesion is a side effect of coherence. If this were the case, it could be dismissed as not important and uninteresting. But there is another point of view: if cohesion is a byproduct of coherence, then perhaps lack of cohesion is an indicator of incoherence. This observation was the starting point for my work.

7.2.2 Our particular approach

Although cohesion is not the only indicator of text coherence, it is a reasonably reliable indicator nonetheless. This work is a step towards automating that indicator, by approximating coherence of a text using one aspect of cohesion, the lexical aspect.

Towards this end, we have designed a new data structure, the lexical graph, which is suitable for analyzing the lexical cohesive structure of texts. The graph consists of lexical items, usually words, as nodes, and lexical relations as arcs. We use an online thesaurus to compute the links. The type of relations depends on the underlying thesaurus, but it always includes the most cohesive relations, such as repetition and synonymy.

Furthermore, for larger texts, we have designed another data structure, a collapsed lexical graph. It has paragraphs as nodes and lexical bonds as arcs. A bond between

two paragraphs is a set of lexical links that span lexical items in those two paragraphs. The strength of a bond depends on the number of links in it.

By examining texts we came up with two hypotheses about how lexical cohesion analysis can detect some coherence problems in texts. One hypothesis is that the lexical graph of a coherent text will have a component that spans all paragraphs of text. This component is called *the main component* of the text, and occurs in virtually all coherent short texts.

For longer texts, the main component test is too restrictive, i.e. some coherent long texts do not have the main component. The reason for this is that longer texts can have more elaborate structure, and there is enough space to present more material and to better deal with topic shifts.

For these longer texts then the collapsed graph is the preferred data structure. By analyzing the collapsed graphs of various texts, We have formulated our second hypothesis: the collapsed graph of a coherent text is usually connected. Such a graph has an interesting property: one can find a paragraph that not only has *paths* to all other paragraphs, but also has the highest number of *lexical bonds*. The presence of such a paragraph, called *the central paragraph* of the text, is a more reliable indicator of coherence for longer texts than the main component is. In other words, if a collapsed lexical graph is not connected, then the text likely has coherence problems that occur at or near the boundaries of the unconnected chunks of text.

The lexical analysis sketched out above can uncover some problems with the coherence of a text. Once we know what the problem is, we can now give some suggestions about how to improve such a less than perfectly coherent text. There are two possible ways to improve a text to make its lexical graph display the desired properties. One is to delete those chunks of text that are lexically unrelated to the rest of the text.

Although this sounds drastic, it nevertheless is appropriate in some cases.

The other way to improve the text is to add some transition sentences or paragraphs to bring the unconnected parts back into the lexical graph. The decision should rest outside of the coherence analyzer, with the human or a system that requested the coherence evaluation in the first place.

The analysis and the improvement suggestions described above can be supported by experimental evidence. Towards this end, we designed an experiment to find out if other people's intuition about where a coherence problem of a text is agrees with the results produced by our lexical cohesion analysis module.

The results of the experiment are encouraging. For most texts presented, both the subjects and our system often found the same locations of coherence problems. This helps to corroborate the analysis provided by the system. The experiment also shows the value of automating coherence evaluation, to provide a consistent, independent analysis which can perform well on both short and long texts.

The subjects were asked specific questions about whether certain paragraphs fit or did not fit, so the task required for the human evaluators was more restrictive than the analysis produced by the system. In addition, the subjects were only asked to evaluate reasonably small texts.

The experimental evidence shows that the method of our system is of value to provide an automated evaluation procedure, which handles both short and long texts.

In addition to the theoretical work, we have also outlined the possible practical applications. For example, in the area of text critiquing, most of the available systems offer only very basic advice. This advice is mostly limited to spelling and syntax, plus a rather unreliable readability index. The current systems don't attempt to analyze the whole text for coherence. With our approach, it is possible to build a lexical

graph of a text, and to find sites of possible coherence problems. The system can then recommend ways to improve the text structure either by deleting those text constituents that don't fit, or by adding transition sentences or paragraphs. The method can be also used in a similar way for second language instruction where a person writing in a second language can have her text evaluated and critiqued.

Similarly, the method can be used for improving the quality of text produced by the natural language generators. Before the output is presented to the user, it could be evaluated for coherence, and if it doesn't pass the criteria, it could be improved.

The method is also applicable to machine translation in two ways. First, we might evaluate the source text, and perhaps rewrite it if it has coherence problems, before the translation begins. Second, if the source text is problem-free, so should the translation be. If we detect coherence problems in the translation of a coherent text, it is an indication that the translation was not quite successful.

With some modifications, the method can be also applied to evaluation of coherence, both local and global, in conversation, and for various information retrieval applications (including web browsing and monitoring television broadcasting).

7.2.3 Summary

To sum up, we believe our thesis has made several important contributions. First, we have introduced two new data structures, the lexical graph and the collapsed lexical graph, to represent all the lexical cohesion information that occurs in texts.

Second, we use breaches in lexical cohesion as an indicator of possible coherence problems in texts. We show that some significant coherence problems can be detected in this way.

Finally, we identify several important areas of computational linguistics to which our model of incoherence detection can be useful. We briefly outline how to apply the model and show some of the advantages of using our approach.

Our work also contrasts with earlier research on the analysis of lexical cohesion by Morris and Hirst, St-Onge and Hirst, Hearst, and Hoey. Hoey's work is not computational, and so it does not result in algorithms which can be implemented. The other, computational, research seems to have a different goal. In our research, we don't just assume that a text is coherent and analyze it, we start out without preconceived notions about text coherence, and aim to evaluate a text from the perspective of coherence. Although there are a number of avenues for future research, we believe that we have indeed made some important contributions to the study of lexical cohesion and its application to evaluating text coherence.

Appendix A

Texts used in the experiment

This appendix contains the complete set of texts we have used for our experiment. We present both texts in each pair, the slightly incoherent version first, followed by the coherent one. For ease of reading, each paragraph is numbered. In the actual texts, we did not include the numbers.

Text 1

Text 1 — the incoherent version

1. If you invest in stocks, bonds, or mutual funds, this may be the most enlightening and useful Investors' Kit you have ever read. Get ready to be better informed and armed with tips for successful investing you probably never knew that could be most helpful.

2. Today there is approximately 3.2 trillion dollars invested in stock funds. A portion of this money comes from huge corporate and state pension plans

like Eastman Kodak, Xerox, IBM, AT&T, Aetna, Exxon, Westinghouse, Chase Manhattan Bank, and the Oregon, Michigan, and North Carolina state retirement systems.

3. These corporate and state pension plans have something else in common. They also have a proportion of their investment portfolios in Professionally Managed Futures. More and more, Professionally Managed Futures are becoming the investment of choice to try to maximize the returns, reduce the risk and help protect the underlying investments. So much so, that managed futures has become the fastest growing segment in the futures industry and one of the fastest growing investments of our time.

4. Any time your financial circumstances change, or when economic conditions shift, you should take a look at your portfolio and see if it needs to be updated. Even if these factors haven't changed, it's a good idea to review your portfolio at least once a year.

5. At this point, those who are unfamiliar with Professionally Managed Futures might be thinking, "How can futures increase performance, reduce risk, and help protect the investments when trading futures is a crap shoot and very risky?" Yes, trading futures can be a crap shoot and very risky. In fact, studies show most amateur investors who trade futures on their own do lose. But there is a logical reason for the amateur losses. From over 11 years of observation, we strongly believe most amateurs trade futures on rumors, tips from friends, gut feelings, part-time research, and for the fun of it. They are not professionals, have insufficient training and experience, and, in our opinion, have no business trading on their own. Most are doomed to fail before they begin, just like you would be if you played Michael Jordan

one-on-one basketball or performed a medical operation and you weren't a doctor.

6. While amateur futures traders usually lose, many Professional Commodity Trading Advisors (CTAs) have achieved highly attractive returns through prudent money management. In fact, Professional CTA's bring to futures trading many of the same benefits that stock fund managers bring to management of stock and bond funds.

7. With the highly attractive returns many CTAs have achieved through prudent money management, managed futures stands on its own merits. However, for most, the real value in Professionally Managed Futures lies in its ability to increase performance and reduce risk.

8. Professionally Managed Futures are a distinct asset class different than securities, uncorrelated, and do not move in lock step with stocks and bonds...which is why managed futures have been shown to increase performance while reducing risk.

9. This should give you a clearer understanding as to why Professionally Managed Futures is one of the fastest growing investments of our time.

Text 1 — the corrected version

1. If you invest in stocks, bonds, or mutual funds, this may be the most enlightening and useful Investors' Kit you have ever read. Get ready to be better informed and armed with tips for successful investing you probably never knew that could be most helpful.

2. Today there is approximately 3.2 trillion dollars invested in stock funds. A portion of this money comes from huge corporate and state pension plans like Eastman Kodak, Xerox, IBM, AT&T, Aetna, Exxon, Westinghouse, Chase Manhattan Bank, and the Oregon, Michigan, and North Carolina state retirement systems.

3. These corporate and state pension plans have something else in common. They also have a proportion of their investment portfolios in Professionally Managed Futures. More and more, Professionally Managed Futures are becoming the investment of choice to try to maximize the returns, reduce the risk and help protect the underlying investments. So much so, that managed futures has become the fastest growing segment in the futures industry and one of the fastest growing investments of our time.

4. As your financial circumstances change, or when economic conditions shift, your portfolio needs to reflect these changes. The next time you do the protfolio review, consider adding managed futures to increase performance and reduce risk.

5. At this point, those who are unfamiliar with Professionally Managed Futures might be thinking, "How can futures increase performance, reduce risk, and help protect the investments when trading futures is a crap shoot and very risky?" Yes, trading futures can be a crap shoot and very risky. In fact, studies show most amateur investors who trade futures on their own do lose. But there is a logical reason for the amateur losses. From over 11 years of observation, we strongly believe most amateurs trade futures on rumors, tips from friends, gut feelings, part-time research, and for the fun of it. They are not professionals, have insufficient training and experience,

and, in our opinion, have no business trading on their own. Most are doomed to fail before they begin, just like you would be if you played Michael Jordan one-on-one basketball or performed a medical operation and you weren't a doctor.

6. While amateur futures traders usually lose, many Professional Commodity Trading Advisors (CTAs) have achieved highly attractive returns through prudent money management. In fact, Professional CTA's bring to futures trading many of the same benefits that stock fund managers bring to management of stock and bond funds.

7. With the highly attractive returns many CTAs have achieved through prudent money management, managed futures stands on its own merits. However, for most, the real value in Professionally Managed Futures lies in its ability to increase performance and reduce risk.

8. Professionally Managed Futures are a distinct asset class different than securities, uncorrelated, and do not move in lock step with stocks and bonds...which is why managed futures have been shown to increase performance while reducing risk.

9. This should give you a clearer understanding as to why Professionally Managed Futures is one of the fastest growing investments of our time.

Text 2

Text 2 — the incoherent version

1. The effects of compounding have the potential to increase your return significantly over the long term. By continuously reinvesting your earnings back into your account, you can keep all of your money working to potentially earn still more.

2. The sooner you begin and the longer you remain invested, the greater the potential benefits of compounding. For example, if 15 years ago you began investing $300 monthly in the S&P 500, the value of your portfolio today would be $208,089.28. The S&P 500 is an unmanaged index, commonly used as a proxy for the U.S. stock markets.

3. On the other hand, if you think you will need the money soon, (for example, you are close to retirement or you are saving up for some large purchase) you might want to put more emphasis on bonds.

Text 2 — the coherent version

1. The effects of compounding have the potential to increase your return significantly over the long term. By continuously reinvesting your earnings back into your account, you can keep all of your money working to potentially earn still more.

2. The sooner you begin and the longer you remain invested, the greater the potential benefits of compounding. For example, if 15 years ago you began investing $300 monthly in the S&P 500, the value of your portfolio today

would be $208,089.28. The S&P 500 is an unmanaged index, commonly used as a proxy for the U.S. stock markets.

Text 3

Text 3 — the incoherent version

1. As governments take an increasingly hard look at universal social programs and growing deficits, Canadians are becoming more and more concerned about their ability to meet their individual income needs during retirement.

2. Rebalancing your portfolio is a powerful disciplinary tool that allows you to manage your investments for profit, by selling high and buying low. For example, you may have decided originally that an investment mix of 60% in stocks and 40% in bonds met your objectives.

3. However, due to a prosperous period in the economy, you find that the stock component of your portfolio grows to 75% while bonds now represent only 25% by value. In this case, we would advise you to take profits and rebalance back to your original mix so that when the economic cycle reverses– as it always does— you will be positioned to take advantage of the change.

Text 3 — the coherent version

1. Rebalancing your portfolio is a powerful disciplinary tool that allows you to manage your investments for profit, by selling high and buying low. For example, you may have decided originally that an investment mix of 60% in stocks and 40% in bonds met your objectives.

2. However, due to a prosperous period in the economy, you find that the stock component of your portfolio grows to 75% while bonds now represent only 25% by value. In this case, we would advise you to take profits and rebalance back to your original mix so that when the economic cycle reverses– as it always does— you will be positioned to take advantage of the change.

Text 4

Text 4 — the incoherent version

1. When developing your financial plan, you first need to consider whether you're an "investor" or a "saver."

2. Investors look to invest some money for the longer haul. They want capital growth over time, some income, and/or tax-free income. Investors have adequate reserves to meet their current needs and can, therefore, stay invested in the market. They are also willing to ride out any short-term market fluctuations and invest instead for long-term growth potential to outpace inflation over time.

3. Savers need to focus on current or short-term needs. Their primary concern is preservation of their capital. Savers also seek liquidity for ready access to cash when necessary.

4. Your financial decisions will be affected by your time frame, your objectives, and your tolerance for risk.

5. You and your financial adviser can determine whether you are a saver or an investor, by looking at your needs and goals. Once this determination is

made, your advisor can help you build an investment portfolio that's right for you.

Text 4 — the coherent version

1. When developing your financial plan, you first need to consider whether you're an "investor" or a "saver."

2. Investors look to invest some money for the longer haul. They want capital growth over time, some income, and/or tax-free income. Investors have adequate reserves to meet their current needs and can, therefore, stay invested in the market. They are also willing to ride out any short-term market fluctuations and invest instead for long-term growth potential to outpace inflation over time.

3. Savers need to focus on current or short-term needs. Their primary concern is preservation of their capital. Savers also seek liquidity for ready access to cash when necessary.

4. Whether you are an investor or a saver, your financial decisions will be affected by your style. Therefore, in addition to your time frame and your objectives, your financial plan should reflect your tolerance for risk.

5. You and your financial adviser can determine whether you are a saver or an investor, by looking at your needs and goals. Once this determination is made, your advisor can help you build an investment portfolio that's right for you.

Text 5

Text 5 — the incoherent version

1. Stocks are shares in a company. When you invest in a company's stock or buy its shares, you own part of a company. If the company makes money, your stock will increase in value. But, just as in short-term investment and bonds, there are pros and cons to stock investments.

2. ...

3. On the other hand, stock prices often go up and down. They are never guaranteed. A shareholder may lose part or all of his money.

4. However, in the long run, stocks have beaten alternative investments such as bank accounts, bonds, real estate, and commodities. A Chicago consulting firm, Ibbotson Associates, has compiled data to show that stocks are the way to go. As shown in the chart below, stocks, represented by the Standard & Poors 500, doubled the compound annual return of T-bonds issued in 1926.

5. If you buy a share or shares of stock in a public company, you become a part owner of that company. As a shareholder of one share of Microsoft, you enjoy the same basic privileges and rights as a Bill Gates who owns millions of shares.

6. As a shareholder, you have the privilege to receive quarterly reports and an annual report informing you of the financial health of the company. These reports are just like report cards you receive from school. The quarterly reports tell how much money the company has made or lost and business activities during the reporting period. The annual report is a combination of

all quarterly reports and is often printed with fancy charts and photographs. It gives detailed business and financial information about the company. As a shareholder, every year you'll be invited to attend the annual shareholders' meeting, where you can ask Mr. Gates questions about Microsoft.

7. In addition, you will have the right to vote for Microsoft's board of directors, the shareholders' representatives who keep track of the important issues of the company. They will, in turn, hire officers such as Chairman Gates to run the company.

8. Most companies use a one-vote-one-share system. Even though your one share of Microsoft does not count much against Mr. Gates's millions of votes, the company takes each vote seriously. If you cannot go to the annual shareholder's meeting, they will send you an absentee ballot.

Text 5 — the coherent version

1. Stocks are shares in a company. When you invest in a company's stock or buy its shares, you own part of a company. If the company makes money, your stock will increase in value. But, just as in short-term investment and bonds, there are pros and cons to stock investments.

2. Stocks have a long historical track record of outperforming other investments, such as bank deposits, money-market funds, CDs, bonds, real estate, and commodities. See the chart below for a comparison from 1945 to 1994. A stockholder or shareholder has voting rights that bondholders and bank depositors do not have.

3. On the other hand, stock prices often go up and down. They are never guaranteed. A shareholder may lose part or all of his money.

4. However, in the long run, stocks have beaten alternative investments such as bank accounts, bonds, real estate, and commodities. A Chicago consulting firm, Ibbotson Associates, has compiled data to show that stocks are the way to go. As shown in the chart below, stocks, represented by the Standard & Poors 500, doubled the compound annual return of T-bonds issued in 1926.

5. If you buy a share or shares of stock in a public company, you become a part owner of that company. As a shareholder of one share of Microsoft, you enjoy the same basic privileges and rights as a Bill Gates who owns millions of shares.

6. As a shareholder, you have the privilege to receive quarterly reports and an annual report informing you of the financial health of the company. These reports are just like report cards you receive from school. The quarterly reports tell how much money the company has made or lost and business activities during the reporting period. The annual report is a combination of all quarterly reports and is often printed with fancy charts and photographs. It gives detailed business and financial information about the company. As a shareholder, every year you'll be invited to attend the annual shareholders' meeting, where you can ask Mr. Gates questions about Microsoft.

7. In addition, you will have the right to vote for Microsoft's board of directors, the shareholders' representatives who keep track of the important issues of the company. They will, in turn, hire officers such as Chairman Gates to run the company.

8. Most companies use a one-vote-one-share system. Even though your one share of Microsoft does not count much against Mr. Gates's millions of votes, the company takes each vote seriously. If you cannot go to the annual shareholder's meeting, they will send you an absentee ballot.

Text 6

Text 6 — the incoherent version

1. If you put off saving until later in life, time becomes your greatest enemy; if you begin saving early, time becomes your greatest ally. Year after year, your assets may earn interest and dividends, and those earnings in turn generate additional earnings, and so on. This "magical" process is called compounding. The sooner you start saving for retirement, even if the amounts you set aside are modest, the greater the benefits you will receive from the power of compounding.

2. Suppose you want to build $100,000 in retirement assets by age 65: If you start at age 35, you will need to save $67 a month to reach that goal, assuming an 8% average annual return. If you wait until age 55, you will need to save a whopping $543 per month.

3. Let's look at another example of the power of compounding. In this hypothetical scenario, Pat and Chris begin contributing to their respective employer's qualified retirement plans.

4. Pat joins her employer's plan at age 30, contributes $1,000 per year, and earns an 8% annual rate of return. Pat continues this program for 10 years

and then stops making contributions. Pat allows her contributions to compound at an 8% annual rate of return until retirement at age 65.

5. Chris waits to join his employer's plan until age 40 (10 years later than Pat). Chris also contributes $1,000 per year and earns an 8% annual rate of return. He continues this program for 25 years until retiring at age 65.

6. Compare the total amount saved by each individual upon reaching age 65 (see table below). As you can see, both Pat and Chris benefited from the power of compounding. Pat, however, used time more effectively and was able to save nearly $30,000 more than Chris – despite the fact that Chris contributed $15,000 more than Pat over the years.

   Pat
   Total Contributed: $10,000
   Total Value At 65: $107,100

   Chris
   Total Contributed: $25,000
   Total Value At 65: $79,000

7. The message is simple: The sooner you start saving, the easier it will be for you to reach your retirement goals.

8. Whether you should choose stock or income funds depends on time frame, your investment objectives, and your tolerance for risk. Before making any decision, you should first define your goals by asking yourself some questions. What is the purpose of the investment? How long is the time

frame for keeping the money invested? What is your tolerance for risk? The bottom line is that you must have a clear idea of your objective.

Text 6 — the coherent version

1. If you put off saving until later in life, time becomes your greatest enemy; if you begin saving early, time becomes your greatest ally. Year after year, any assets that you invest may earn interest and dividends, and those earnings in turn generate additional earnings, and so on. This "magical" process is called compounding. The sooner you start saving for retirement, even if the amounts you set aside are modest, the greater the benefits you will receive from the power of compounding.

2. Suppose you want to build $100,000 in retirement assets by age 65: If you start at age 35, you will need to save $67 a month to reach that goal, assuming an 8% average annual return. If you wait until age 55, you will need to save a whopping $543 per month.

3. Let's look at another example of the power of compounding. In this hypothetical scenario, two investors, Pat and Chris, begin investing in their respective employer's qualified retirement plans.

4. Pat joins her employer's plan at age 30, invests $1,000 per year, and earns an 8% annual rate of return. Pat continues this program for 10 years and then stops making contributions. Pat allows her contributions to compound at an 8% annual rate of return until retirement at age 65.

5. Chris waits to join his employer's plan until age 40 (10 years later than Pat). Chris also invests $1,000 per year and earns an 8% annual rate of return.

He continues this program for 25 years until retiring at age 65.

6. Compare the total amount saved by each individual upon reaching age 65 (see table below). As you can see, both Pat and Chris benefited from the power of compounding. Pat, however, used time more effectively and was able to save nearly $30,000 more than Chris – despite the fact that Chris contributed $15,000 more than Pat over the years.

   Pat

   Total Contributed: $10,000

   Total Value At 65: $107,100

   Chris

   Total Contributed: $25,000

   Total Value At 65: $79,000

7. The message is simple: The sooner you start saving, the easier it will be for you to reach your retirement goals.

Appendix B

Lexical analysis for the experiment

B.1 Text 1

We will first analyze the incoherent version of text 1 presented in appendix A. The lexical graph of this text contains so many links that it is difficult to reproduce it on these pages. In addition, because at 9 paragraphs the text is rather long, we apply the central paragraph test and not the main component one when judging coherence.

The collapsed lexical graph for text 1 is shown in Figure B.1. In this figure, paragraph 4 is shown as lexically unrelated to the rest of the text. In other words, it contains no lexical items that are lexically related to any other lexical item in any paragraph of the text. For this reason, our method classified paragraph 4 as a site of a potential coherence problem.

This is in agreement with our experimental scores.

Let us now turn our attention to the coherent version of this text. There, we have changed paragraph 4 so that it now relates to both the preceding and the subsequent

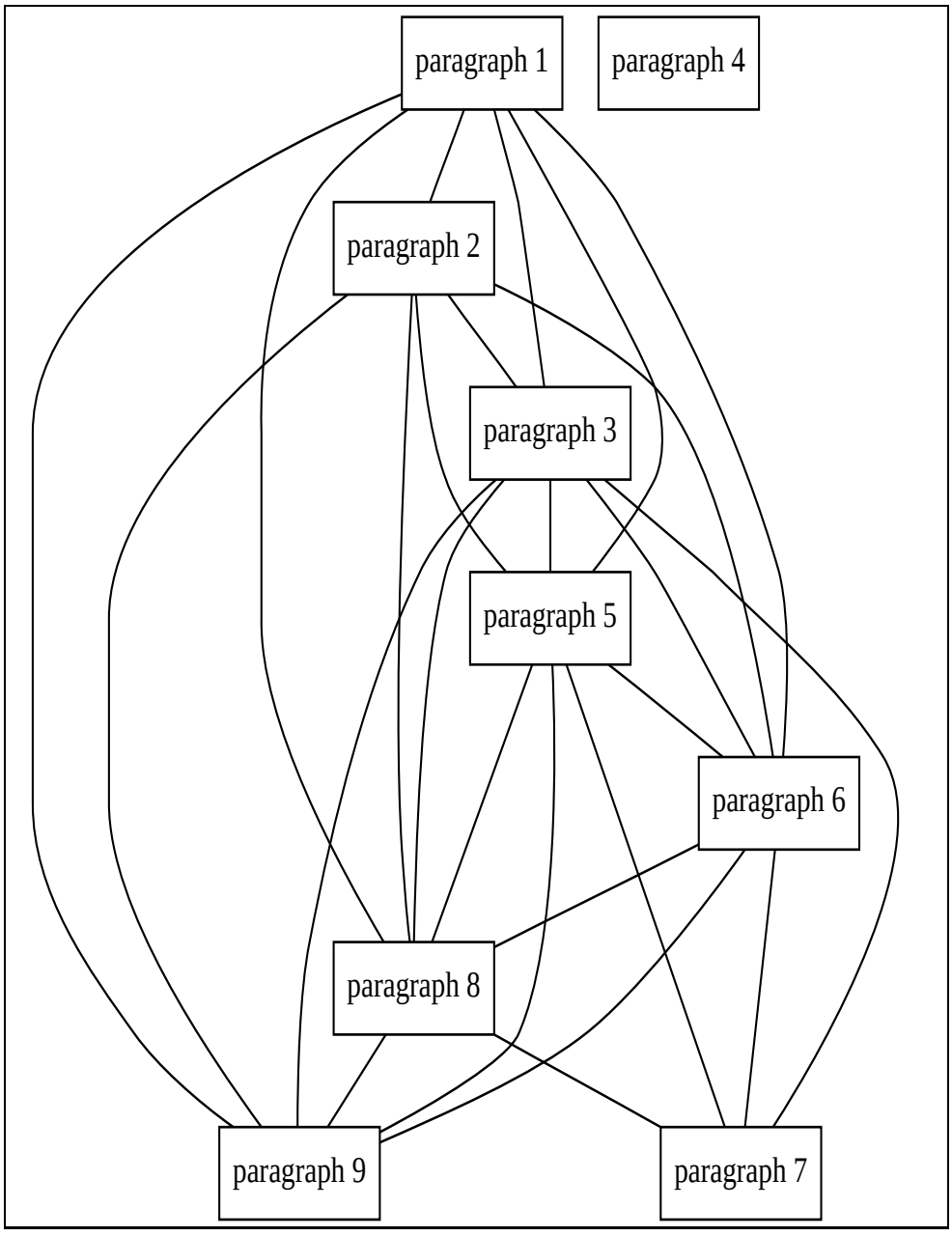


Figure B.1: The collapsed lexical graph for the incoherent version of text 1.

paragraph. As a result, the text now reads much more smoothly, i.e. its coherence has been improved.

This is reflected in the collapsed lexical graph for this text. The graph shown in Figure B.2 now has the central paragraph, paragraph 3. Paragraph 4 is linked to paragraph 3 by the repetition of the word *risk*, and to paragraph 5 by the words *risk* and *risky*.

B.2 Text 2

Text 2 is a short example that in the incoherent version lacks the main component, but it has the central paragraph, i.e. paragraph 3. This example illustrates the importance of the Main Component Hypothesis. The lexical graph for the incoherent version is shown in Figure B.3.

The lexical graph for the coherent version of text 2 is shown in Figure B.4.

B.3 Text 3

This text again is a short text that lacks the main component. In addition, it also lacks the central paragraph. The incoherence occurs at paragraph 1 (we have already seen this text in chapter 3, Figure 3.3), as we discovered when we attempted to traverse the graph in search of the main component. For this text, no paragraphs are reachable from the first paragraph, hence the problem is at the beginning of the text, in paragraph 1.

The lexical graph of this text is shown in Figure B.5. Note the sole node, *income*, in the first paragraph.

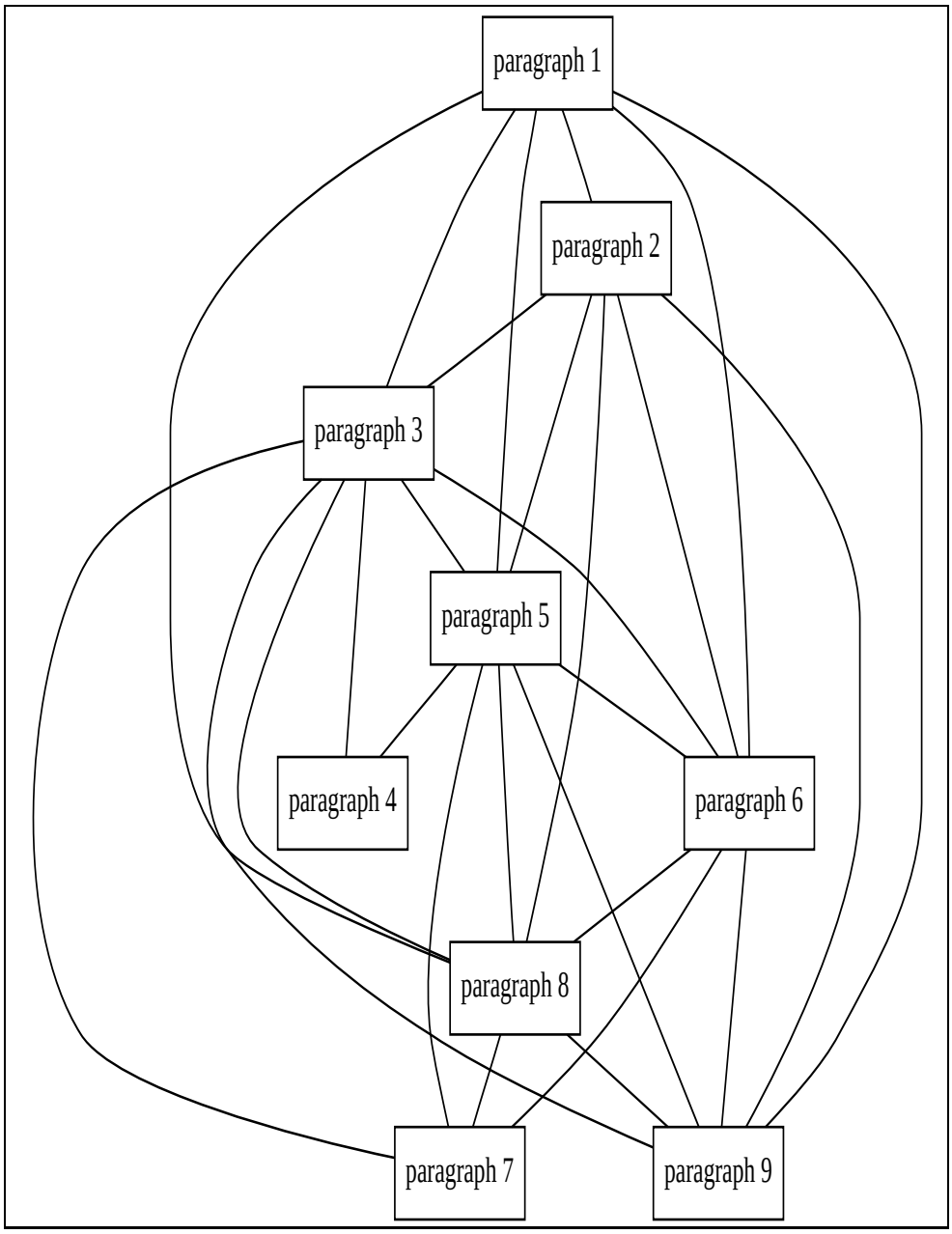


Figure B.2: The collapsed lexical graph for the coherent version of text 1.

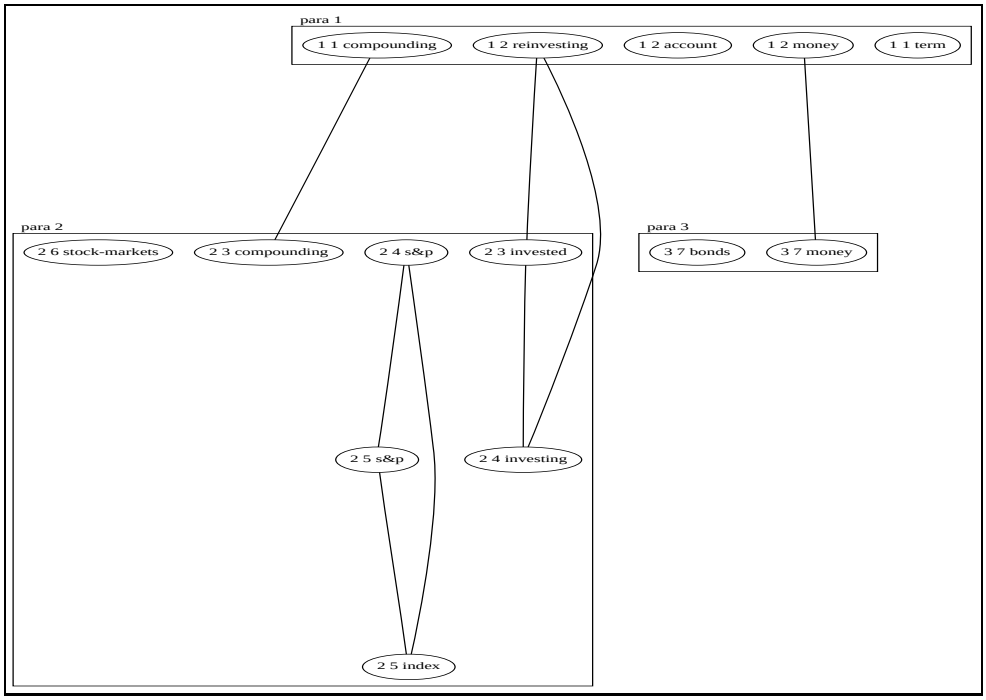


Figure B.3: The lexical graph for the incoherent version of text 2.

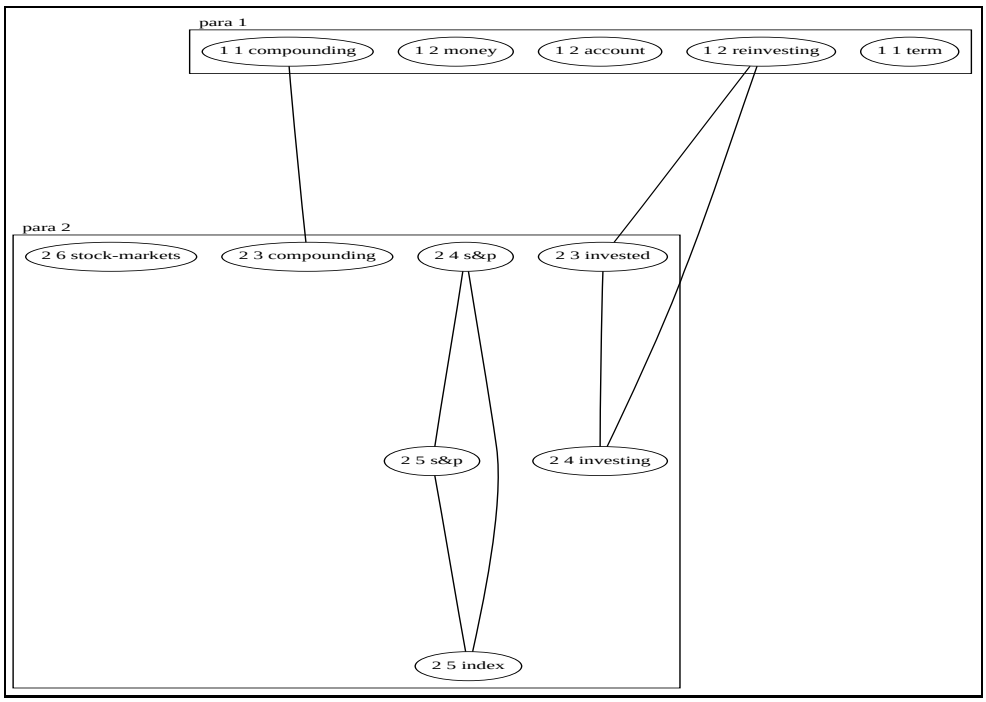


Figure B.4: The lexical graph for the coherent version of text 2.

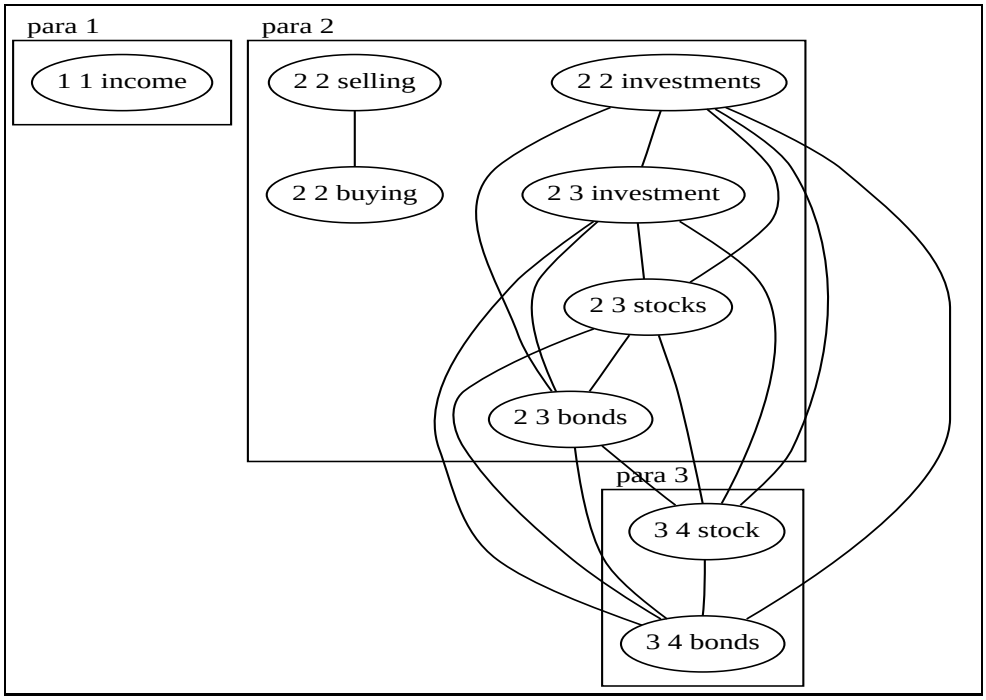


Figure B.5: The lexical graph for the incoherent version of text 3.

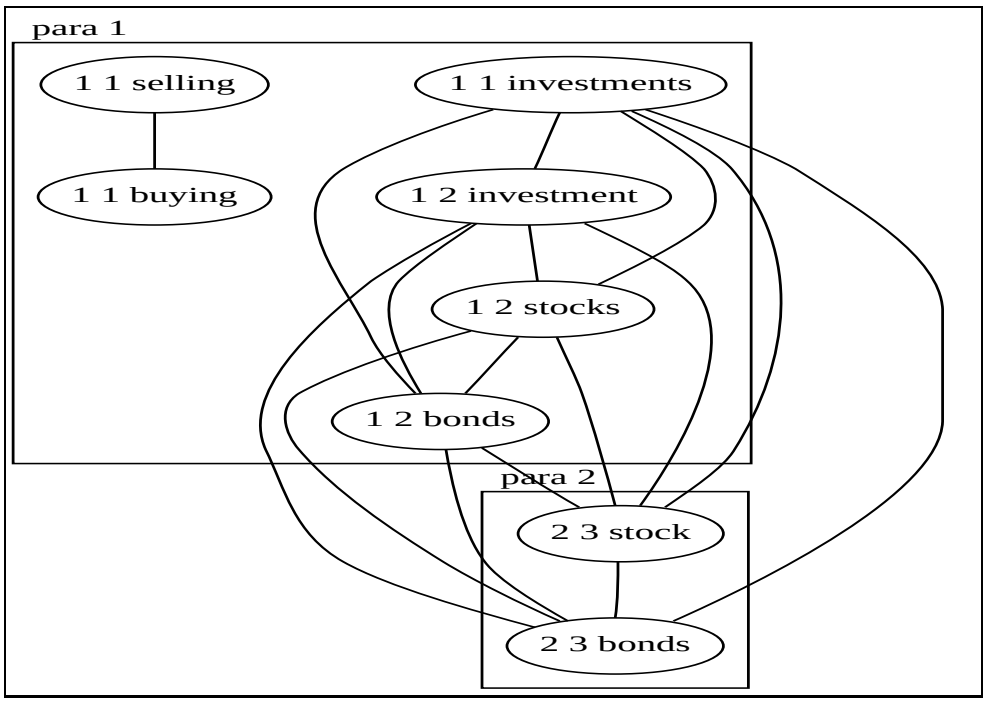


Figure B.6: The lexical graph for the coherent version of text 3.

The lexical graph for the coherent version of this text is shown in Figure B.6. The main component is tightly connected and consists of all the lexical items in the graph except *buying*, which is not related to the other items and is therefore placed in a singleton component.

B.4 Text 4

This is a slightly longer text, consisting of five paragraphs, and also familiar from chapter 3 (cf. Figure 3.9). Paragraph 4 of this text is not lexically related to any other paragraph of the text. Hence, the text lacks the central paragrah.

The lexical graph for this text is shown in Figure B.7. Since the text contains a few more words, its lexical graph is somewhat detailed. To enhance the readability of this example, we also include the collapsed graph. It is shown in Figure B.8.

The lexical graph for the coherent version of this text is presented in Figure B.9. The collapsed graph of this text is shown in Figure B.10.

B.5 Text 5

This is an incoherent text for which our analysis was unable to find the problem. In this text, we have arbitrarily deleted one paragraph, paragraph 2, and left the remaining text intact. This kind of problem might occur while carelessly editing by cut and paste. Our lexical analysis would find the lexical relation between *stocks* in paragraph 3 and *stocks* in paragraph 1, therefore linking these together.

Since this text is long and has many lexical links, the lexical graph is too large to reproduce it here. We are including the collapsed lexical graph (Figure B.11), which is connected.

The collapsed lexical graph for the original version of text 5 is shown in Figure B.12. This version, too, is connected and hence deemed coherent by our model.

B.6 Text 6

Even though this text is seven paragraphs long, it contains relatively few domain-specific words and relatively few lexical links. For this reason, we decided to reproduce here the full lexical graph of this text.

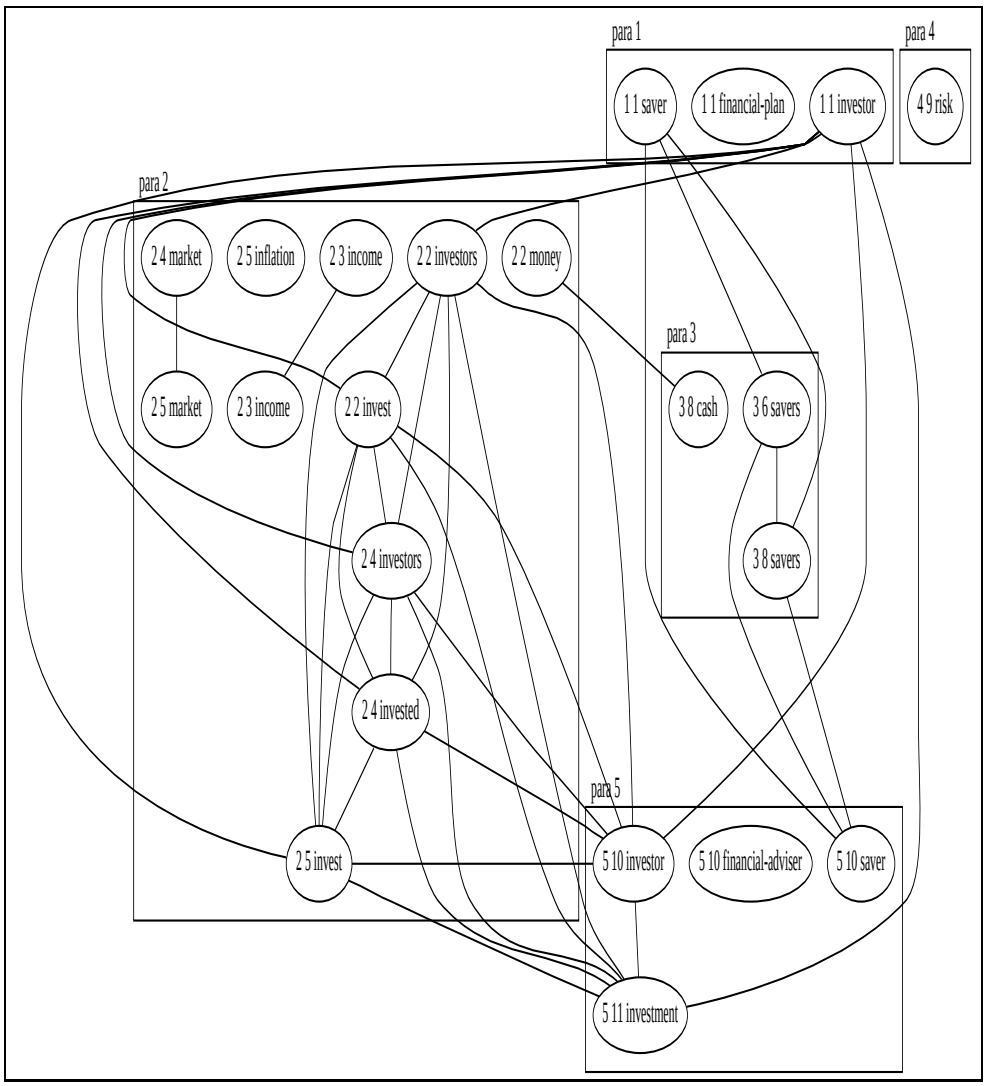


Figure B.7: The lexical graph for the incoherent version of text 4.

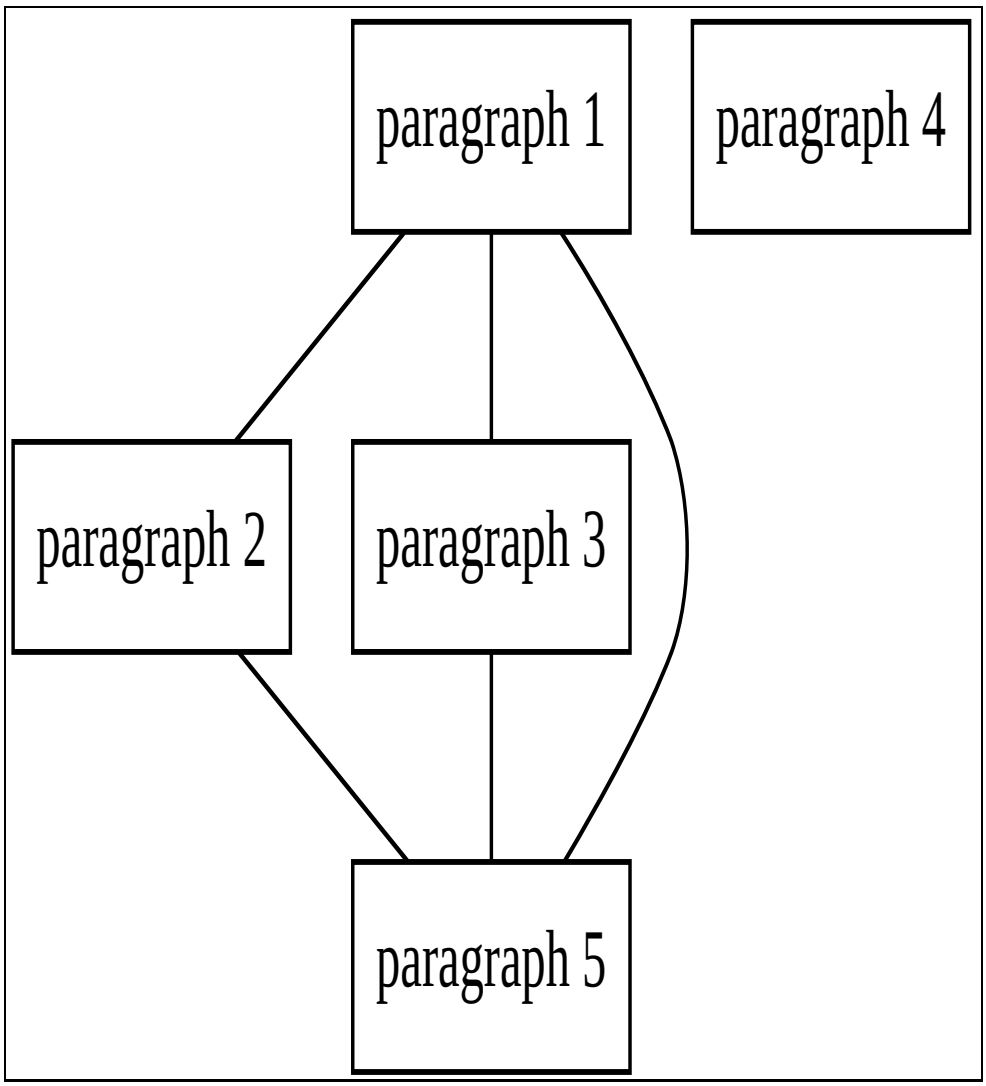


Figure B.8: The collapsed lexical graph for the incoherent version of text 4.

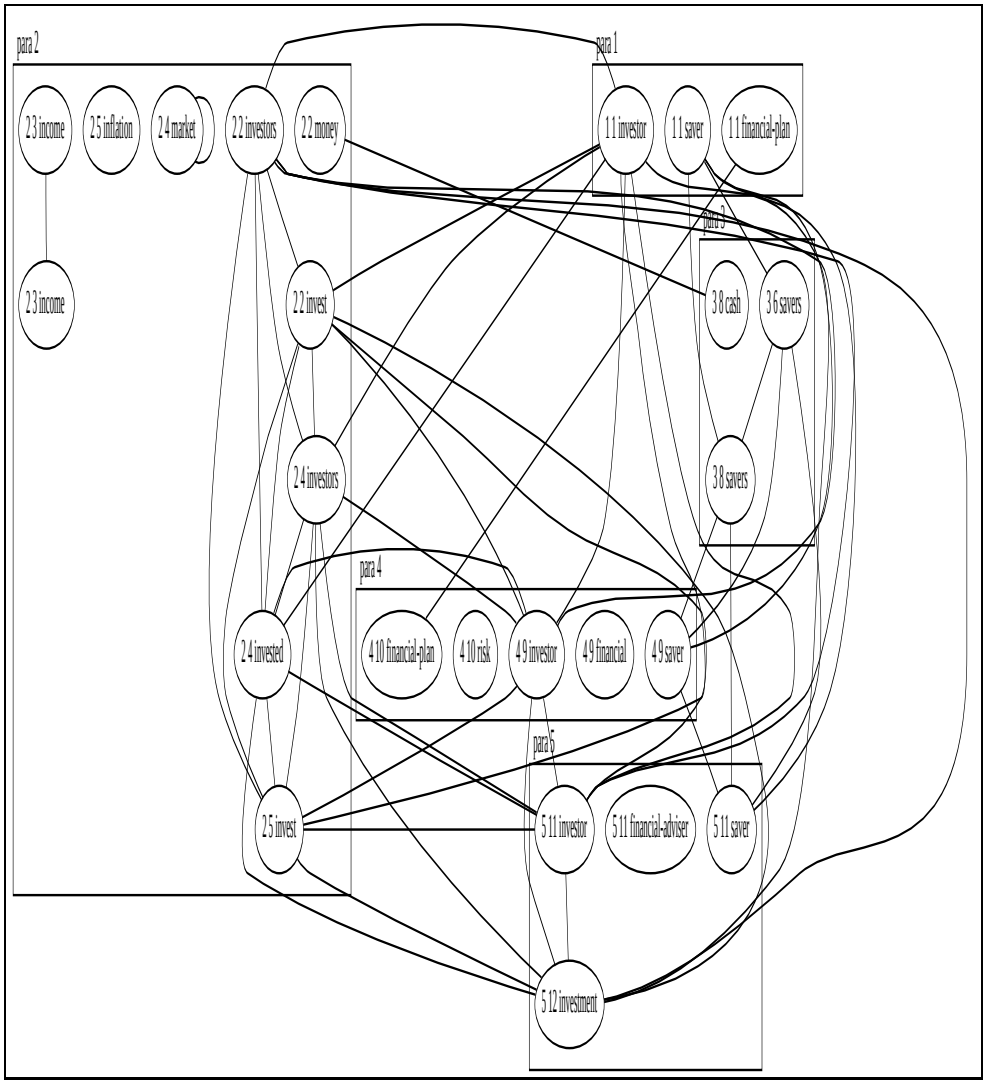


Figure B.9: The lexical graph for the coherent version of text 4.

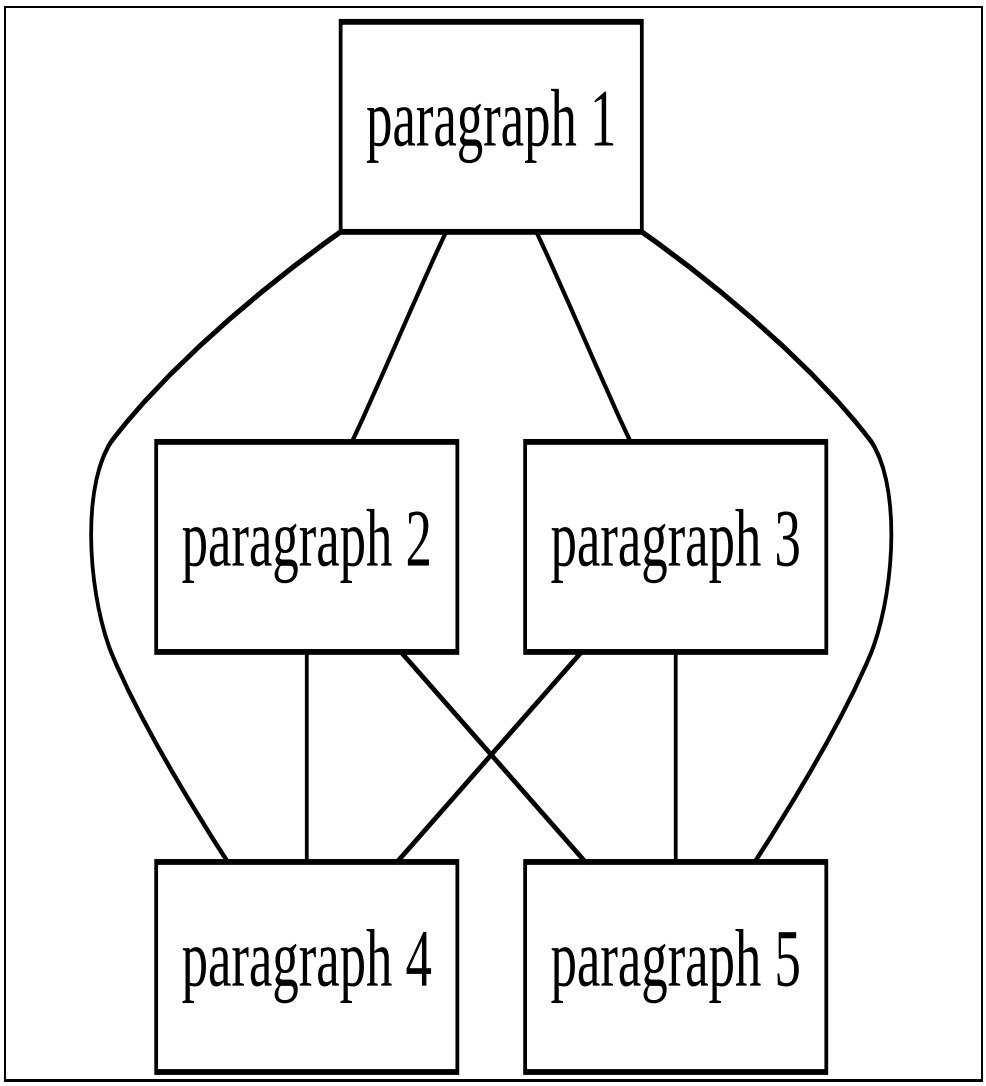


Figure B.10: The collapsed lexical graph for the coherent version of text 4.

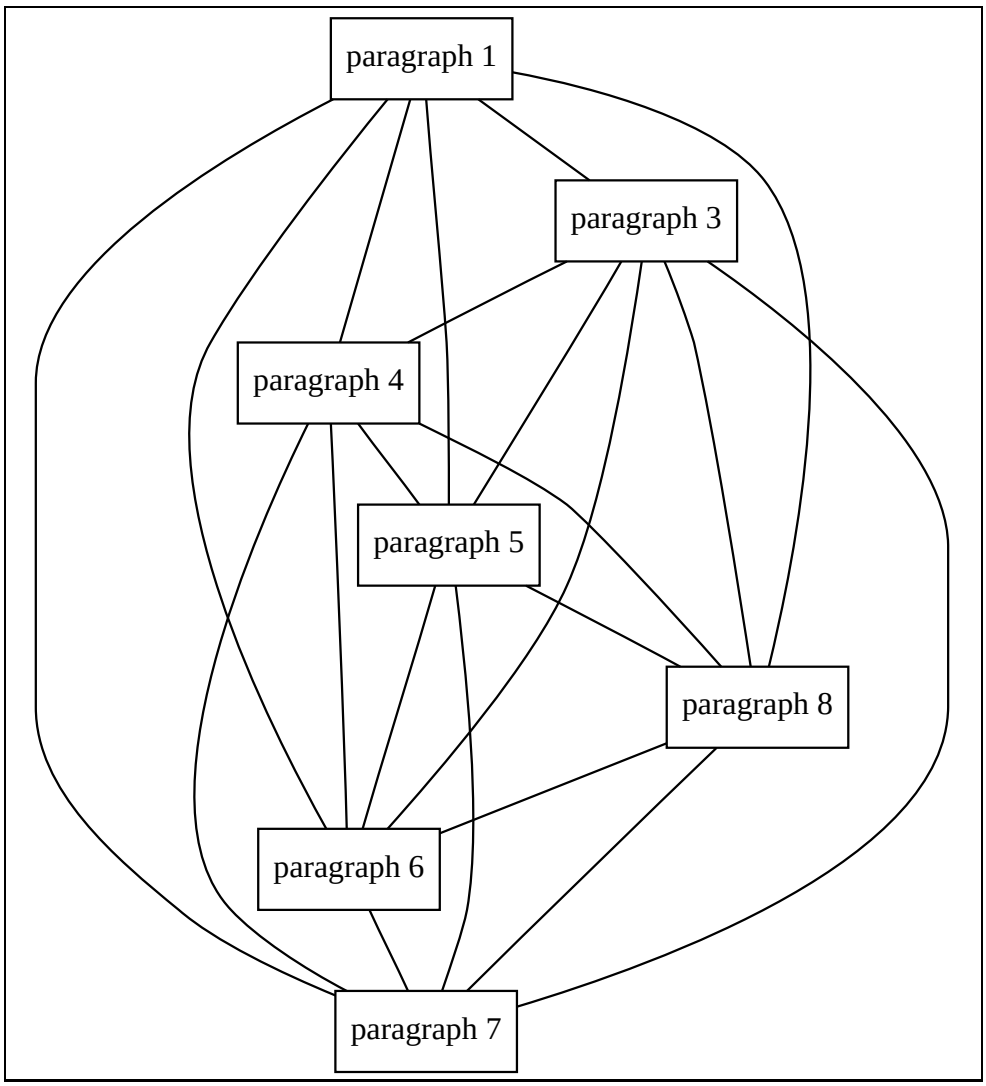


Figure B.11: The lexical graph for the incoherent version of text 5.

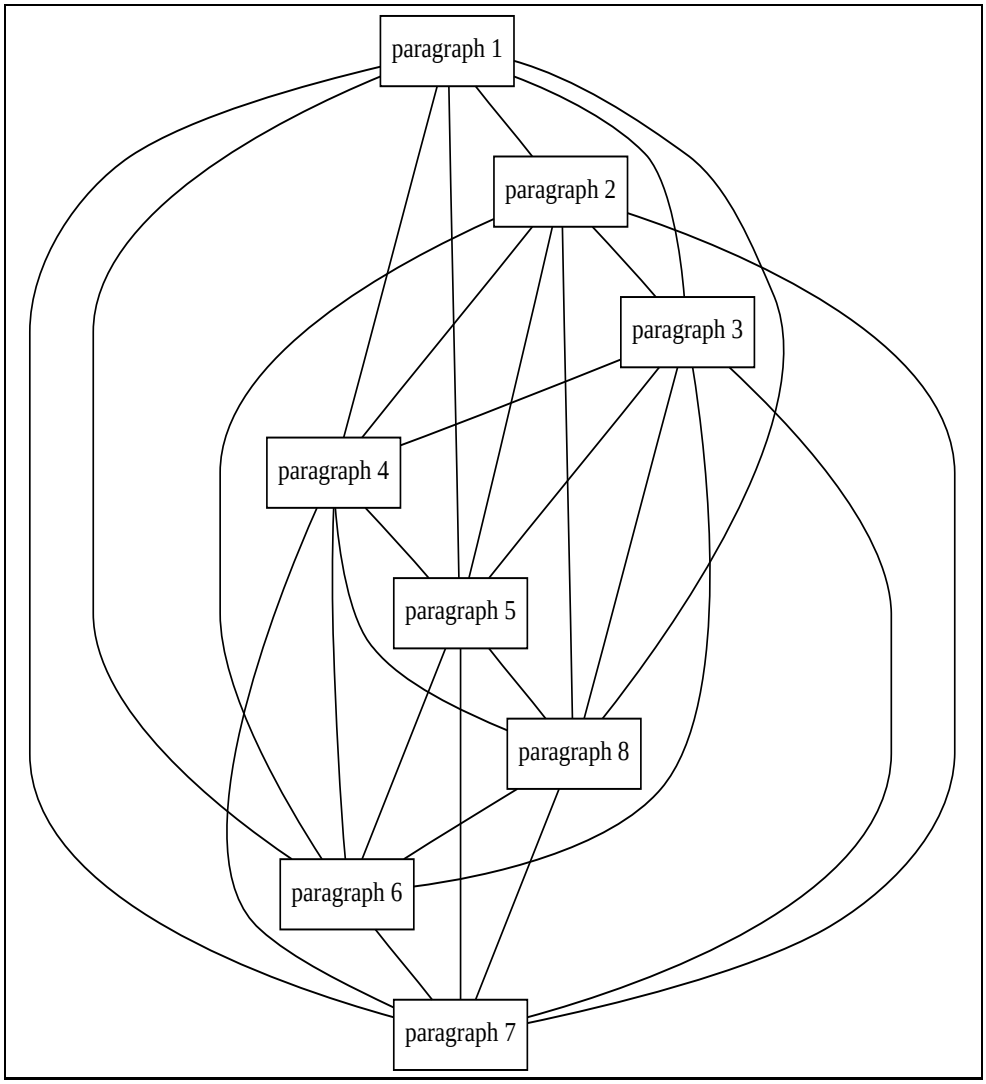


Figure B.12: The lexical graph for the coherent version of text 5.

The lexical graph of the incoherent version is shown in Figure B.13. The last paragraph of this text is not lexically related to any other paragraph in the text.

The collapsed graph, shown in Figure B.14 confirms this.

The lexical graph of the coherent version is shown in Figure B.15.

The collapsed lexical graph for this text is shown in Figure B.16. Paragraph 1 is the central paragraph of this text.

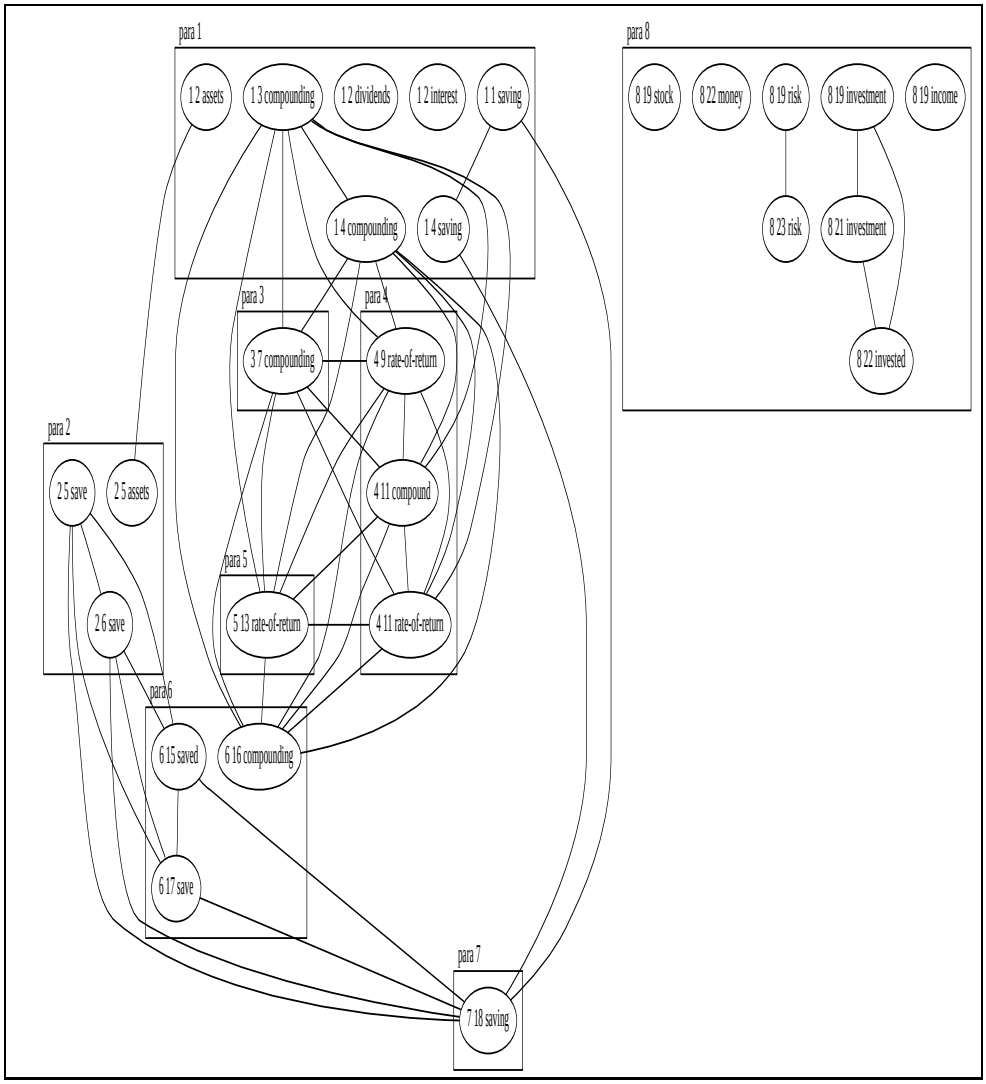


Figure B.13: The lexical graph for the incoherent version of text 6.

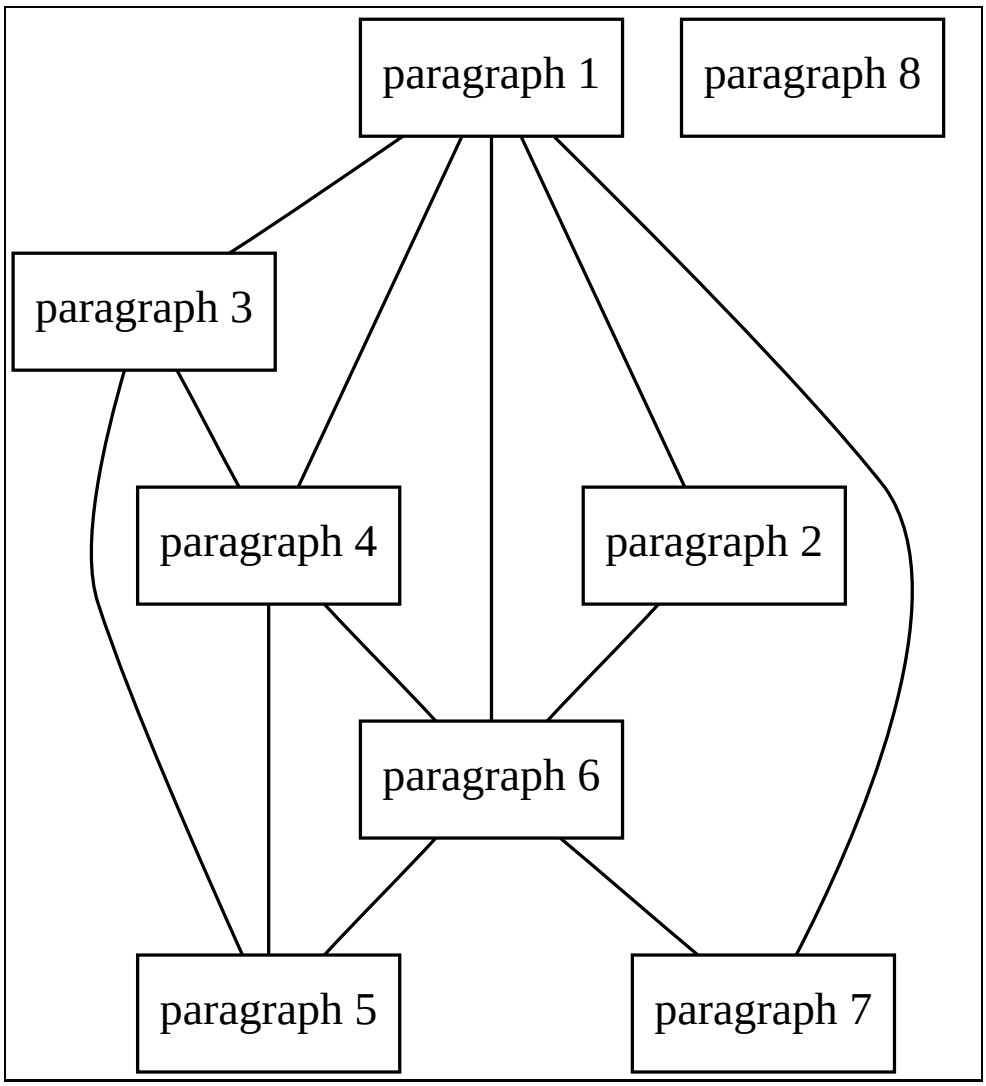


Figure B.14: The lexical graph for the incoherent version of text 6.

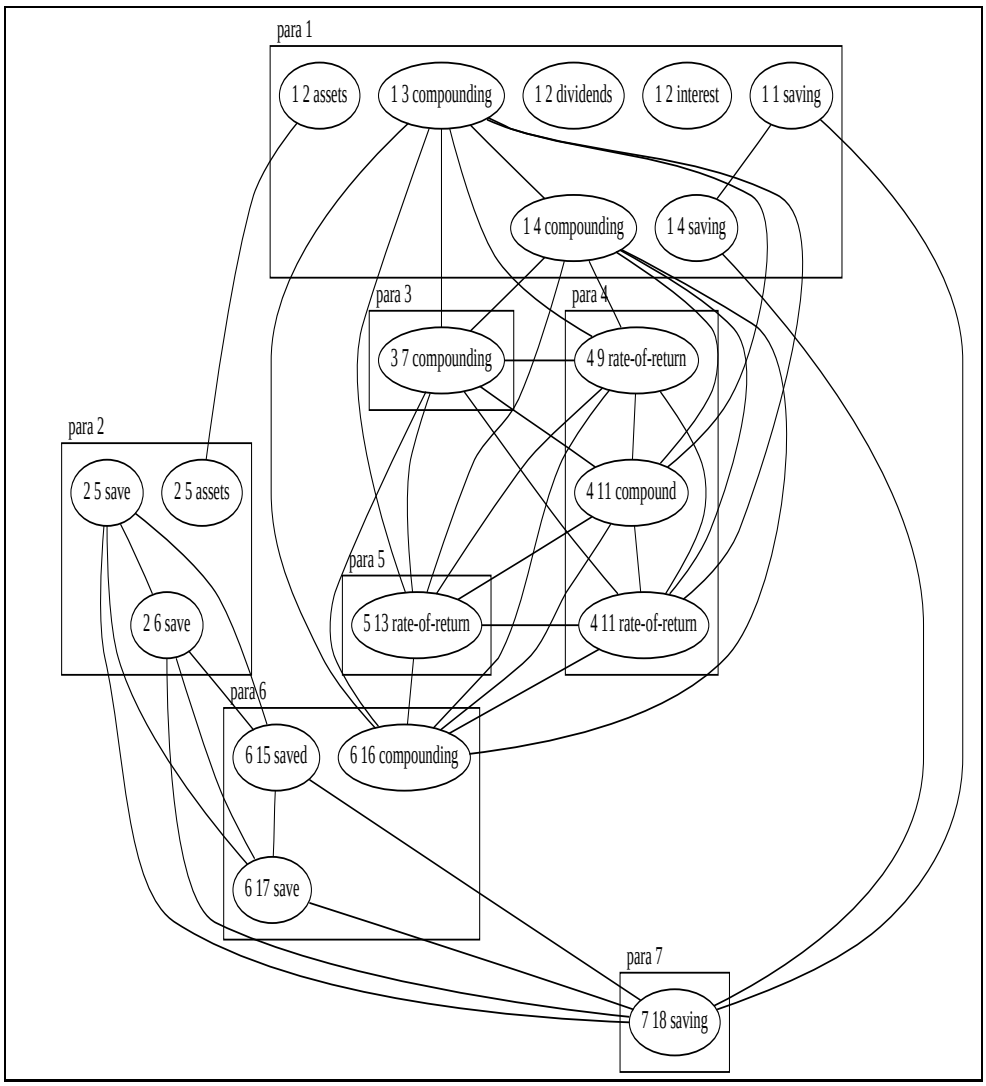


Figure B.15: The lexical graph for the coherent version of text 6.

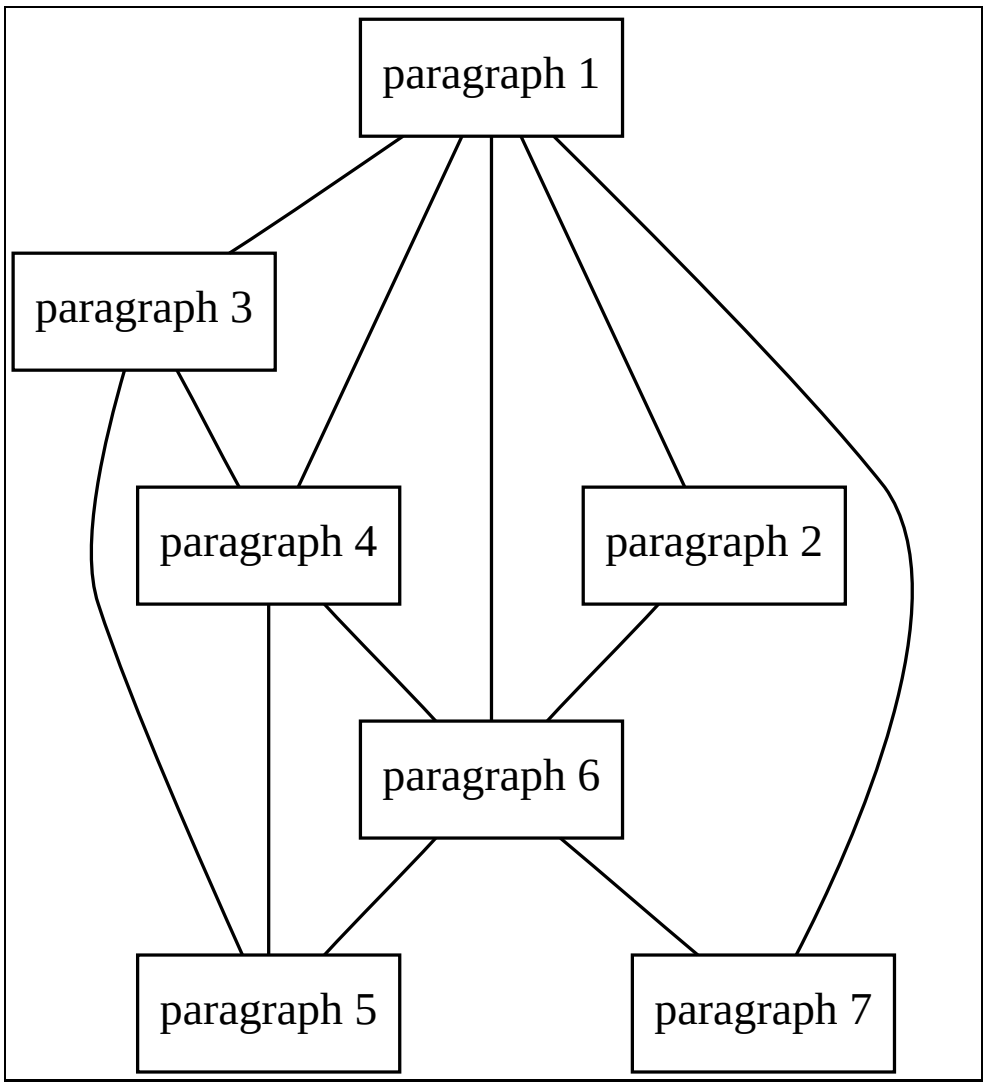


Figure B.16: The lexical graph for the coherent version of text 6.

Appendix C

The financial thesaurus

This appendix contains the financial thesaurus that we used for analyzing our texts.

Since our chosen domain is financial advice texts aimed at individual investors, certain terms that might at first seem appropriate for the thesaurus have not been included. Terms such as *economy* or *deficit*, while having to do with finances, are not really related to the domain of investing. Hence, these terms are not included in our thesaurus.

We discovered that some words became ubiquitous in our domain. For example, words such as *profit* are not represented in the thesaurus at all.

We only included the root forms in the thesaurus. Hence, the word *save* is included while *savings* is not. The derived forms are handled by morphological analysis (described in section 3.4), and the appropriate relation is computed properly.

Certain lexical items are represented with hyphens in our thesaurus. These items can be recognized in texts whether a hyphen or a blank exists between the connecting words.

This appendix shows words which are related and the type of lexical relation between them. *scr* stands for systematically classifiable relation, described in section 2.3. We include this type of relation if both elements are in an IS-A relation with a third item, such as *gold* and *silver* with *precious metals*.

word 1	word 2	relation
account	bank	PART-OF
asset	investment	IS-A
asset	bond	IS-A
asset	stock	IS-A
asset	security	IS-A
asset	real-estate	IS-A
asset	asset-allocation	PART-OF
asset-allocation	asset	PART-OF
balanced-fund	mutual-fund	IS-A
balanced-fund	bond-fund	scr
balanced-fund	dividend-fund	scr
balanced-fund	equity-fund	scr
balanced-fund	income-fund	scr
balanced-fund	index-fund	scr
balanced-fund	money-market-fund	scr
balanced-fund	mortgage-fund	scr
balanced-fund	stock-fund	scr
balanced-fund	sector-fund	scr
balanced-fund	specialty-fund	scr

word 1	word 2	relation
bank	account	PART-OF
bear-market	bull-market	antonym
bear-market	market-trend	IS-A
bear-market	stock-market	IS-A
blue-chip	stock	IS-A
bond	asset	IS-A
bond	investment	IS-A
bond	stock	scr
bond	interest	PART-OF
bond	yield	PART-OF
bond	fixed-income	IS-A
bond	junk-bond	IS-A
bond	convertible-bond	IS-A
bond-fund	mutual-fund	IS-A
bond-fund	balanced-fund	scr
bond-fund	dividend-fund	scr
bond-fund	equity-fund	scr
bond-fund	income-fund	scr
bond-fund	index-fund	scr
bond-fund	money-market-fund	scr
bond-fund	mortgage-fund	scr
bond-fund	stock-fund	scr
bond-fund	sector-fund	scr

word 1	word 2	relation
bond-fund	specialty-fund	scr
bond-fund	mutual-fund	IS-A
broker	brokerage-house	PART-OF
broker	financial-adviser	synonym
brokerage-firm	brokerage-house	synonym
brokerage-house	broker	PART-OF
brokerage-house	brokerage-firm	synonym
budget	financial-plan	PART-OF
bull-market	market-trend	IS-A
bull-market	stock-market	IS-A
bull-market	bear-market	antonym
buy	sell	antonym
capital-gains	income	IS-A
cash	money	synonym
cheap	expensive	antonym
cheap	inexpensive	synonym
commodity	investment	IS-A
common-share	share	IS-A
compound	interest	PART-OF
compound	rate-of-return	PART-OF
convertible-bond	bond	IS-A
convertible-bond	stock	IS-A
currency	dollar	IS-A
currency	currency-exchange	PART-OF

word 1	word 2	relation
currency-exchange	currency	PART-OF
deflation	inflation	antonym
distribution	mutual-fund	PART-OF
dividend	share	PART-OF
dividend	stock	PART-OF
dividend	income	IS-A
dividend-fund	mutual-fund	IS-A
dividend-fund	balanced-fund	scr
dividend-fund	bond-fund	scr
dividend-fund	equity-fund	scr
dividend-fund	income-fund	scr
dividend-fund	index-fund	scr
dividend-fund	money-market-fund	scr
dividend-fund	mortgage-fund	scr
dividend-fund	stock-fund	scr
dividend-fund	sector-fund	scr
dividend-fund	specialty-fund	scr
djia	index	IS-A
dollar	currency	IS-A
dollar	yen	scr
dollar	franc	scr
dollar	pound	scr
dollar-cost-averaging	financial-plan	PART-OF
eps	share	PART-OF

word 1	word 2	relation
equity	stock	IS-A
equity-fund	mutual-fund	IS-A
equity-fund	balanced-fund	scr
equity-fund	bond-fund	scr
equity-fund	dividend-fund	scr
equity-fund	income-fund	scr
equity-fund	index-fund	scr
equity-fund	money-market-fund	scr
equity-fund	mortgage-fund	scr
equity-fund	stock-fund	scr
equity-fund	sector-fund	scr
equity-fund	specialty-fund	scr
estate	estate-plan	PART-OF
estate-plan	estate	PART-OF
estate-plan	financial-plan	IS-A
expensive	cheap	antonym
expensive	inexpensive	antonym
family-of-funds	mutual-fund	PART-OF
family-of-funds	fund-family	synonym
finance	financial	pleonym
financial	finance	pleonym
financial-adviser	broker	synonym
financial-goal	financial-plan	PART-OF

word 1	word 2	relation
financial-plan	financial-goal	PART-OF
financial-plan	budget	PART-OF
financial-plan	dollar-cost-averaging	PART-OF
financial-plan	value-averaging	PART-OF
financial-plan	tax-plan	IS-A
financial-plan	estate-plan	IS-A
fixed-income	bond	IS-A
franc	dollar	scr
fund-family	family-of-funds	synonym
fund-family	mutual-fund	PART-OF
gold	precious-metal	IS-A
gold	silver	scr
hedge	inflation	scr
homeowners-insurance	insurance	IS-A
income	capital-gains	IS-A
income	dividend	IS-A
income	interest	IS-A
income	interest	IS-A
income-fund	mutual-fund	IS-A
income-fund	balanced-fund	scr
income-fund	bond-fund	scr
income-fund	dividend-fund	scr

word 1	word 2	relation
income-fund	equity-fund	scr
income-fund	index-fund	scr
income-fund	money-market-fund	scr
income-fund	mortgage-fund	scr
income-fund	stock-fund	scr
income-fund	sector-fund	scr
income-fund	specialty-fund	scr
index	s&p	IS-A
index	djia	IS-A
index	tse100	IS-A
index-fund	mutual-fund	IS-A
index-fund	balanced-fund	scr
index-fund	bond-fund	scr
index-fund	dividend-fund	scr
index-fund	equity-fund	scr
index-fund	income-fund	scr
index-fund	money-market-fund	scr
index-fund	mortgage-fund	scr
index-fund	stock-fund	scr
index-fund	sector-fund	scr
index-fund	specialty-fund	scr
inexpensive	cheap	synonym
inexpensive	expensive	antonym
inflation	hedge	scr

word 1	word 2	relation
inflation	interest	scr
inflation	deflation	antonym
insurance	insure	pleonym
insurance	term	IS-A
insurance	whole-life	IS-A
insurance	life-insurance	IS-A
insurance	homeowners-insurance	IS-A
insure	insurance	pleonym
interest	bond	PART-OF
interest	compound	PART-OF
interest	income	IS-A
interest	inflation	scr
interest	income	IS-A
interest	interest-rate	PART-OF
interest-rate	interest	PART-OF
invest	reinvest	pleonym
invest	investment	pleonym
invest	investor	pleonym
investment	invest	pleonym
investment	investor	pleonym
investment	asset	IS-A
investment	bond	IS-A

word 1	word 2	relation
investment	stock	IS-A
investment	security	IS-A
investment	real-estate	IS-A
investment	commodity	IS-A
investor	invest	pleonym
investor	investment	pleonym
investor	trader	antonym
junk-bond	bond	IS-A
life-insurance	insurance	IS-A
load-fund	mutual-fund	IS-A
load-fund	no-load-fund	antonym
market	stock-market	synonym
market-trend	bear-market	IS-A
market-trend	bull-market	IS-A
money	cash	synonym
money-market-fund	mutual-fund	IS-A
money-market-fund	balanced-fund	scr
money-market-fund	bond-fund	scr
money-market-fund	dividend-fund	scr
money-market-fund	equity-fund	scr
money-market-fund	income-fund	scr
money-market-fund	index-fund	scr

word 1	word 2	relation
money-market-fund	mortgage-fund	scr
money-market-fund	stock-fund	scr
money-market-fund	sector-fund	scr
money-market-fund	specialty-fund	scr
mortgage	mortgage-backed-security	scr
mortgage-backed-security	mortgage	scr
mortgage-fund	mutual-fund	IS-A
mortgage-fund	balanced-fund	scr
mortgage-fund	bond-fund	scr
mortgage-fund	dividend-fund	scr
mortgage-fund	equity-fund	scr
mortgage-fund	income-fund	scr
mortgage-fund	index-fund	scr
mortgage-fund	money-market-fund	scr
mortgage-fund	stock-fund	scr
mortgage-fund	sector-fund	scr
mortgage-fund	specialty-fund	scr
mutual-fund	equity-fund	IS-A
mutual-fund	stock-fund	IS-A
mutual-fund	dividend-fund	IS-A
mutual-fund	bond-fund	IS-A

word 1	word 2	relation
mutual-fund	money-market-fund	IS-A
mutual-fund	balanced-fund	IS-A
mutual-fund	income-fund	IS-A
mutual-fund	mortgage-fund	IS-A
mutual-fund	specialty-fund	IS-A
mutual-fund	sector-fund	IS-A
mutual-fund	index-fund	IS-A
mutual-fund	fund-family	PART-OF
mutual-fund	family-of-funds	PART-OF
mutual-fund	load-fund	IS-A
mutual-fund	no-load-fund	IS-A
mutual-fund	distribution	PART-OF
no-load-fund	mutual-fund	IS-A
no-load-fund	load-fund	antonym
penny-stock	stock	IS-A
pension	retirement-plan	PART-OF
pound	dollar	scr
precious-metal	gold	IS-A
precious-metal	silver	IS-A
preferred-share	share	IS-A
rate-of-return	compound	PART-OF
real-estate	asset	IS-A

word 1	word 2	relation
real-estate	investment	IS-A
reinvest	invest	pleonym
retirement-plan	pension	PART-OF
retirement-plan	rrsp	PART-OF
risk	risky	pleonym
risky	risk	pleonym
rrsp	retirement-plan	PART-OF
s&p	index	IS-A
save	saver	pleonym
saver	save	pleonym
sector-fund	mutual-fund	IS-A
sector-fund	balanced-fund	scr
sector-fund	bond-fund	scr
sector-fund	dividend-fund	scr
sector-fund	equity-fund	scr
sector-fund	income-fund	scr
sector-fund	index-fund	scr
sector-fund	money-market-fund	scr
sector-fund	mortgage-fund	scr
sector-fund	stock-fund	scr
sector-fund	specialty-fund	scr
security	asset	IS-A
security	investment	IS-A
sell	buy	antonym

word 1	word 2	relation
share	shareholder	pleonym
share	stock	IS-A
share	dividend	PART-OF
share	common-share	IS-A
share	preferred-share	IS-A
share	eps	PART-OF
shareholder	share	pleonym
silver	gold	scr
silver	precious-metal	IS-A
specialty-fund	mutual-fund	IS-A
specialty-fund	balanced-fund	scr
specialty-fund	bond-fund	scr
specialty-fund	dividend-fund	scr
specialty-fund	equity-fund	scr
specialty-fund	income-fund	scr
specialty-fund	index-fund	scr
specialty-fund	money-market-fund	scr
specialty-fund	mortgage-fund	scr
specialty-fund	stock-fund	scr
specialty-fund	sector-fund	scr
stock	asset	IS-A
stock	convertible-bond	IS-A
stock	investment	IS-A
stock	share	IS-A

word 1	word 2	relation
stock	equity	IS-A
stock	dividend	PART-OF
stock	blue-chip	IS-A
stock	penny-stock	IS-A
stock	bond	scr
stock-fund	mutual-fund	IS-A
stock-fund	balanced-fund	scr
stock-fund	bond-fund	scr
stock-fund	dividend-fund	scr
stock-fund	equity-fund	scr
stock-fund	income-fund	scr
stock-fund	index-fund	scr
stock-fund	money-market-fund	scr
stock-fund	mortgage-fund	scr
stock-fund	sector-fund	scr
stock-fund	specialty-fund	scr
stock-market	market	synonym
stock-market	bull-market	IS-A
stock-market	bear-market	IS-A
tax	tax-plan	PART-OF

word 1	word 2	relation
tax	taxable	pleonym
tax	tax-exempt	antonym
tax-exempt	tax	antonym
tax-exempt	taxable	antonym
tax-plan	financial-plan	IS-A
tax-plan	tax	PART-OF
taxable	tax	pleonym
taxable	tax-exempt	antonym
term	insurance	IS-A
trade	trader	pleonym
trader	investor	antonym
trader	trade	pleonym
tse100	index	IS-A
value-averaging	financial-plan	PART-OF
whole-life	insurance	IS-A
yen	dollar	scr
yield	bond	PART-OF

References

[Agar and Hobbs, 1981] M. Agar and J. Hobbs. Interpreting discourse: Coherence and the analysis of entographic interviews, 1981.

[Bateman and Paris, 1989] J. Bateman and C. Paris. Phrasing a text in terms the user can understand. In *Eleventh International Joint Conference on Artificial Intelligence*, 1989.

[Beckwith *et al.*, 1991] Richard Beckwith, Christian Fellbaum, Derek Gross, and George Miller. Wordnet: A lexical database organized on psycholinguistic principles. In Uri Zernik, editor, *Lexical Acquisition: Exploiting on- line resources to build a lexicon*, pages 211–232. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

[Bencze, 1985] L. Bencze. Conscious tradition, unconscious construction or subconscious metaphors? Certain levels of text cohesion and coherence. In E. Sožer, editor, *Papers in textlinguistics*. Helmut Buske Verlag Hamburg, 1985.

[Blakemore, 1987] D. Blakemore. *Semantic constraints on relevance.* Basil Blackwell, 1987.

[Clark and Marshall, 1983] H. H. Clark and C. R. Marshall. Definite reference and mutual knowledge. In A. K. Joshi, B. L. Weber, and I. Sag, editors, *Elements of discourse understanding.* Cambridge University Press, 1983.

[Cohen, 1987] R. Cohen. Analyzing the structure of argumentative discourse. *Computational linguistics*, 13(1-2), 1987.

[Davis, 1994] S. Davis. *C++ for dummies.* IDG Books, 1994.

[Donaldson *et al.*, 1996] T. Donaldson, M. Makuta, and R. Cohen. An integrated approach to evaluating text coherence and its application to the prevention of reader misconceptions. *Proc. of AAAI Workshop on Preventing, Detecting, and Repairing Miscommunication in Discourse*, 1996.

[Flesch, 1948] Flesch. A new readability yardstick. *Journal of Applied Psychology*, 34, 1948.

[Friedriksen, 1981] C. H. Friedriksen. Inference in preschool children's conversation—a cognitive perspective. In *Ethnography and language in educational settings.* Norwood, NJ, 1981.

[Givón, 1992] T. Givón. The grammar of referential coherence as mental processing instructions. *Linguistics*, 30(1), 1992.

[GramCheck, 1997] GramCheck. Gramcheck: A bilingual grammar and style checker, 1997.

[Green, 1992] S. J. Green. A functional theory of style for natural language generation. Master's thesis, Faculty of Mathematics, University of Waterloo, 1992. Also available as University of Waterloo Faculty of Mathematics Technical Report CS-92-48.

[Grosz and Sidner, 1986] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), November 1986.

[Halliday and Hasan, 1976] M. A. K. Halliday and R. Hasan. *Cohesion in English.* Longman, 1976.

[Hearst, 1994] M. Hearst. *Multi-paragraph segmentation of expository text.* ACL, 1994.

[Hearst, 1997] M. Hearst. Texttiling: segmenting text into multiple paragraph subtopic passages. *Computational linguistics*, 1997.

[Heidorn *et al.*, 1982] G. E. Heidorn, K. Jensen, L. A. Miller, R. J. Byrd, and M. S. Chodorow. The epistle text-critiquing system. *IBM Systems journal*, 21(3):305–326, 1982.

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

[Hirst and St-Onge, 1995] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet*. MIT Press, 1995.

[Hobbs, 1976] J. Hobbs. A computational approach to discourse analysis. *Technical report No. 76-2*, 1976.

[Hobbs, 1985] J. Hobbs. On the coherence and structure of discourse. *Technical report SLI-85-37, Center for the Study of Language and Information, Stanford University*, 1985.

[Hoey, 1991] M. Hoey. *Patterns of lexis in text.* Cambridge University Press, 1991.

[Hörmann, 1981] H. Hörmann. *Enführung in die Psycholinguistic.* Darmstadt, Wissenschaftliche Buchgesselschaft, 1981.

[Hovy, 1988] E. H. Hovy. *Generating natual language under pragmatic constraints.* Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, 1988.

[Johnson-Laird, 1983] P. N. Johnson-Laird. *Mental models.* Cambridge University Press, 1983.

[Kerrigan, 1974] W. J. Kerrigan. *Writing to the point: Six basic steps.* Harcourt Brace Jovanovic Inc., 1974.

[Kintsch and Greene, 1978] W. Kintsch and E. Greene. The role of culture-specific schemata in the comprehension and recall of stories. *Discourse processes*, 1, 1978.

[Kintsch and van Dijk, 1977] W. Kintsch and T. A. van Dijk. Toward a model of text comprehension and production. *Psychological review*, 85:363–394, 1977.

[Kipfer, 1995] B. A. Kipfer. Personal communications, 1995.

[Kozima, 1993] H. Kozima. Text segmentation based on similarity between words. *Proceedings of the 31st annual meeting of the Association of Computational Linguistics*, 1993.

[Malone, 1988] J. L. Malone. *The science of linguistics in the art of translation: Some tools from linguistics for the analysis and practice of translation.* SUNY Press, 1988.

[Mann and Thompson, 1983] W. C. Mann and S. A. Thompson. Relational propositions in discourse. *Information Sciences Institute, University of Southern California*, November 1983.

[Mann and Thompson, 1986] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Description and construction of text structures. *Technical report: ISI*, October 1986.

[Marcu, 1996] Daniel Marcu. Building up rhetorical structure trees. In *Proceedings of the AAAI 13th national conference on artificial intelligence*, pages 1069–1074, August 1996.

[Marcu, 1997] Daniel Marcu. From local to global coherence: A bottom-up approach to text planning. In *Proceedings of the AAAI 13th national conference on artificial intelligence*, 1997.

[McCoy and Cheng, 1988] K. F. McCoy and J. Cheng. Focus of attention: Constraining what can be said next. In *Natural language generation in artificial language and computational linguistics*. Kluwer Academic Publishers, 1988.

[McKeown, 1985] K. R. McKeown. Discourse strategies for generating natural-language text. *Artificial Intelligence*, 1985.

[Moore and Paris, 1989] J. Moore and C. Paris. Planning text for advisory dialogues. In *Association for computational linguistics*, 1989.

[Moore and Pollock, 1992] J. Moore and M. Pollock. A problem for rst: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4), 1992.

[Morris and Hirst, 1991] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of the text. *Computational Linguistics*, 17(1), 1991.

[Paris, 1988] C. Paris. Tailoring object descriptions to a user's level of expertise. *Computational Linguistics*, 14(3), 1988.

[Paris, 1993] C. Paris. *User Modelling in Text Generation.* Frances Pinter, 1993.

[Passoneau and Litman, 1993] R. Passoneau and D. Litman. Intention-based segentation: human reliability and correlation with linguistic cues. In *ACL*, 1993.

[Polanyi, 1988] L. Polanyi. A formal model of the structure of discourse. *Journal of pragmatics*, 12, 1988.

[Reichman, 1978] R. Reichman. Conversational coherency. *Cognitive Science*, 2, 1978.

[Richardson and Braden-Harder, 1988] S. D. Richardson and L. C. Braden-Harder. The experience of developing a large-scale natural language processing system: Critique. In *Proceedings of the 2nd conference on applied NLP*. Austin, 1988.

[Rickheit and Stroher, 1986] G. Rickheit and H. Stroher. Towards a functional approach to text connectedness. In J. Petröfi, editor, *Text connectedness from psychological point of view*. Helmut Buske Verlag Hamburg, 1986.

[Rickheit, 1991] G. Rickheit. *Kohärenzprozesse. Modellierung von Sprachverarbeitung in Texten und Discursen*. Westdeutscher Verlag, 1991.

[RightWriter, 1991] RightWriter. Improving readability of extension materials, 1991.

[Sanders *et al.*, 1992] T. J. M. Sanders, W. P. M. Spooren, and L. G. M. Noordman. Toward a taxonomy of coherence relations. *Discourse Processes*, 15, January–March 1992.

[Sanford and Garrod, 1981] A. J. Sanford and S. C. Garrod. *Understanding written language*. Chichester, NY, Willey, 1981.

[Sastry and Sastry, 1997] L. Sastry and V. Sastry. *Tcl/Tk cookbook*. http://www.dcc.ufba.br/ eprates/cookbook, 1997.

[Schank and Abelson, 1977] R. Schank and R. P. Abelson. *Scripts, goals, and understanding*. Hillside, NJ., Lawrence Erlbaum, 1977.

[Shavelson, 1988] R. J. Shavelson. *Statistical reasoning for behavioral sciences*. Allyn and Bacon, Inc., 1988.

[Skorochod'ko, 1972] E. F. Skorochod'ko. Adaptive method of automatic avstracting and indexing. In C. V. Frelman, editor, *Information Processing 71, Proceedings of the IFIP Congress 1971*. North-Holland Publishing Company, 1972.

[Sperling, 1997] D. Sperling. Dave's esl cafe, 1997.

[Srinivasdan, 1992] P. Srinivasdan. Thesaurus construction. In Frakes W. B. and Baeza-Yates R., editors, *Information retrieval. Data structures and algorightms.* Prentice Hall, 1992.

[Stoddard, 1991] S. Stoddard. *Text and texture: Patterns of cohesion.* Ablex Publishing Corporation, Norwood, NJ, 1991.

[StyleWriter, 1996] StyleWriter. http://www.isnnet.com/editor.html, 1996.

[Van Dijk, 1972] T. A. Van Dijk. *Some aspects of text grammars.* Mouton, The Hague, 1972.

[WinProof, 1997] WinProof. Methods and assumptions, 1997.

[WordNet, 1995] WordNet. *WordNet: A Lexical Database for English.* http://www.ito.darpa.mil/Summaries95/B370–Princeton.html, 1995.

[Zadrozny and Jensen, 1991] W. Zadrozny and B Jensen. Semantics of paragraphs. *Computational Linguistics*, 17(2), 1991.