Coherence in natural language: Data structures and applications

by

Florian Wolf

M.A. Linguistics

Ludwig-Maximilians-University of Munich, 2000

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN COGNITIVE SCIENCES
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2005

Signature of Author: ______________________________________
Department of Brain and Cognitive Sciences
1 November, 2004

Certified by: ______________________________________
Edward Gibson
Associate Professor of Cognitive Sciences
Thesis Supervisor

Accepted by: ______________________________________
Earl K. Miller
Picower Professor of Neuroscience
Chairman, Department Graduate Committee

Coherence in natural language: Data structures and applications

by

Florian Wolf

Submitted to the Department of Brain and Cognitive Sciences
on November 1, 2004 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Cognitive Sciences

ABSTRACT

The general topic of this thesis is coherence in natural language, where coherence refers to informational relations that hold between segments of a discourse. More specifically, this thesis aims to (1) develop criteria for a descriptively adequate data structure for representing discourse coherence; (2) test the influence of coherence on psycholinguistic processes, in particular, pronoun processing; (3) test the influence of coherence on the relative saliency of discourse segments in a text.

In order to address the first aim, a method was developed for hand-annotating a database of naturally occurring texts for coherence structures. The thus obtained database of coherence structures was used to test assumptions about descriptively adequate data structures for representing discourse coherence. In particular, the assumption that discourse coherence can be represented in trees was tested, and results suggest that more powerful data structures than trees are needed (labeled chain graphs, where the labels represent types of coherence relations, and an ordered array of nodes represents the temporal order of discourse segments in a text).

The second aim was addressed in an on-line comprehension and an off-line production experiment. Results from both experiments suggest that only a coherence-based account predicted the full range of observed data. In that account, the observed preferences in pronoun processing are not a result of pronoun-specific mechanisms, but a byproduct of more general cognitive mechanisms that operate when establishing coherence.

In order to address the third aim, layout-, word-, and coherence-based approaches to discourse segment ranking were compared to human rankings. Results suggest that word-based accounts provide a strong baseline, and that some coherence-based approaches best predict the human data. However, coherence-based algorithms that operate on trees did not perform as well as coherence-based algorithms that operate on more general graphs. It is suggested that that might in part be due to the fact that more general graphs are more descriptively adequate than trees for representing discourse coherence.

Thesis Supervisor: Edward Gibson

Title: Associate Professor of Cognitive Sciences

Biographical note for Florian Wolf

Education

May 1995 – July 2000:	Ludwig-Maximilians-University, Munich, Germany: Studies for M.A. (Linguistics / Computational Linguistics / Psycholinguistics) Final Grade: 1.05 (grades ranging from 1 to 6; 1 = best) M.A. thesis: "Psycholinguistic processes under dual-task conditions" (advisor: Prof. Dr. Janney; second advisor: Dr. Ziegler), grade for thesis: 1.00
July 1997 – July 1998:	University of Stellenbosch, South Africa: Non-degree 'Special International Student' in General Linguistics
September 2000 - :	Massachusetts Institute of Technology, Cambridge, U.S.A.: PhD student at the Department of Brain and Cognitive Sciences Advisor: Prof. Edward Gibson; GPA: 4.9/5.0

Awards

March, 2003: *Angus MacDonald Award for Excellence in Undergraduate Teaching.* Department of Brain and Cognitive Sciences, MIT, Cambridge, USA.

2003-2004: *Eli and Dorothy Berman Fund Fellowship.* MIT, Cambridge, USA.

Publications

Wolf, F & Gibson, E (2003). Parsing: An overview. In: Nadel, L (Ed.), *Encyclopedia of Cognitive Science*, pp.465-476. London: Macmillan Publishers.

Jones, DA, Wolf, F, Gibson, E, Williams, E, Fedorenko, E, Reynolds, D & Zissman, M (2003). Measuring the readability of automatic speech-to-text transcripts. In: *Proceedings of the 8th European Conference on Speech Communication and Technology.* Geneva, Switzerland, September 2003.

Wolf, F, Gibson, E, Fisher, A & Knight, M (to appear). *The Discourse GraphBank – a database of texts annotated with coherence relations.* Linguistic Data Consortium, University of Pennsylvania, PA, USA.

Wolf, F, Gibson, E & Desmet, T (in press). Coherence and pronoun resolution. *Language and Cognitive Processes.*

Wolf, F & Gibson, E (2004). Paragraph-, word-, and coherence-based approaches to sentence ranking: A comparison of algorithm and human performance. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics.* Barcelona, Spain, July 2004.

Wolf, F & Gibson, E (2004). Representing discourse coherence: A corpus-based analysis. In: *Proceedings of the 20th International Conference on Computational Linguistics.* Geneva, Switzerland, August 2004.

Acknowledgements

I would like to thank the people in my lab, Mara Breen, Evan Chen, Evelina Fedorenko, Dan Grodner, John Hale, Doug Rohde, Tessa Warren, and Duane Watson, for great discussions and for asking really good questions. I would also like to thank Dan, Doug, and Duane for sometimes losing in foosball.

My collaborators, Timothy Desmet and Emmanuel Bellengier, suffered with me through writing items for reading experiments. Timothy was in our lab for about one year, and no screw-up in experiments or hypotheses could escape him. Thanks to his great sense of humor, he also started the tradition of "good riddance"-cakes in our lab.

David Caplan was the first person in the Boston area I talked to before I came to MIT. Despite his extremely busy schedule, he has somehow always managed to provide excellent and detailed feedback.

Several undergraduate students in our lab contributed to the work reported in this thesis. In particular, Amy Fisher and Meredith Knight annotated text after text, and are now busy developing a version of Trivial Pursuit based on what they have learned from 1986-89 news texts.

Nancy Kanwisher made the students in my year feel welcome when we arrived at MIT. Denise Heintze has provided great guidance through all kinds of administrative jungles. I have always enjoyed talking with Pawan Sinha, partly because of his great sense of humor, partly because of his amazing creativity.

The MIT community in general and the audiences at the CogLunch seminars in particular have provided constructive criticism in the true sense of these words.

Doug Jones, Doug Reynolds, and Marc Zissman in the Communication Systems group at Lincoln Labs have provided me with the chance to explore other areas of natural language processing, besides discourse coherence.

Every time after talking with Regina Barzilay, I walked away with a "great bag full of new ideas".

The same is true for my committee members, Josh Tenenbaum, Michael Collins, and Andy Kehler. It would be nice if my days had significantly more than 24 hours, so that I could explore all their great suggestions and ideas!

My advisor, Ted Gibson, has taught me that it is OK to say "I don't know" (of course, if you don't know, then you go and try to find out!). He has also taught me the difference between "doing research" and "pushing an agenda". More people should have the opportunity to work with and learn from people like Ted.

I will miss our friends in the Boston area, in particular Simon, Nathan, Emily, David, Sejal, Phil, Cristina, Marie, David, Thomas, Sylvie, Mikael, and Yo.

My wife, Jennifer, has kept me happy and sane. I love her for that, for everything we do together, and for all the other things.

Jennifer's parents and her brother are the most generous people one could imagine. I am extraordinarily happy to be able to call them "family".

The energy of my brother, Markus, is an inspiration.

My parents have always encouraged me and their patience, trust, and emphasis on curiosity have made me what I am. Everybody should have parents like that.

Table of contents

Figures

Tables

1 Introduction

The general topic of this thesis is discourse coherence in natural language. Discourse coherence here is assumed to be a representation of informational relations between sentences or other segments in a text. The following example illustrates coherence:

(1) There was bad weather at the airport. Therefore our flight got delayed.

(2) The first flight to Frankfurt this morning was delayed. The second one was on time.

The sentences in (1) are in a causal coherence relation (bad weather was the reason for a flight delay), whereas the sentences in (2) are in a contrastive coherence relation (a flight that was delayed vs. a flight that was on time).

Discourse coherence is a crucial part of natural language production and comprehension. Therefore, a more detailed understanding of how discourse coherence can be represented and processed would be a contribution to research on natural language in general. Additionally, it has been argued that discourse coherence can influence psycholinguistic processes (e.g. Kehler (2002)), and that it can influence the relative saliency of segments of a text or discourse (e.g. Marcu (2000)). This thesis aims to address some of these questions, and the specific goals are:

- to develop a descriptively adequate representation for discourse structure
- to test the influence of discourse structure on psycholinguistic processes, in particular pronoun processing
- to test the influence of coherence structures on the relative saliency of discourse segments

The first aim of this thesis is to specify a descriptively adequate data structure for representing discourse coherence[1]. More specifically, we will discuss types of discourse relations as well as whether tree graphs are descriptively adequate for representing discourse structures. Results of this work were presented in Wolf & Gibson (2004b). In addition to specifying a descriptively adequate data structure for representing discourse coherence, a procedure will be developed for annotating naturally occurring discourses with coherence structures, based on the representation and data structure proposed in this thesis. This procedure is also described in Wolf et al. (2003).

The second aim of this thesis is to test how coherence structure can influence psycholinguistic processes. We will discuss a proposal by Hobbs (1979) and Kehler (2002) about the relation of coherence structure and pronoun processing, and present on-line experimental data that support their proposal. This work is also presented in Wolf et al. (in press).

The third aim of this thesis is to test to what extent the relative salience of segments in a discourse can be predicted from the structure of the graph representing the coherence structure of the discourse. Our computational method will be compared to human judgments of discourse segment salience. We will also compare the performance of our model to other models of discourse segment salience. Some results on discourse segment salience were presented in Wolf & Gibson (2004a).

1.1 Representations and data structures for discourse coherence

As pointed out above, the first goal of this thesis is to specify a descriptively adequate data structure for representing discourse coherence. In particular, it will be evaluated whether trees are a descriptively adequate data structure for representing coherence, or whether more powerful data structures are needed. Most accounts of discourse coherence assume trees as a data structure (e.g. Carlson et al. (2002); Corston-Oliver (1998); Grosz & Sidner (1986); Mann & Thompson (1988); Marcu (2000); Polanyi (1996)). Some accounts allow nodes with multiple parents but no crossed dependencies (e.g. Lascarides & Asher (1993); Webber et al. (1999b)). Other accounts assume data structures that

[1] The terms “coherence structure” and “discourse structure” will be used interchangeably.

allow both nodes with multiple parents and crossed dependencies (e.g. Danlos (2004); Hobbs (1985); McKeown (1985)).

So far the issue of descriptively adequate data structures for representing coherence has not been empirically evaluated. In order to conduct such an empirical evaluation, we tested whether discourse structures of naturally occurring texts contain violations of tree structure constraints, i.e. nodes with multiple parents or crossed dependencies. Our results indicate that this is the case.

Another empirical question is whether crossed dependencies and nodes with multiple parents occur frequently. If not, one could still assume a tree-based representation and accept a certain (low) error rate. However, if they occur frequently, just ignoring these violations is not an option.

A further empirical question is whether there are any constraints on where crossed dependencies and nodes with multiple parents can occur. If yes, one could assume an augmented tree structure in order to accommodate certain well-specified violations of tree constraints. If not, however, more powerful data structures than trees are needed.

In order to address the questions of frequency and constraints, we collected a database of 135 texts from the Wall Street Journal and the AP Newswire that were hand-annotated for coherence structures by two independent annotators. Our results indicate that crossed dependencies and nodes with multiple parents are too frequent to be ignored, and that there are no obvious constraints that would allow maintaining an augmented tree structure.

Chapter 2 describes the annotation procedure we used in order to collect the database of hand-annotated texts. That chapter will also argue against trees and for chain graphs as a data structure for representing discourse coherence.

1.2 Discourse coherence and psycholinguistic processing

The second goal of this thesis is to test whether processes underlying the establishment of coherence structure can influence psycholinguistic processes. As an example of such a psycholinguistic process, we examined pronoun processing.

Most accounts of pronoun processing are based on structural principles that are specific to referential processing (e.g. Chambers & Smyth (1998); Grosz et al. (1995), cf.

Wundt (1911)). Some accounts, however, treat pronoun processing as a byproduct of more general inference processes that are used when establishing coherence in a discourse (e.g. Hobbs (1979); Kehler (2002)).

We tested these accounts in an on-line comprehension study and in an off-line production study. The results from both studies support accounts that are based on general inference processes, rather than structural pronoun-specific accounts. These results, particularly from the on-line study, also support hypotheses about cognitive processes argued to be underlying the processing of coherence (Kehler (2002)). Chapter 3 will discuss these issues in more detail.

1.3 Discourse coherence and the importance of document segments

The third issue addressed in this thesis is the relative saliency of discourse segments in a text[2]; usually texts have some parts that are more and some that are less salient or important with respect to what the text is about. We will test whether coherence structures can be used to predict the relative saliency of discourse segments. This is a task that has applications in information extraction and summarization.

Besides coherence-based approaches, we also tested word- and layout-based approaches to determining discourse segment saliency. The basic idea behind word-based approaches is that a discourse segment is the more important the more important words it contains[3]. Different approaches have different definitions of what makes a word important (e.g. Luhn (1958); Salton & Buckley (1988)).

Layout-based approaches are based on the idea that layout information, such as paragraph structure in written discourse, makes discourse segments more or less important. A simple layout-based method was tested that ranks the first sentences of the first four paragraphs in a document as important.

[2] "Relative discourse segment saliency" here refers to the relative importance of a discourse segment. "Relative importance" refers to how important the content conveyed by a discourse segment is with respect to the content conveyed by the whole discourse. In what follows, we will use the terms "relative saliency", "relative importance" and "importance" interchangeably.

[3] Notice that we use "discourse segment importance" and "discourse segment salience" interchangeably.

Coherence-based approaches assume that the saliency of a discourse segment is based on its position in the informational structure of a discourse. What kind of position in a discourse structure makes a discourse segment more or less salient depends on the data structure that is assumed for representing discourse coherence. Tree-based approaches assume that a discourse segment is more important if it is in a higher position in the tree (e.g. Marcu (2000)). Chain graph-based approaches are based on the idea that a discourse segment is more important if more other discourse segments relate to it (e.g. Sparck-Jones (1993)).

The experimental results reported in Chapter 4 indicate that coherence-based methods that operate on chain graphs performed best. These results also indicate that the choice of data structure for representing discourse coherence is important because tree-based methods performed worse than chain graph-based methods. Chapter 4 will also discuss possible reasons why tree-based representations might systematically lead to errors in estimating discourse segment salience.

2 Representing discourse coherence: A corpus-based analysis

2.1 Introduction

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

There have been two kinds of approaches to defining and representing discourse structure and coherence relations. These approaches differ with respect to what kinds of discourse structure they are intended to represent. Some accounts aim to represent the intentional-level structure of a discourse; in these accounts, coherence relations reflect how the role played by one discourse segment with respect to the interlocutors' intentions relates to the role played by another segment (e.g. Grosz & Sidner (1986)). Other accounts aim to represent the informational structure of a discourse; in these accounts, coherence relations reflect how the meaning conveyed by one discourse segment relates to the meaning conveyed by another discourse segment (e.g. Hobbs (1985); Marcu (2000); Webber et al. (1999b)). Furthermore, accounts of discourse structure vary greatly with respect to how many discourse relations they assume, ranging from two (Grosz & Sidner (1986)) to over 400 different coherence relations, reported in Hovy & Maier (1995). However, Hovy & Maier (1995) argue that, at least for informational-level accounts, taxonomies with more relations represent subtypes of taxonomies with fewer relations. This means that different informational-level-based taxonomies can be compatible with each other; they differ with respect to how detailed or fine-grained they represent informational structures of texts. Going beyond the question of how different informational-level accounts can be compatible with each other, Moser & Moore (1996) discuss the compatibility of rhetorical structure theory (RST; Mann & Thompson (1988)) to the theory of Grosz & Sidner (1986). However, notice that Moser & Moore (1996) focus on the question of how compatible the claims are that Mann & Thompson (1988) and Grosz & Sidner (1986) make about intentional-level discourse structure.

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

This chapter is organized as follows. Section 2.2 reviews current approaches to discourse structure. Section 2.3 describes the procedure we used to collect a database of 135 texts annotated with coherence relations. Section 2.4 describes in detail the descriptional inadequacy of tree structures for representing discourse coherence, and Section 2.5 provides statistical evidence from our database that supports these claims. Section 2.6 contains some concluding remarks.

2.2 Current approaches to representing discourse structure

This section will review current approaches to discourse structure. For each approach, we will discuss its central features and point out if or how it is relevant to the approach that we propose. Some of the issues brought up in the next session, particularly the question of whether trees are descriptively adequate for representing discourse structures, will be discussed at length in subsequent sections.

2.2.1 Discourse Lexicalized Tree Adjoining Grammar (D-LTAG)

Webber et al. (1999a), Webber et al. (1999b), and Webber et al. (2003) aim to develop a representation of discourse structure within the formalism of tree-adjoining grammars (e.g. Joshi & Schabes (1997)). They point out that often discourse segments relate to more than one other discourse segments. An example of such a discourse structure is the structure of (3), where discourse segment 0 relates to both segments 1 and 2 (to segment 1 via some coherence relation R_2 and to segment 2 via some coherence relation R_1). Figure 1 shows the discourse structure for (3)[4].

(3) Example text (constructed)

0. There is a train on Platform A.
1. There is another train on Platform B.
2. The destination of the train on Platform A is Berlin.

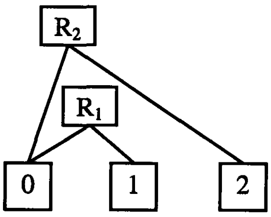


Figure 1. Discourse structure for (3).

Structures such as the one shown in Figure 1 pose a problem for tree structures because there would be nodes with multiple parents; in the discourse structure of (3), discourse segment 0 has two parents (R_1 and R_2). In order to still keep the tree structure, Webber et al. (1999a) introduce a secondary structure that augments the tree structure. Webber et al. (1999a)'s account distinguishes between *structural* coherence relations and *non-structural* or *anaphoric* coherence relations. The mechanics that are argued to be underlying these different kinds of coherence relations are termed *semantic composition* on the one and *anaphoric links* and *general inference* on the other hand (Webber et al. (2003)). Structural relations are represented by the tree structure itself, whereas anaphoric relations are represented in a secondary structure that is not subject to tree structure constraints. However, Webber et al. (2003) do not explain why they think that allowing nodes with multiple parents is acceptable but crossed dependencies would be "too costly" (p. 547) and less costly than developing an account that distinguishes

[4] Notice that a structure like "R_2 (R_1 (0, 1), 2)" would imply that discourse segment 2 relates to both discourse segments 0 and 1. However, in the example above, discourse segment 2 relates to discourse segment 0 but not to discourse segment 1.

structural and anaphoric relations. They do not make it clear what they mean with "cost". For example, it might be that the search space for possible discourse structures increases if one moved from trees to more general graphs, but there might be other constraints on graphs (i.e. not tree constraints) that could limit the search space for general graphs in different ways than a tree constraint. On the other hand, it might be costly in account like Webber et al. (2003)'s to develop a unification mechanism for structural and anaphoric levels of representation (cf. a similar argument for sentence structures by Skut et al. (1997)). These issues should be investigated in future research. Subsequent sections will come back to Webber et al. (2003)'s account in more detail, in particular to the issue of distinguishing structural from non-structural coherence relations.

2.2.2 Segmented Discourse Representation Theory (SDRT)

The goal of Lascarides & Asher (1993) is to provide an account of how coherence relations between discourse segments can be determined in a defeasible reasoning process, based on a theory of semantics such as Dowty (1986). Coherence relations in their account refer to informational relations that hold between what is stated by discourse segments, as in Hobbs (1985), not to communicative goals and intentions, as in Grosz & Sidner (1986). Defeasible or non-monotonic reasoning means that rules are subject to exceptions (e.g. Russell & Norvig (1995)).

Lascarides & Asher (1993) argue that the mechanics they describe for determining coherence relations can be seen as manifestations of Grice (1975)'s conversational maxims. Some of the mechanisms of determining coherence relations described by Lascarides & Asher (1993) were also used, with some modifications, to collect the database of coherence-labeled texts described here (for details of the annotating process, see Wolf et al. (2003)). The following are some mechanisms and principles for determining coherence relations from Lascarides & Asher (1993):

- *Penguin Principle*[5]: prefer a more specific coherence relation over a less specific one. For example, if two events are in a temporal sequence and there is also a causal relation between the events, prefer *cause-effect* over *temporal sequence.*
- *Narration Principle*: by default, events should be described in the order in which they happened. This principle corresponds to Grice's Maxim of Manner.
- *Causal Law*: for a causal coherence relation to hold between the events described by two discourse segments, the event that is the cause has to precede completely the event that is the effect.

The *Narration Principle* is an empirical claim that remains to be tested; the *Penguin Principle* and the *Causal Law* were used for determining coherence relations when we collected our database of hand-annotated texts. However, in order to be practically useful, *Narration Principle* and *Causal Law* should be defined less strictly. The following example from Bateman & Rondhuis (1994) illustrates this:

(4) The film was boring. John yawned.

Probably most people would agree that there is a causal relation between the two sentences in (4) – John yawned because he was bored by the film. However, the film event does not conclude before the yawning effect. The Causal Law as stated above demands that, and since this condition is not met in (4), no causal relation between the two sentences can be established. The same issue, one event not concluding before the other one starts, keeps a *temporal sequence* relation from applying. Following Lascarides & Asher (1993), the sentences in (4) would be in a *background / elaboration* relation. Bateman & Rondhuis (1994) show that in analyses of naturally occurring discourses,

[5] The *Penguin Principle* is called *Penguin Principle* because of the following example: assume one has a knowledge base that says that (1) birds can fly, (2) penguins are birds, (3) penguins cannot fly. In order to solve the contradiction here, one needs to decide between (1) and (3). (3) is selected because penguins are a more specific class than birds. By analogy to that example, the *Penguin Principle* with respect to discourse coherence says that a more specific relation is preferred over a less specific relation if the conditions for both the less and the more specific relation are met.

almost all coherence relations end up being *background / elaboration*, since Lascarides & Asher (1993)'s mechanisms for determining coherence relations are formulated so strictly.

Lascarides & Asher (1993) assume that only the right frontier of a discourse structure tree is open for attachment of new discourse segments. This means that they do not allow crossed dependencies. However, like Webber et al. (1999a), they appear to allow nodes with multiple parents (Lascarides & Asher (1991)). Neither crossed dependencies nor nodes with multiple parents are allowed by Polanyi (1996) and Polanyi et al. (2004). However, none provide any empirical evidence for tree structures or for constraining possible attachment sites to the right frontier of a discourse tree. For further discussion and empirical evaluation of these issues, see subsequent sections.

2.2.3 Linguistic Discourse Model

Polanyi (1996)'s Linguistic Discourse Model is based on the hypothesis that the sequence of clauses in a discourse reflects the sequence of events in the semantic model of the discourse (cf. Dowty (1986), Kamp & Rohrer (1983), Labov (1972), and Longacre (1983)). In addition, Polanyi (1996) hypothesizes that processing mechanisms as well as data structures for representing discourse structures should be the same as sentence-level processing mechanisms and data structures. According to Polanyi (1996), discourse segmentation should be clause-based, and discourse structures should be represented by tree structures, with the right frontier of a discourse tree being the only permissible attachment site for new discourse segments. Furthermore, Polanyi (1996) argues that the same ambiguity resolution mechanisms as for sentence-level structures should apply to discourse-level structures. However, she provides no details as to what these mechanisms might be or how exactly sentence-level mechanisms and data structures should be implemented on a discourse level. It is therefore hard to see how Polanyi (1996)'s account could be implemented or tested. Notice here also that Polanyi et al. (2004) has neither been implemented nor tested. Important for now is the point that Polanyi (1996) assumes trees for representing discourse structures, an assumption that will be empirically evaluated in subsequent sections.

2.2.4 Discourse semantics

The goal of Martin (1992)'s discourse semantics is to describe the relationship between coherence relations and how they are explicitly realized in linguistic structures (i.e. the choice of words, sentence, and paragraph structures). Martin (1992) distinguishes between external and internal coherence relations.

External relations refer to informational relations that hold between events in the world. Martin (1992)'s taxonomy of coherence relations is more fine-grained than e.g. Hobbs (1985)'s. For example, Martin (1992) distinguishes different kinds of causal relations, such as *cause:consequence*, as in (5), and *cause:contingency:condition: exclusive*, as in (6).

(5) *cause:consequence*
We arrived late, so we didn't have much time to prepare

(6) *cause:contingency:condition:exclusive*
The new software has to work. Otherwise everyone will be upset.

Martin (1992)'s internal relations are largely based on Halliday (1985), and refer to textual organization or forms of linguistic realization of external relations (i.e. the same external relation can be realized through different kinds of internal relations). For instance, there are different ways of dismissing objections:

(7) Dinner's ready. – But I'm not hungry.

(8) The weather might be bad, and there are no good movies showing right now. In any case, I'm going out.

We will be concerned with what Martin (1992) calls "external" relations, although we will use a less fine-grained taxonomy of relations than Martin (1992) (cf. Section 2.3.3). In order to represent coherence structures, Martin (1992) assumes tree structures. We will empirically evaluate that claim in subsequent sections.

2.2.5 Grosz & Sidner (1986)

Grosz & Sidner (1986)'s account is based on the belief that speaker intentions determine discourse structure. Their account is primarily intended as an account of multi-speaker dialogs, and it assumes three levels of structure: linguistic, intentional, and attentional structure.

Grosz & Sidner (1986) argue that linguistic structure is determined by cue words (*but*, *first*, *now*, *back to*, etc). Grosz & Sidner (1986) furthermore assume that the linguistic structure is isomorphic to the intentional structure of a discourse, and that therefore determining the intentional structure should help determine the linguistic structure. However, no further details are provided as to how this might be implemented, how exactly intentional structure might be determined, or how linguistic structure might be determined if there are no cue phrases that could determine a coherence relation (cf. Schauer (2000)). For representing linguistic structure, Grosz & Sidner (1986) assume tree structures.

Intentional structure refers to relations between the purposes of different discourse segments. The purpose of a discourse segment refers to the intention that a speaker has when he utters the discourse segment under consideration. Examples of discourse segment purposes are "intend that some agent believe some fact", or "intend that some agent know some property of some object". It is unclear, however, how intentional structure could be determined in practice, since it requires knowledge of speakers' intentions.

Attentional structure refers to a stack that contains the discourse elements on which participants focus their attention at a given point in the discourse. This stack also forms the basis of a pronoun resolution mechanism in Centering Theory (Joshi & Kuhn (1979), Grosz et al. (1995)); the entities on top of the attentional stack are assumed to be more easily accessible for pronominal reference than entities further down the stack (Grosz et al. (1995)). However, the idea of attentional structure as a stack has been challenged e.g. by Ariel (1990).

For our purposes here, Grosz & Sidner (1986)'s linguistic structure is the most relevant. Unfortunately, however, Grosz & Sidner (1986) do not provide a detailed enough description of that level of representation. Therefore the only claim we can evaluate empirically about Grosz & Sidner (1986)'s account is the question of whether tree structures are descriptively adequate (cf. subsequent sections). However, in Chapter 3, we will also evaluate attentional structure-based claims about pronoun processing that are made by Grosz et al. (1995) and Gordon et al. (1993), based on Grosz & Sidner (1986)'s theory.

2.2.6 Rhetorical Structure Theory (RST)

RST (Mann & Thompson (1988); Marcu (2000)) is intended to be a descriptive account of the organization of a discourse. The organization of a discourse is assumed to be a function of informational and intentional relations between discourse segments. The number of different relations varies greatly between different versions of RST, ranging from 25 (Mann & Thompson (1988)) to over 400[6] (Marcu (2000)), but commonly found examples of informational relations are *cause-effect*, where one discourse segment states the cause for the effect stated by another discourse segment; *elaboration*, where what is state by one discourse segment elaborates on what is stated by another segment; *contrast*, where what is stated by one discourse segment contrasts with what is stated by another discourse segment. In some versions of RST, *Cause-effect* has subcategories that can be differentiated by underlying intentions – causes can be *volitional* ("I dropped the glass because I wanted to see how is shatters on the ground") or *non-volitional* ("I dropped the glass because I fell down").

A central feature of RST is the intuition that some coherence relations are *symmetrical* and others are *asymmetrical.* Coherence relations are symmetrical if the discourse segments between which the coherence relation holds are equally important; this is true for *similarity* or *contrast* relations as in (9) or (10).

(9) Similarity

(9a) There is a train on Platform A.

(9b) There is another train on Platform B.

[6] The 25 coherence relations are a superset of the 400.

(10) Contrast

(10a) John supported Schwarzenegger during the campaign

(10b) but Susan opposed him.

In asymmetrical coherence relations, one participating segment is more important than the other one; this is true, for example, for *elaboration* relations, where the elaborating discourse segment is less important (*satellite*) than the elaborated segment (*nucleus*), cf. Mann & Thompson (1988). An example of an *elaboration* relation is (11), where (11b) is less important than (11a).

(11) Elaboration

(11a) A probe to Mars was launched from the Ukraine last week.

(11b) The European-built "Mars Express" is scheduled to reach Mars by late December.

Another central feature of RST is that it assumes the rhetorical or discourse structure of a discourse to be a tree. Marcu (2000) claims that trees are a descriptively adequate representation of most discourse structures and that they are easier to formalize and derive than less constrained graphs. We will evaluate this claim in subsequent sections.

2.2.7 Hobbs-relations

In Hobbs (1985)'s account, discourse coherence is closely related to more general inference mechanisms (cf. Hobbs et al. (1993)). Similar to Lascarides & Asher (1993), Mann & Thompson (1988), and Marcu (2000), Hobbs (1985) aims to account for informational relations that hold between what is stated in a discourse. Unlike these other accounts, however, Hobbs (1985) does not assume a tree structure for representing discourse structures. Instead, he assumes a labeled directed graph, where the nodes represent discourse segments, and the labeled directed arcs represent the coherence relations that hold between the discourse segments. In subsequent chapters we will test whether trees can be used to represent coherence structure, or whether more general graphs are needed.

Furthermore, Hobbs (1985) assumes a smaller number of different coherence relations than Mann & Thompson (1988) and Marcu (2000). Notice that Hovy & Maier (1995) showed that taxonomies with more relations represent subtypes of taxonomies with fewer relations (cf. Section 2.1). For the account of discourse structure that we will present in subsequent chapters, we assume a smaller set of coherence relations because they are easier to code and allow for a more abstract representation of discourse structure.

Hobbs (1985)'s set of coherence relations has also been adopted and adapted by Kehler (2002). Kehler (2002) shows that coherence relations influence linguistic processes such as pronoun resolution and ellipsis interpretation. The influence of coherence on psycholinguistic processes, in particular pronoun resolution, will be discussed in Chapter 3, where we test a hypothesis proposed by Kehler (2002) that is based on Hobbs (1979). Chapter 4 will test how coherence structure and text segment salience are related.

2.2.8 "Psychologically motivated" coherence relations

The taxonomies of informational coherence relations proposed by the accounts reviewed so far basically all go back to taxonomies that have been proposed by philosophers such as Hume (1748). Hume (1748) assumes that there are only three ways ideas can connect with each other (here, ideas would be equivalent to the information conveyed by discourse segments): Resemblance, Contiguity, and Cause-Effect. The ideas behind this ontology go back to Aristotle's associationism (e.g. Nestle (1977)).

However, instead of assuming these basic taxonomies from philosophy, Sanders et al. (1992) aimed to develop a taxonomy of coherence relations that mirrors basic cognitive mechanisms used for determining coherence. Sanders et al. (1992) argue that their proposed cognitive mechanisms are what underlies taxonomies like those proposed by Grosz & Sidner (1986), Hobbs (1985), or Polanyi & Scha (1984). In particular, Sanders et al. (1992) argue for the following "cognitive primitives":

- *Additive* vs. *causal*: some coherence relations establish a causal link between what is stated by two discourse segments, whereas all other coherence relations are additive.
- *Semantic* vs. *pragmatic*: coherence can be inferred from propositional content or from illocutionary force. However, it is not specified what exactly is meant by "illocutionary force".
- *Negative* vs. *positive*: negative coherence relations are *contrast* or *violated expectation*; all other coherence relations are positive.
- *Basic* vs. *non-basic order* (only applies to causal relations): in basic coherence relations, the cause is stated to the left side; in non-basic coherence relations, the cause is stated to the right side. This is claimed to relate to Grice (1975)'s Maxim of Manner (cf. also Lascarides & Asher (1993)).

However, there are problems with Sanders et al. (1992)'s account. The first problem is that it is unclear how what Sanders et al. (1992) propose is more than just regrouping different types of coherence relations from already existing taxonomies. For example, the "additive vs. causal" distinction just claims that causal coherence relations are somehow different from all other kinds of coherence relations. "Semantic vs. pragmatic" claims that informational coherence relations are somehow different from coherence relations that are not informational (although, as mentioned above, it is unclear what exactly the other, "non-informational" or "pragmatic", coherence relations describe). "Negative vs. positive" claims that *violated expectation* is related to *cause-effect*, and that *contrast* is related to *similarity* (cf. Section 2.3.3 for definitions of coherence relations). And "basic vs. non-basic order" is a claim specific to causal coherence relations. Sanders et al. (1992) do not justify why they decided to choose these "cognitive primitives" and not others. For example, it is unclear why there is no "symmetrical vs. asymmetrical" primitive (cf. Mann & Thompson (1988); Section 2.2.6). Another primitive might be "temporal sequence vs. no temporal sequence". Yet another primitive might be "structural processing involved vs. no structural processing involved", as proposed by Kehler (2002); consequences of such a distinction on pronoun processing are tested in Chapter 3. And it is unclear why Sanders et al. (1992)'s "basic vs. non-basic order" primitive is applied to causal relations, but not, for example, to *elaboration* relations (so that in a basic order, the elaborated statement comes before the elaborating one, cf. Section 2.3.3).

Another problem with Sanders et al. (1992) is that the items in their experiments were not controlled for plausibility. Nor do they report any statistics on their data. Thus it is not clear whether the differences they observed in their experiments were significant, and whether they were due to their independent variables or some other uncontrolled variable(s).

In the light of the issues pointed out above it is unclear why Sanders et al. (1992)'s cognitive primitives should have more merit to them than other taxonomies that are based on the ideas of Aristotle. However, the question of what psychological mechanisms underlie the process of inferring different kinds of coherence relations is an important one. Chapter 3 will come back to that question.

2.2.9 Summary on current accounts of discourse coherence

The previous sections have summarized and evaluated some current accounts of discourse coherence. The key points discussed were:

- **Claims about descriptively adequate data structures for representing discourse coherence:** Most of the reviewed accounts assume trees as a data structure for representing discourse structure. Some accounts relax tree assumptions and allow nodes with multiple parents (e.g. Webber et al. (1999b)). Other accounts assume more general graphs (e.g. Hobbs (1985)). This issue will be discussed in the subsequent sections.
- **Claims about the taxonomies of coherence relations:** In addition to informational coherence relations, some accounts also include intentional or attentional coherence relations (e.g. Grosz & Sidner (1986)). Most informational relation taxonomies are, though more or less fine grained, eventually based on taxonomies proposed by philosophers such as Nestle (1977). Sanders et al. (1992) claim that their account is based on psychological processes underlying

coherence relation taxonomies. However, Section 2.2.8 pointed out a number of problems with Sanders et al. (1992).

- **Claims about how coherence relations are inferred, and about how they influence other linguistic processes:** Kehler (2002), based on Hobbs (1979), makes claims about how different kinds of coherence relations are inferred, and about how these inference processes influence, for example, pronoun resolution. In a similar vein, Grosz & Sidner (1986) make predictions about pronoun processing, based on aspects of their account of discourse coherence. Some of these predictions will be tested in Chapter 3.

2.2.10 Alternative assumptions about data structures for representing discourse coherence

Here is a brief summary of assumptions made by different accounts of discourse coherence with respect to data structures:

- **Trees**
    - Binary branching
    - No augmenting or secondary structures (except Webber et al. (1999b))
    - Intermediate nodes: represent semantics that results from applying coherence relations to the semantics of discourse segments (cf. Marcu (2000)'s compositionality criterion)
    - Intermediate nodes can have other intermediate nodes as children
    - Coherence relations: represent kinds of inferences made over the semantic content of discourse segments or intermediate nodes
    - Can be represented by context-free grammars

- **More general graphs**
    - Directed and undirected, labeled, arcs (the labels represent types of coherence relations that hold between (groups of) discourse segments)
    - Discourse segments are represented by an ordered array of nodes; the order represents the sequence of discourse segments in a text

- No augmenting or secondary structures
- Intermediate nodes: represent semantics of topically related contiguous discourse segments
- Intermediate nodes can have other intermediate nodes as children
- Coherence relations: represent kinds of inferences made over the semantic content of discourse segments or intermediate nodes
- Cannot be represented by context-free grammars

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

2.3 Collecting a database of texts annotated with coherence relations

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

2.3.1 Discourse segments

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

For our database, we mostly adopted a clause unit-based definition of discourse segments (although restrictive relative clauses did not constitute their own discourse segments). We originally intended to use a sentence unit-based definition of discourse segments, following Hobbs (1985). But it turned out that newspaper texts often have very long sentences with several clauses that are in some kind of coherence relation. In order not to loose that information, we decided to use a more fine-grained, clause unit-based, definition of discourse segments. However, we also assume that contentful coordinating and subordinating conjunctions (cf. Table 1) can delimit discourse segments.

cause-effect	because; and so
violated expectation	although; but; while
condition	if...(then); as long as; while
similarity	and; (and) similarly
contrast	by contrast; but
temporal sequence	(and) then; first, second,...; before; after; while
attribution	according to...; ...said; claim that...; maintain that...; stated that...
example	for example; for instance
elaboration	also; furthermore; in addition; notice (furthermore) that; (for, in, on, against, with,...) which; who; (for, in, on, against, with,...) whom
generalization	in general

Table 1. Contentful conjunctions used to illustrate coherence relations.

Notice that we did not classify "and" as delimiting discourse segments if it was used to conjoin nouns in a conjoined noun phrase, like "dairy plants and dealers" in (12) (example from wsj_0306; Wall Street Journal 1989 corpus; Harman & Liberman (1993)), or if it was used to conjoin verbs in a conjoined verb phrase, like "snowed and rained" in (13) (constructed example):

(12) Milk sold to the nation's dairy plants and dealers averaged $14.50 for each hundred pounds.
(from wsj_0306; Wall Street Journal 1989 corpus; Harman & Liberman (1993))

(13) It snowed and rained all day long.

We classified full-stops, semicolons, and commas as delimiting discourse segments. However, in cases like (14) (constructed example), where they conjoin a complex noun phrase, commas were not classified as delimiting discourse segments.

(14) John bought bananas, apples, and strawberries.

We furthermore treat attributions ("John said that...") as discourse segments. This is empirically motivated. The texts used here are taken from news corpora, and there, attributions can be important carriers of coherence structures. For instance, consider a case where some Source A and some Source B both comment on some Event X. It should be possible to distinguish between a situation where Source A and Source B make basically the same statement about Event X, and a situation where Source A and Source B make contrasting comments about Event X. Notice, however, that we treated cases like (15) as one discourse segment and not as two separate ones ("...cited" and "transaction costs..."). We only separated attributions if the attributed material was a complementizer phrase, a sentence, or a group of sentences. This is not the case in (15) – the attributed material is a complex NP ("transaction costs from its 1988 recapitalization").

(15) The restaurant operator cited transaction costs from its 1988 recapitalization.
(from wsj_0667; Wall Street Journal 1989 corpus; Harman & Liberman (1993))

2.3.2 Discourse segment groupings

Adjacent discourse segments could be grouped together. For example, discourse segments were grouped if they all stated something that could be attributed to the same source (cf. Section 2.3.3 for a definition of *attribution* coherence relations). Furthermore, discourse segments were grouped if they were topically related. For example, if a text discusses inventions in information technology, there could be groups of a few discourse segments each talking about inventions by specific companies. There might also be subgroups, consisting of several discourse segments each talking about specific inventions at specific companies. Thus, marking groups can determine a partially hierarchical structure for the text.

Other examples of discourse segment groupings included cases where several discourse segments described an event or a group of events that all occurred before another event or another group of events described by another (group of) discourse segments. In that case, what is described by a group of discourse segments is in a *temporal sequence* relation with what is described by another (group of) discourse segments (cf. Section 2.3.3 for a definition of *temporal sequence* coherence relations). Notice furthermore that in cases where one topic requires one grouping and a following topic requires a grouping that is different from the first grouping, both groupings were annotated.

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

2.3.3 Coherence relations

As pointed out in Section 2.1, we aim to develop a representation of informational relations between discourse segments. Notice a difference between schema-based approaches (McKeown (1985)) and coherence relations like we use them: whereas schemas are instantiated from information contained in a knowledge base, coherence relations like we use them do not make (direct) reference to a knowledge base.

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

The kinds of coherence relations we used include *cause-effect* relations, as in the constructed example (16), where (16a) states the cause for the effect that is stated in (16b).

(16) Cause-Effect

(16a) There was bad weather at the airport

(16b) and so our flight got delayed.

Our *cause-effect* relation subsumes the *cause* as well as the *explanation* relation in Hobbs (1985). A *cause* relation holds if a discourse segment stating a cause occurs before a discourse segment stating an effect; an *explanation* relation holds if a discourse segment stating an effect occurs before a discourse segment stating a cause. We encoded this difference by adding a direction to the *cause-effect* relation. In a graph, this can be represented by a directed arc going from cause to effect.

Another kind of causal relation is *condition*. Hobbs (1985) does not distinguish *condition* relations from either *cause* or *explanation* relations. However, we felt that it might be important to distinguish between causal relation describing an actual causal event (*cause-effect*, cf. above) on the one hand, and a causal relation describing a possible causal event (*condition*, cf. below) on the other hand. In the constructed example (17), (17b) states an event that will only take place if the event described by (17a) also takes place.

(17) Condition

(17a) If the new software works,

(17b) everyone should be happy.

In a third type of causal relation, the *violated expectation* relation (also *violated expectation* in Hobbs (1985)), a causal relation between two discourse segments that normally would be present is absent. In (18) (constructed example), (18a) normally would be a cause for everyone being happy. This expectation is violated by what is stated by (18b).

(18) Violated Expectation

(18a) The new software works great,

(18b) but nobody was happy.

Other possible coherence relations include *similarity* (*parallel* in Hobbs (1985)) or *contrast* relations, where similarities or contrasts are determined between corresponding sets of entities or events (also *contrast* in Hobbs (1985)), such as between (19a) and (19b) and (20a) and (20b) respectively (examples (9) and (10), reprinted here as (19) and (20), respectively).

(19) Similarity

(19a) There is a train on Platform A.

(19b) There is another train on Platform B.

(20) Contrast

(20a) John supported Schwarzenegger during the campaign

(20b) but Susan opposed him.

Discourse segments might also *elaborate* (also *elaboration* in Hobbs (1985)) on other sentences, as in (21) (constructed example), where (21b) elaborates on (21a).

(21) Elaboration

(21a) A probe to Mars was launched from the Ukraine this week.

(21b) The European-built "Mars Express" is scheduled to reach Mars by late December.

Discourse segments can provide examples for what is stated by another discourse segment. In (22) (constructed example), (22b) states an *example* (*exemplification* in Hobbs (1985)) for what is stated in (22a).

(22) Example

(22a) There have been many previous missions to Mars.

(22b) A famous example is the Pathfinder mission.

Hobbs (1985) also includes an *evaluation* relation, as in (23) (example from Hobbs (1985)), where (23b) states an evaluation of what is stated in (23a). We decided to call such relations *elaborations*, since we found it too difficult in practice to reliably distinguish *elaborations* from *evaluations* (according to our annotation scheme, what is stated in (23b) elaborates on what is stated in (23a)).

(23) Elaboration – labeled as Evaluation in Hobbs (1985)

(23a) (A story.)

(23b) It was funny at the time.

(from Hobbs (1985))

Unlike Hobbs (1985), we also did not have a separate *background* relation as in (24) (example modified from Hobbs (1985)), where what is stated in (24a) provides the background for what is stated in (24b). Similarly to the *evaluation* relation, we found the *background* relation too difficult to reliably distinguish from *elaboration* relations (according to our annotation scheme, what is stated in (24a) elaborates on what is stated in (24b)).

(24) Elaboration – labeled as Background in Hobbs (1985)

(24a) T is the pointer to the root of a binary tree.

(24b) Initialize T.

(modified from Hobbs (1985))

In a *generalization* relation, as in (25) (constructed example), one discourse segment states a generalization, here (25b), for what is stated by another discourse segment, here (25a).

(25) Generalization

(25a) Two missions to Mars in 1999 failed.

(25b) There are many missions to Mars that have failed.

Furthermore, discourse segments can be in an *attribution* relation, as in (26) (constructed example), where (26a) states the source of what is stated in (26b) (cf. Bergler (1991) for a more detailed semantic analysis of *attribution* relations). Hobbs (1985) does not include an *attribution* relation. However, we decided to include *attribution* as a relation because, as pointed out in Section 2.3.1, the texts we annotated are taken from news corpora. There, attributions can be important carriers of coherence structures.

(26) Attribution

(26a) John said that

(26b) the weather would be nice tomorrow.

In a *temporal sequence* relation, as in (27) (constructed example), one discourse segment, here (27a), states an event that takes place before another event stated by the other discourse segment, here (27b). In contrast to *cause-effect* relations, there is no causal relation between the events described by the two discourse segments. The *temporal sequence* relation is equivalent to the *occasion* relation in Hobbs (1985).

(27) Temporal Sequence

(27a) First, John went grocery shopping.

(27b) Then he disappeared in a liquor store.

The *same* relation, illustrated by (28) (constructed example), is not an actual coherence relation but an epiphenomenon of assuming contiguous distinct elements of text (Hobbs (1985) does not include a *same* relation). A *same* relation holds if a subject NP is separated from its predicate by an intervening discourse segment. For example, in (28), (28a) is the subject NP of a predicate in (28c), and so there is a *same* relation between (28a) and (28c); (28a) is the first and (28c) is the second segment of what is actually one single discourse segment, separated by the intervening discourse segment (28b), which is in an *attribution* relation with (28a) (and therefore also (28c), since (28a) and (28c) are actually one single discourse segment).

(28) Same

(28a) The economy,

(28b) according to some analysts,

(28c) is expected to improve by early next year.

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

Hobbs (1985)	**Current annotation scheme**
occasion	temporal sequence
cause	cause-effect: cause stated first, then effect; directionality indicated by directed arcs in a coherence graph
explanation	cause-effect: effect stated first, then cause; directionality indicated by directed arcs in a coherence graph
–	condition
evaluation	elaboration
background	elaboration
exemplification: example stated first, then general case; directionality indicated by directed arcs in a coherence graph	example
exemplification: general case stated first, then example; directionality indicated by directed arcs in a coherence graph	generalization
elaboration	elaboration
parallel	similarity
contrast	contrast
violated expectation	violated expectation
–	attribution
–	same

Table 2. Correspondence between the set of coherence relations in Hobbs (1985) and our set of coherence relations.

We distinguish between asymmetrical or directed relations on the one hand and symmetrical or undirected relations on the other hand (Mann & Thompson (1988); Marcu (2000); cf. Section 2.2.6). *Cause-effect, condition, violated expectation, elaboration, example, generalization,* and *attribution* are asymmetrical or directed relations, whereas *similarity*, *contrast*, and *same* are symmetrical or undirected relations. In asymmetrical or directed relations, the directions of relations are as follows:

- *cause-effect*: from the discourse segment stating the cause to the discourse segment stating the effect
- *condition*: from the discourse segment stating the condition to the discourse segment stating the consequence
- *violated expectation*: from the discourse segment stating the cause to the discourse segment describing the absent effect
- *elaboration*: from the elaborating discourse segment to the elaborated discourse segment
- *example*: from the discourse segment stating the example to the discourse segment stating the exemplified
- *generalization*: from the discourse segment stating the special case to the discourse segment stating the general case
- *attribution*: from the discourse segment stating the source to the attributed discourse segment
- *temporal sequence*: from the discourse segment stating the event that happened first to the discourse segment stating the event that happened second

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

Notice also that in our annotation project we decided to annotate a coherence relation either if there was a coherence relation between the complete content of two discourse segments, or if there was a relation between parts of the content of two discourse segments. Consider the following example:

(29)

(29a) [Difficulties have arisen] [in enacting the accord for the independence of Namibia]

(29b) for which SWAPO has fought many years,

(from ap890104-0003; AP Newswire corpus; Harman & Liberman (1993))

For this example we would annotate an *elaboration* relation from (29b) to (29a) ((29b) provides additional details about the accord mentioned in (29a)), although the relation actually only holds between (29b) and the second part of (29a), indicated by angle brackets. While it is beyond the scope of the current project, future research should investigate annotations with discourse segmentations that allow annotating relations only between parts of discourse segments that "are responsible for" a coherence relation. For example, consider 0, where angle brackets indicate how more fine-grained discourse segments might be marked:

(30)

(30a) [for which] [SWAPO] [has fought many years,]

(30b) referring to the acronym of the South-West African Peoples Organization nationalist movement.

(from ap890104-0003; AP Newswire corpus; Harman & Liberman (1993))

In our current project, we annotated an *elaboration* relation from (30b) to (30a) ((30b) provides additional details, the full name, for SWAPO, which is mentioned in (30a)). A future, more detailed, annotation of coherence relations could then annotate this *elaboration* relation to hold only between (30b) and the word "SWAPO" in (30a).

2.3.4 Discourse coherence and general inferencing

Hobbs (1985) in particular points out that there is a relation between discourse coherence and general inference processes (cf. also Hobbs et al. (1993)). Notice that causal coherence relations require the most general inferencing (they require knowledge about what are possible causal relations in the world and what are not). The other coherence relations require less general inferencing. For example, in order to infer a *similarity* or a *contrast* relation, it is enough to be able to determine parallel linguistic structure between two discourse segments, and to make inferences about similar or contrasting entities or events. Such inferences might be based on resources such as WordNet (Fellbaum (2001)) or FrameNet (Johnson et al. (2003)), which encode the knowledge necessary to infer whether certain entities or events are similar or contrasting. Other coherence relations require even less general inferencing: for example, inferences for *elaboration*, *example*, or *generalization* relations do not require being able to determine parallel linguistic structures but only knowledge about super- or subclass relations between entities or events (cf. Fellbaum (2001) or Johnson et al. (2003)).

This hierarchy of how much inferencing is required by different kinds of coherence relations has implications on the discourse structure coding process we developed. When trying to determine coherence relations between (groups of) discourse segments, one should first try those kinds of coherence relations that place the strictest conditions on possible general inferences (i.e. causal coherence relations), followed by those that have less strict requirements on possible general inferences (i.e. *similarity*, *contrast*; *attribution*; *temporal sequence*; *elaboration*, *example*, *generalization*).

The following section will describe in detail the coding procedure we used for annotating texts. Section 2.3.6 will provide an example of how the coding procedure can be applied to a short text from the Wall Street Journal corpus (Harman & Liberman (1993)).

2.3.5 Coding procedure

In order to code the coherence relations of a text, we used a procedure consisting of three steps. In the first step, a text is segmented into discourse segments (cf. Section 2.3.1).

In the second step, adjacent discourse segments that are topically related are grouped together. The criteria for this step are described in Section 2.3.2.

In the third step, coherence relations (cf. Section 2.3.3) are determined between discourse segments and groups of discourse segments. Each previously unconnected (group of) discourse segment(s) is tested to see if it connects to any of the (groups of) discourse segments that have already been connected to the already existing representation of discourse structure.

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

The following list shows in more detail how the annotations were done. This list was also used by the annotators to guide them with their task.

A. Segment the text into discourse segments:

1. Insert segment boundaries at every full-stop that marks a sentence boundary (i.e. not at full-stops like in "Mrs." or in "Dr.")
2. Insert segment boundaries at every semicolon that marks a sentence or clause boundary
3. Insert segment boundaries at every double colon that marks a sentence or clause boundary
4. Insert segment boundaries at every comma that marks a sentence, clause, or modifying prepositional phrase (PP) boundary (modifying PPs are an important part of discourse structure in newspaper texts, which is why we decided to segment modifying PPs as their own discourse segments); do

not insert segment boundaries at commas that conjoin complex noun or verb phrases

5. Insert segment boundaries at every quotation mark, if there is not already a segment boundary based on Steps 1-4
6. Insert segment boundaries at the contentful coordinating conjunctions listed in Table 1, if there is not already a segment boundary based on Steps 1-5. For "and", do not insert a segment boundary if it is used to conjoin verbs or nouns in a conjoined verb or noun phrase

B. Generate groupings of related discourse segments:

1. Group contiguous discourse segments that are enclosed by pairs of quotation marks
2. Group contiguous discourse segments that are attributed to the same source
3. Group contiguous discourse segments that belong to the same sentence (marked by full-stops, commas, semicolons, or double colons)
4. Group contiguous discourse segments that are topically centered around the same entities or events

C. Determine coherence relations between discourse segments and groups of discourse segments. For each previously unconnected (group of) discourse segment(s), test if it connects to any of the (groups of) discourse segments that have already been connected to the already existing representation of discourse structure. Use the following steps for each decision:

1. Use pairs of quotation marks as a signal for *attribution*
2. For pairs of (groups of) discourse segments that are already connected with one of the contentful coordinating conjunction from Table 1: choose the coherence relation that corresponds to the coordinating conjunction
3. For pairs of (groups of) discourse segments that are not connected with one of the contentful coordinating conjunction from Table 1:

   a. Mentally connect the (groups of) discourse segments with one of the coordinating conjunctions from Table 1 and judge if the resultant passage sounds acceptable
   b. If the passage sounds acceptable, choose the coherence relation that corresponds to the coordinating conjunction
   c. If the passage does not sound acceptable, repeat Step 3a until an acceptable coordinating conjunction is found
   d. If the passage does not sound acceptable with any of the coordinating conjunctions from Table 1, assume that the (groups of) discourse segments under consideration are not related by a coherence relation

4. Iterative procedure for Steps 2 and 3:
   a. Start with any of the unambiguous coordinating conjunctions from Table 1 ("because", "although", "if...then", "...said", "for example")
   b. If none of the unambiguous coordinating conjunctions results in an acceptable passage, use the more ambiguous coordinating conjunctions ("and", "but", "while", "also", etc.)
5. Important distinctions for Steps 2 and 3 (this is based on practical issues that came up during the annotation project):
   a. *Example* vs. *elaboration*: An *example* relation sets up an additional entity or event (the example), whereas an *elaboration* relation provides more details about an already introduced entity or event (the one on which one elaborates)
   b. *Cause-effect* vs. *temporal sequence*: both *cause-effect* and *temporal sequence* describe a temporal order of events (in *cause-effect*, the cause has to precede the effect). However, only *cause-effect* relations have a causal relation between what is stated by the (groups of) discourse segments under consideration. Thus, if there is a causal relation between the (groups of) discourse segments

under consideration, assume *cause-effect* rather than *temporal sequence* (cf. Lascarides & Asher (1993))

2.3.6 Example of the coding procedure

In order to illustrate our approach to determining discourse structures, we use the following extract of a text from the Wall Street Journal (Harman & Liberman (1993), text wsj_0607):

(31)

> Three new issues began trading on the New York Stock Exchange today, and one began trading on the Nasdaq/National Market System last week. On the Big Board, Crawford & Co., Atlanta, (CFD) began trading today. Crawford evaluates health care plans, manages medical and disability aspects of worker's compensation injuries and is involved in claims adjustments for insurance companies.

(from wsj_0607; Wall Street Journal corpus; Harman & Liberman (1993))

The following sections will illustrate how the coding procedure outlined in Section 2.3.5 can be applied to (31).

2.3.6.1 Segmenting the text into discourse segments

The first step of annotating a text with a discourse structure is to segment the text into discourse segments (Steps A1 to A6 in the coding procedure in 2.3.5). Here is how this procedure is applied to (31):

Step A1: Insert segment boundaries at every full-stop that marks a sentence boundary.

(32) Example text after applying Step A1:

0. Three new issues began trading on the New York Stock Exchange today, and one began trading on the Nasdaq/National Market System last week.
1. On the Big Board, Crawford & Co., Atlanta, (CFD) began trading today.
2. Crawford evaluates health care plans, manages medical and disability aspects of worker's compensation injuries and is involved in claims adjustments for insurance companies.

Step A2: Insert segment boundaries at every semicolon that marks a sentence or clause boundary. There are no semicolons in the example text, so this step is skipped.

Step A3: Insert segment boundaries at every double colon that marks a sentence or clause boundary. There are no double colons in the example text, so this step is skipped.

Step A4: Insert segment boundaries at every comma that marks a sentence, clause, or modifying PP boundary. Notice that the comma in discourse segment 4 in (33) conjoins elliptic clauses ("Crawford evaluates health care plans, [Crawford] manages medical and disability aspects of worker's compensation injuries and [Crawford] is involved in claims adjustment...").

(33) Example text after applying Step A4:

0. Three new issues began trading on the New York Stock Exchange today,
1. and one began trading on the Nasdaq/National Market System last week.
2. On the Big Board,
3. Crawford & Co., Atlanta, (CFD) began trading today.
4. Crawford evaluates health care plans,
5. manages medical and disability aspects of worker's compensation injuries and is involved in claims adjustments for insurance companies.

Step A5: Insert segment boundaries at every quotation mark. There are no quotations in the example text, so this step can be skipped.

Step A6: Insert segment boundaries at contentful coordinating conjunctions. There is one contentful coordinating conjunction, “and”, in discourse segment 5 in (33); it conjoins two clauses (notice that “Crawford” is the elided subject for both discourse segments 5 and 6 in (34).

(34) Example text after applying Step A6:

0. Three new issues began trading on the New York Stock Exchange today,
1. and one began trading on the Nasdaq/National Market System last week.
2. On the Big Board,
3. Crawford & Co., Atlanta, (CFD) began trading today.
4. Crawford evaluates health care plans,
5. manages medical and disability aspects of worker's compensation injuries
6. and is involved in claims adjustments for insurance companies.

2.3.7 Groupings of discourse segments

In the second step of annotating a text with a discourse structure, related contiguous discourse segments are grouped together (Steps B1 to B4 in the coding procedure in 2.3.5). Here is how this procedure is applied to (34):

Step B1: Group contiguous discourse segments that are enclosed by pairs of quotation marks. There are no quotation marks in the example text, so this step can be skipped.

Step B2: Group contiguous discourse segments that are attributed to the same source. Because there are no sources in the example text, this step can be skipped.

Step B3: Group contiguous discourse segments that belong to the same sentence. Discourse segments 0 and 1 are in the same sentence, so there is a group including these two discourse segments. Furthermore, discourse segments 2 and 3 as well as 4, 5, and 6 respectively, are in the same sentence, so there are two more groups: one including discourse segments 2 and 3, and one including discourse segments 4, 5, and 6.

Step B4: Group contiguous discourse segments that are topically centered around the same entities or events. Discourse segments 2 through 6 are about a company, Crawford. Therefore these discourse segments are grouped together.

2.3.7.1 Coherence relations between discourse segments and groups of discourse segments

After grouping related contiguous discourse segments, the partial discourse structure for (34) looks like shown in Figure 2:

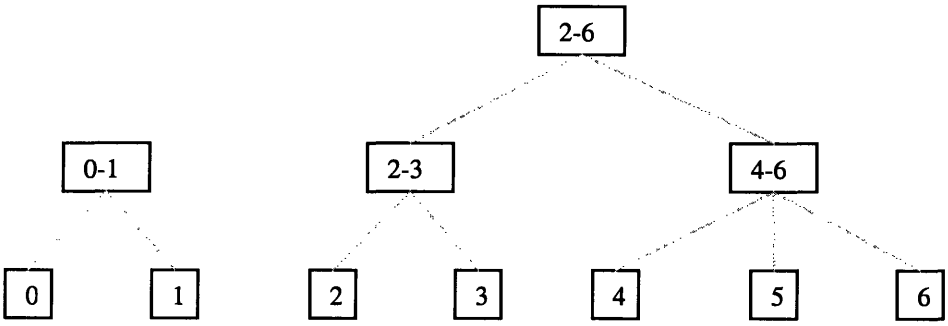


Figure 2. Partial discourse structure for (34) after discourse segment groupings.

The third step of annotating a text with a discourse structure consists of determining coherence relations between (groups of) discourse segments (Steps C1 to C5 in the coding procedure in 2.3.5). Here is how this procedure is applied to (34), reprinted here as (35):

(35) Example text:

0. Three new issues began trading on the New York Stock Exchange today,
1. and one began trading on the Nasdaq/National Market System last week.
2. On the Big Board,
3. Crawford & Co., Atlanta, (CFD) began trading today.
4. Crawford evaluates health care plans,
5. manages medical and disability aspects of worker's compensation injuries
6. and is involved in claims adjustments for insurance companies.

Step C1: Use pairs of quotation marks as a signal for *attribution.* This step is skipped because there are no quotation marks in the example text.

Step C2: Choose coherence relations for pairs of (groups of) discourse segments that are connected with a contentful coordinating conjunction. Discourse segments 0 and 1 and 5 and 6 respectively are connected with "and". Because of that, and because of their parallel sentence structures, these discourse segments are in *similarity* relations (discourse segments 0 and 1: both are about new issues that began trading on an exchange; discourse segments 5 and 6: both describe activities of the company Crawford). This results in the partial discourse structure shown in Figure 3:

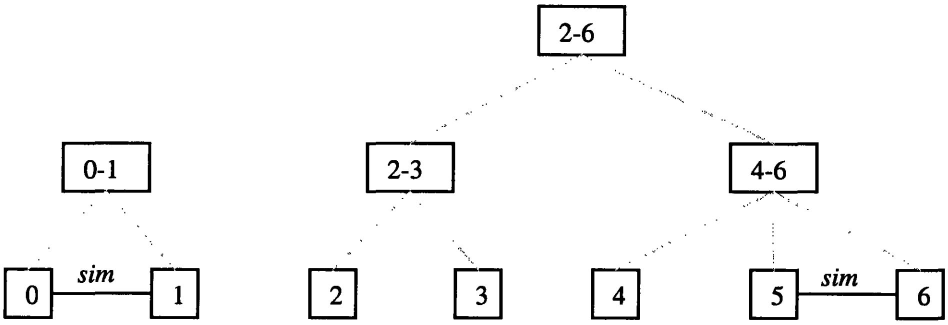


Figure 3. Partial discourse structure for (35) after applying Step C2.

Steps C3-C5: Determine coherence relations for the (groups of) discourse segments that are not conjoined with contentful coordinating conjunctions. Since discourse segments 0 and 1 have already been connected, building a discourse structure for (35) will continue with integrating discourse segment 2 into the discourse structure built so far.

- Discourse segment 2: "Big Board" refers to "New York Stock Exchange" in discourse segment 0, resulting in an *elaboration* relation between discourse segments 2 and 0. There are no other coherence relations between discourse segment 2 and the discourse structure built so far. The result is the partial discourse structure shown in Figure 4.

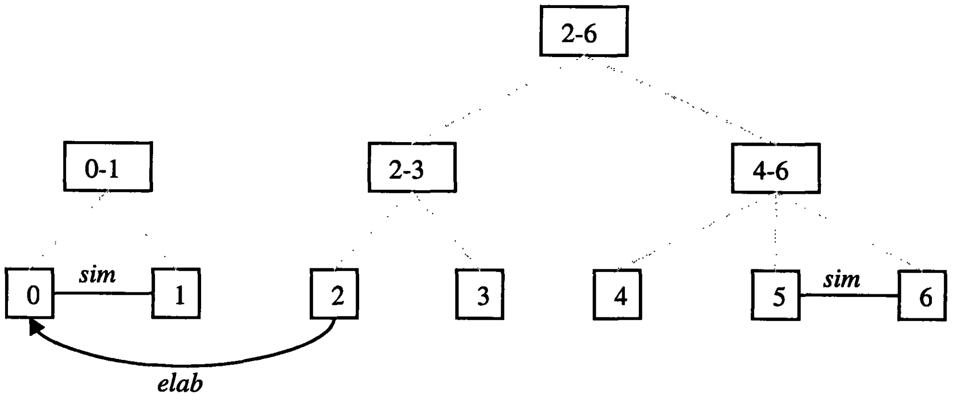


Figure 4. Partial discourse structure for (35) after integrating discourse segment 2.

- Discourse segment 3: discourse segment 2 is a prepositional phrase that modifies discourse segment 3 (discourse segment 2 provides additional detail about where the trading described in discourse segment 3 takes place). Therefore an elaboration relation holds between discourse segments 2 and 3. There are no other coherence relations between discourse segment 3 and the discourse structure built so far. The result is the partial discourse structure shown in Figure 5.

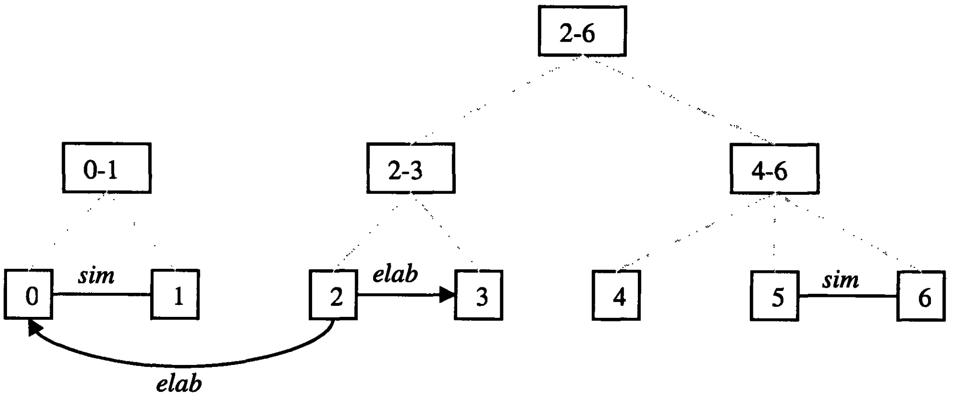


Figure 5. Partial discourse structure for (35) after integrating discourse segment 3.

- Group of discourse segments 2-3: this group of discourse segments provides additional detail about discourse segment 0 (2-3 provide details about a company that began being traded on the New York Stock Exchange, mentioned in discourse segment 0), so there is an *elaboration* relation between 2-3 and 0. There are no other coherence relations between 2-3 and the discourse structure built so far. The result is the partial discourse structure shown in Figure 6.

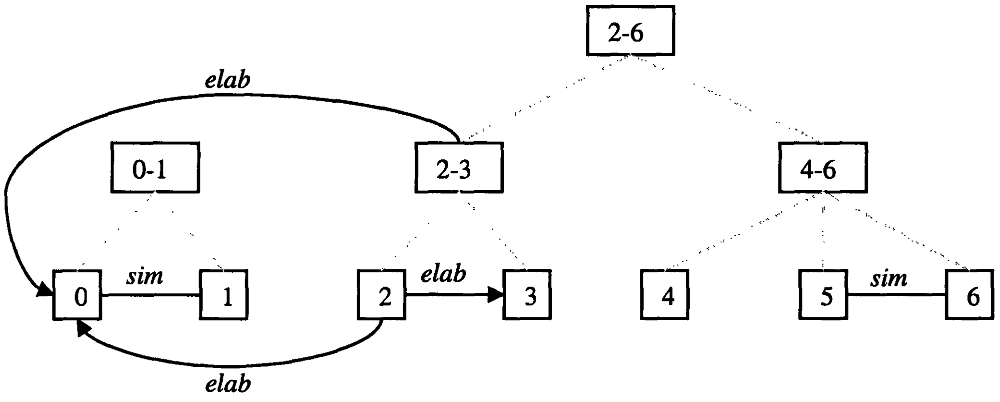


Figure 6. Partial discourse structure for (35) after integrating the group of discourse segments 2-3.

- Discourse segment 4: discourse segments 4 and 5 have parallel structure and both describe things that the company Crawford does. Therefore there is a *similarity* relation between these two discourse segments. There is furthermore a similarity relation between discourse segments 4 and 6. This results in the partial discourse structure shown in Figure 7.

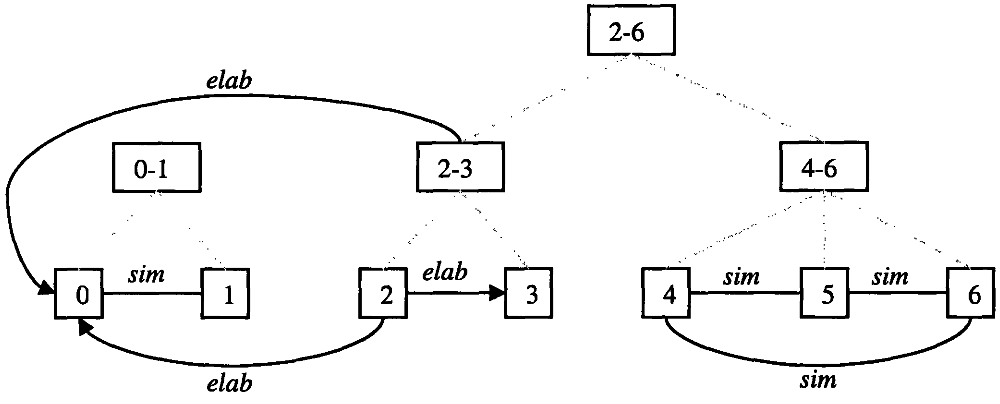


Figure 7. Partial discourse structure for (35) after integrating discourse segment 4.

- Group of discourse segments 4-6: instead of putting individual coherence relations between discourse segments 4, 5, and 6 and the already built structure, these coherence relations are put between the group of discourse segments 4-6 and the already built structure: for the example text considered here, all relations that hold for the individual discourse segments in the group 4-6 also hold for that group. There is an *elaboration* relation between 4-6 and discourse segment 3: 4-6 provide additional detail (business activities) about the company Crawford that is mentioned in discourse segment 3. The result is the partial discourse structure shown in Figure 8.

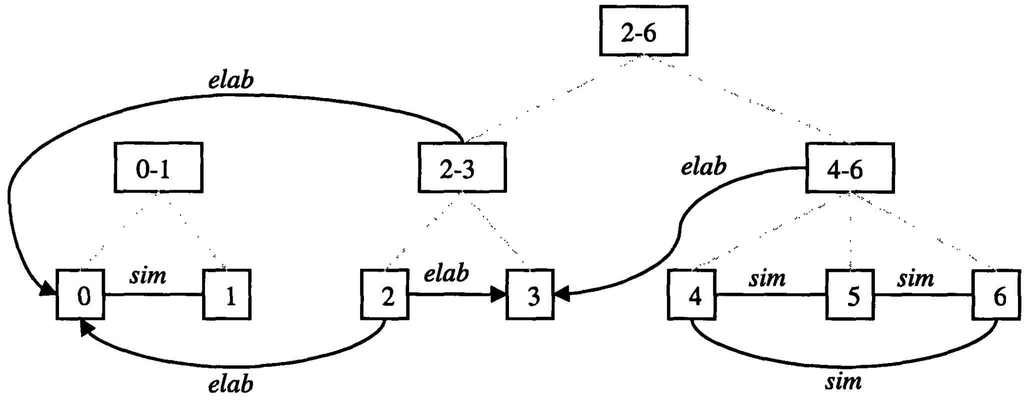


Figure 8. Partial discourse structure for (35) after integrating the group of discourse segments 4-6.

- Group of discourse segments 2-6: this group of discourse segments provides detail about one of the companies that began being traded on the New York Stock Exchange. Therefore there is an elaboration relation between 2-6 and discourse segment 0 (which mentions the New York Stock Exchange). The result is the discourse structure shown in Figure 9.

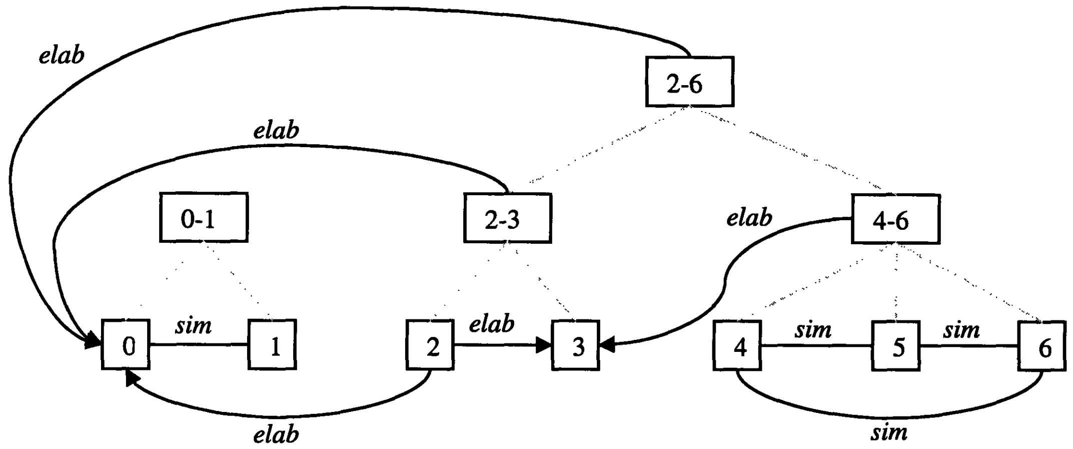
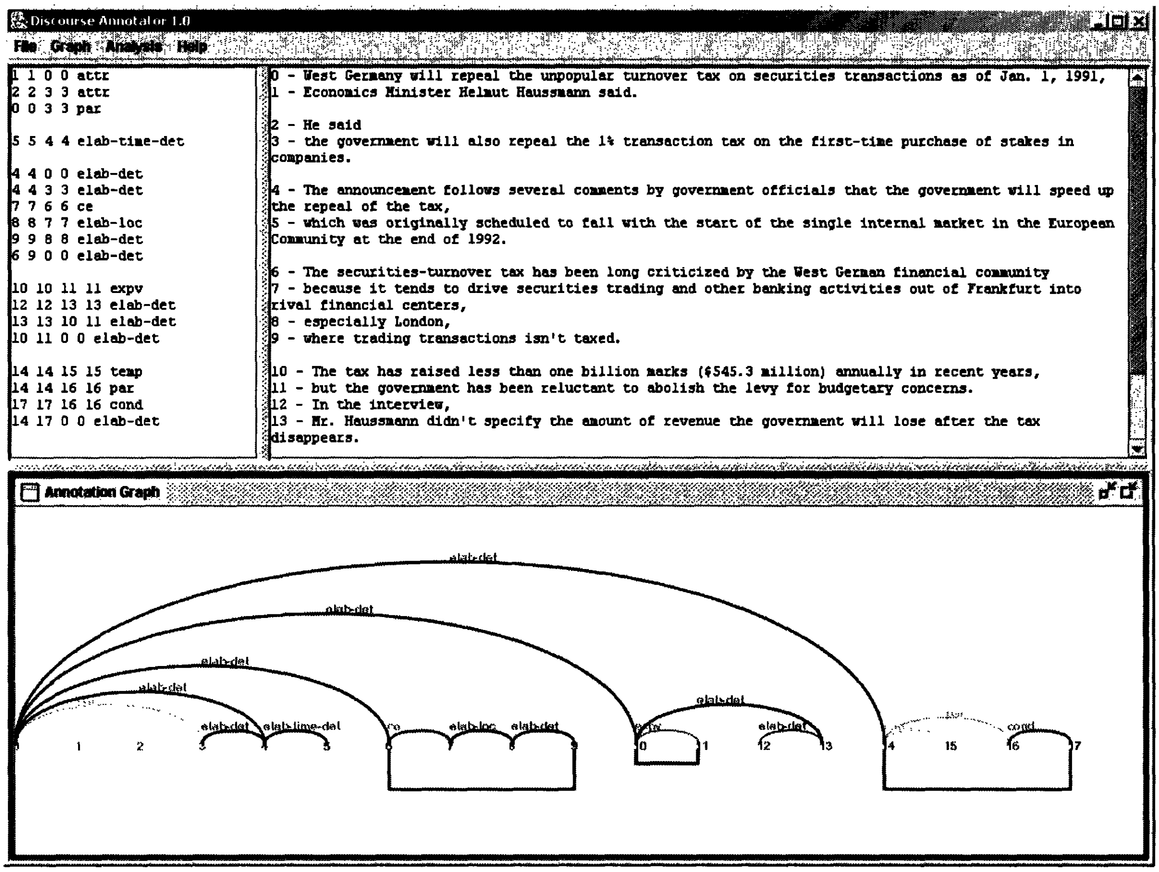


Figure 9. Discourse structure for (35) after integrating the group of discourse segments 2-6.

2.3.8 Annotators and annotation tool

The annotators for the database were MIT undergraduate students who worked in our lab as research students. For training, the annotators received a manual that describes the background of the project, discourse segmentation, coherence relations and how to recognize them, and how to use the annotation tools that we developed in our lab (Wolf et al. (2003). The author of this thesis provided training for the annotators. Training consisted of explaining the background of the project and the annotation method, and of annotating example texts (these texts are not included in our database). Training took about 8-10 hours in total, distributed over 5 days of a week. After that training, annotators worked independently.

In order to do the annotations, annotators used a simple Java-based tool that displayed the text to be annotated, the annotators' annotation commands, and a very simple display of the coherence graph annotated thus far. The tool used different colors for different kinds of coherence relations (green for *similarity* and *contrast*; blue for *example*, *generalization*, and *elaboration*; red for *cause-effect*, *violated expectation*, and *condition*; cyan for *temporal sequence*; orange for *attribution*; grey for *same*). Groupings of discourse segments were indicated with open boxes that grouped together discourse segments in the coherence graph display. Figure 10 shows a screenshot of the annotation tool.



```
1 1 0 0 attr
2 2 3 3 attr
0 0 3 3 par

5 5 4 4 elab-time-det

4 4 0 0 elab-det
4 4 3 3 elab-det
7 7 6 6 ce
8 8 7 7 elab-loc
9 9 8 8 elab-det
6 9 0 0 elab-det

10 10 11 11 expv
12 12 13 13 elab-det
13 13 10 11 elab-det
10 11 0 0 elab-det

14 14 15 15 temp
14 14 16 16 par
17 17 16 16 cond
14 17 0 0 elab-det
```

```
0 - West Germany will repeal the unpopular turnover tax on securities transactions as of Jan. 1, 1991,
1 - Economics Minister Helmut Haussmann said.

2 - He said
3 - the government will also repeal the 1% transaction tax on the first-time purchase of stakes in
companies.

4 - The announcement follows several comments by government officials that the government will speed up
the repeal of the tax,
5 - which was originally scheduled to fall with the start of the single internal market in the European
Community at the end of 1992.

6 - The securities-turnover tax has been long criticized by the West German financial community
7 - because it tends to drive securities trading and other banking activities out of Frankfurt into
rival financial centers,
8 - especially London,
9 - where trading transactions isn't taxed.

10 - The tax has raised less than one billion marks ($545.3 million) annually in recent years,
11 - but the government has been reluctant to abolish the levy for budgetary concerns.
12 - In the interview,
13 - Mr. Haussmann didn't specify the amount of revenue the government will lose after the tax
disappears.
```

Figure 10. Screenshot of the annotation tool. Upper right window pane: text to be annotated, with numbered discourse segments; upper left window pane: annotation code entered by the annotators; lower window pane: simple display of the coherence graph annotated thus far.

2.3.9 Statistics on annotated database

In order to evaluate hypotheses about appropriate data structures for representing coherence structures, we have collected a database of texts where the relations between discourse segments are labeled with the coherence relations described above. Table 3 shows statistics for a database of 135 texts from the Wall Street Journal 1987-1989 and the AP Newswire 1989 (both from Harman & Liberman (1993)) that have been annotated with coherence relations.

number of words		number of discourse segments	
mean	545	mean	61
minimum	161	minimum	6
maximum	1409	maximum	143
median	529	median	60

Table 3. Database statistics for 135 texts from AP Newswire 1989 (105 texts) and Wall Street Journal 1989 (30 texts).

Steps Two (discourse segment grouping) and Three (coherence relation annotation) of the coding procedure were performed independently by two annotators. For Step One (discourse segmentation), a pilot study on 10 texts showed that agreement on this step, *number of common segments / (number of common segments + number of differing segments)*, was never below 90%. Therefore, all 135 texts were segmented by two annotators together, resulting in segmentations that both annotators could agree on.

In order to determine inter-annotator agreement for Step Two of the coding procedure for the database of annotated texts, we calculated kappa statistics (Carletta (1996)). We used the following procedure to construct a confusion matrix: first, all groups marked by either annotator were extracted. Annotator 1 had marked 2616 groups, and Annotator 2 had marked 3021 groups in the whole database. The groups marked by the annotators consisted of 536 different discourse segment group types (for example, groups that included the first two discourse segments of each text were marked 31 times by both annotators; groups that included the first three discourse segments of each text were marked 6 times by both annotators). Therefore, the confusion matrix had 536 rows and columns. For all annotations of the 135 texts, the agreement was 0.8449, per chance agreement was 0.0161, and kappa was 0.8424. Annotator agreement did not differ as a function of text length, arc length, or kind of coherence relation (all χ^2s < 1).

We also calculated kappa statistics to determine inter-annotator agreement for Step Three of the coding procedure for the database of annotated texts[7]. For all

[7] Notice that inter-annotator agreement for Step Three was influenced by inter-annotator agreement for Step Two. For example, one annotator might mark a group of discourse segments 2 and 3, whereas the second annotator might not mark that group of discourse segments. If the first annotator then marks e.g. a *cause-*

annotations of the 135 texts, the agreement was 0.8761, per chance agreement was 0.2466, and kappa was 0.8355. Annotator agreement did not differ as a function of text length ($\chi^2 = 1.27$; $p < 0.75$), arc length ($\chi^2 < 1$), or kind of coherence relation ($\chi^2 < 1$). Table 4 shows the confusion matrix for the database of 135 annotated texts that was used to compute the kappa statistics. Table 4 shows, for example, that much of the inter-annotator disagreement seems to be driven by disagreement over how to annotate *elaboration* relations (in the whole database, Annotator$_1$ marked 260 *elaboration* relations where Annotator$_2$ marked no relation; Annotator$_2$ marked 467 *elaboration* relations where Annotator$_1$ marked no relation).

	Annotator$_2$													
Annotator$_1$	*contr*	*expv*	*ce*	*none*	*gen*	*cond*	*examp*	*temp*	*attr*	*elab*	*same*	*sim*	**sum**	**percent**
contr	383	11	0	34	0	0	0	2	0	0	0	0	**430**	**4.47**
expv	4	113	0	7	0	0	0	0	0	0	0	0	**124**	**1.29**
ce	0	0	446	14	0	0	0	0	0	5	0	0	**465**	**4.83**
none	66	24	42	0	0	2	27	16	6	467	1	64	**715**	**7.43**
gen	0	0	0	1	21	0	0	0	0	1	0	0	**23**	**0.24**
cond	0	0	0	2	0	127	0	1	0	1	0	0	**131**	**1.36**
examp	0	0	1	18	0	0	219	0	0	3	0	0	**241**	**2.51**
temp	1	1	2	7	0	0	0	214	0	1	0	0	**226**	**2.35**
attr	0	0	0	5	0	0	0	0	1387	0	0	0	**1392**	**14.47**
elab	0	0	17	260	0	3	0	3	0	3913	1	0	**4197**	**43.63**
same	0	0	2	5	0	0	0	1	0	0	530	1	**539**	**5.60**
sim	7	0	3	43	0	0	0	6	0	0	3	1074	**1136**	**11.81**
sum	**461**	**149**	**513**	**396**	**21**	**132**	**246**	**243**	**1393**	**4391**	**535**	**1139**		
percent	**4.79**	**1.55**	**5.30**	**4.12**	**0.20**	**1.37**	**2.56**	**2.53**	**14.50**	**45.60**	**5.56**	**11.80**		

Table 4. Confusion matrix of annotations for the database of 135 annotated texts (*contr = contrast*; *expv = violated expectation*; *ce = cause-effect*; *none = no coherence relation*; *gen = generalization*; *cond = condition*; *examp = example*; *ts = temporal sequence*; *attr = attribution*; *elab = elaboration*; *sim = similarity*).

effect coherence relation between discourse segment 4 and the group of discourse segments 2 and 3, whereas the second annotator marks a *cause-effect* coherence relation between discourse segment 4 and discourse segment 3, this would count as a disagreement. Thus, our measure of inter-annotator agreement for Step Three is conservative.

The only other comparable discourse annotation project that we are currently aware of is Carlson et al. (2002)[8]. Since they use trees and split up the annotation process into different sub-steps compared to our procedure, their annotator agreement figures are not directly comparable to ours. Furthermore, notice that Carlson et al. (2002) do not report annotator agreement figures for their database as a whole, but for different subsets of four to seven documents that were each annotated by different pairs of annotators. For discourse segmentation, Carlson et al. (2002) report kappa values ranging from 0.951 to 1.00; for annotation of discourse tree spans, their kappa values ranged from 0.778 to 0.929; for annotation of coherence relation nuclearity (whether a node in a discourse tree is a nucleus or a satellite, cf. Section 2.3.3 for the definition of nucleus vs. satellite), kappa values ranged from 0.695 to 0.882; for assigning types of coherence relations, Carlson et al. (2002) report kappa values ranging from 0.624 to 0.823.

2.4 Data structures for representing coherence relations

In order to represent the coherence relations between discourse segments in a text, most accounts of discourse coherence assume tree structures (Britton (1994); Carlson et al. (2002); Corston-Oliver (1998); Longacre (1983); Grosz & Sidner (1986); Mann & Thompson (1988); Marcu (2000); Polanyi & Scha (1984); Polanyi (1996); Polanyi et al. (2004); van Dijk & Kintsch (1983); Walker (1998)); some accounts do not allow crossed dependencies but appear to allow nodes with multiple parents (Lascarides & Asher (1991))[9]. Other accounts assume less constrained graphs that allow crossed dependencies as well as nodes with multiple parents (e.g. Bergler (1992); Birnbaum (1982); Danlos (2004); Hobbs (1985); McKeown (1985); Reichman (1985); Zukerman & McConachy (1995); for dialogue structure: Penstein Rose et al. (1995)).

[8] Notice that Miltsakaki et al. (2004) report results on annotating connectives but not on annotating whole discourse structures.

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

Some proponents of tree structures assume that trees are easier to formalize and to derive than less constrained graphs (Marcu (2000); Webber et al. (2003); but cf. Vogel et al. (1996)). We demonstrate that in fact many coherence structures in naturally occurring texts cannot be adequately represented by trees. Therefore we argue for less constrained graphs as an appropriate data structure for representing coherence, where nodes represent discourse segments and labeled directed arcs represent the coherence relations that hold between these discourse segments.

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

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

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

Some accounts of informational discourse structure do not assume tree structures, e.g. Bergler (1992) and Hobbs (1985) for monologue and Penstein Rose et al. (1995) for dialogue structure. However, none of these accounts provides systematic empirical support for using more general graphs rather than trees. Providing a systematic empirical study of whether trees are descriptively adequate for representing discourse coherence is the goal of this thesis.

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

Notice that there seems to be a problem with the way Webber et al. (2003) define the difference between "structural" and "non-structural" discourse structure. Webber et al. (2003)'s definition of "structural" vs. "non-structural" is based on the semantics and certain distributional patterns of "discourse adverbials" (e.g. *then*, *also*, *otherwise*, *instead*, etc). But it is unclear how their account generalizes to cases where there is a coherence relation between two discourse segments but no discourse adverbial. However, Schauer (2000) points out that in fact only about 15-20% of all coherence relations in a discourse are signaled by what Webber et al. (2003) would call "discourse adverbials".

Knott (1996) might provide a way to empirically formalize the claims in Webber et al. (2003), or at least claims that seem to be very similar to Webber et al. (2003): based on the observation that he cannot identify characteristic cue phrases for *elaboration* relations (e.g., "because" would be a characteristic cue phrase for *cause-effect*), Knott (1996) argues that *elaboration* relations are more permissive than other types of coherence relations (e.g. *cause-effect*, *parallel*, *contrast*). As a consequence, Knott (1996) argues, *elaboration* relations would better be described in terms of focus structures (cf. Grosz & Sidner (1986)), which Knott (1996) argues are less constrained, than in terms of rhetorical relations (cf. Hobbs (1985); Mann & Thompson (1988)), which Knott (1996) argues are more constrained. This hypothesis makes testable empirical claims: *elaboration* relations should in some way pattern differently from other coherence relations. We will come back to this issue in Sections 2.5.1 and 2.5.2.

In this thesis we present evidence against trees as a data structure for representing discourse coherence. Notice though that the evidence does not support the claim that discourse structures are completely arbitrary. The goal of our research program is to first determine which constraints on discourse structure are empirically viable. To us, the work we present here seems to be the crucial first step in order to avoid arbitrary constraints on inferences for building discourse structures. In other words, the point we wish to make here is that while there might be other constraints on possible discourse annotations that will have to be identified in future research, tree structure constraints do not seem to be the right kinds of constraints. This appears to be a crucial difference between approaches like Knott (1996)'s, Marcu (2000)'s, or Webber et al. (2003)'s on the one hand, and our approach on the other hand. The goal of the former approaches seems to be to first specify a set of constraints on possible discourse annotations, and then to annotate texts with these constraints in mind.

The following two sections will illustrate problems with trees as a representation of discourse coherence structures. Section 2.4.1 will show that the discourse structures of naturally occurring texts contain crossed dependencies, which cannot be represented in trees. Another problem for trees, in addition to crossed dependencies, is that many nodes in coherence graphs of naturally occurring texts have multiple parents. This is shown in Section 2.4.2. Because of these problems for trees, we will argue for a representation such as chain graphs (cf. Frydenberg (1989), Lauritzen & Wermuth (1989)), where directed arcs represent asymmetrical or directed coherence relations, and undirected arcs represent symmetrical or undirected coherence relations (this is equivalent to arguing for directed graphs with cycles). For all the examples in Sections 2.4.1 and 2.4.2, chain-graph-based analyses will be given. RST analyses will only be given for those examples that are also annotated by Carlson et al. (2002) (in that case, the RST analyses are those by Carlson et al. (2002)).

2.4.1 Crossed dependencies

Consider the text passage in (36) (modified from SAT practicing materials). Figure 11 shows the coherence graph for (36). Notice that the arrowheads of the arcs represent directionality for asymmetrical relations (*elaboration*) and bidirectionality for symmetrical relations (*similarity*, *contrast*).

(36)

7. Schools tried to teach students history of science.
8. At the same time they tried to teach them how to think logically and inductively.
9. Some success has been reached in the first of these aims.
10. However, none at all has been reached in the second.

(modified from SAT practicing materials)

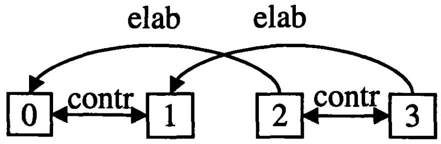


Figure 11. Coherence graph for (36). Abbreviations used: *contr = contrast*; *elab = elaboration*.

The coherence structure for (36) can be derived as follows:

- *Contrast* relation between discourse segments 0 and 1: discourse segments 0 and 1 describe teaching different things to students.
- *Contrast* relation between discourse segments 2 and 3: discourse segments 2 and 3 describe varying degrees of success (some vs. none).
- *Elaboration* relation between discourse segments 2 and 0: discourse segment 2 provides more details (the degree of success) about the teaching described in discourse segment 0.
- *Elaboration* relation between discourse segments 3 and 1: discourse segment 3 provides more details (the degree of success) about the teaching described in discourse segment 1.

In the resultant coherence structure for (36), there is a crossed dependency between {2, 0} and {3, 1}[10].

In order to be able to represent a structure like the one for (36) in a tree without violating validity assumptions about tree structures (Diestel (2000)), one might consider augmenting a tree either with feature propagation (Shieber (1986)) or with a coindexation mechanism (Chomsky (1973)).

There is a problem for both feature propagation and coindexation mechanisms: Both the tree structure itself as well as the features and coindexations represent the same kind of information (coherence relations). It is unclear how a dividing line could be drawn between tree structures and their augmentation. That is, it is unclear how one could decide which part of a text coherence structure should be represented by the tree structure and which part should be represented by the augmentation. Other areas of linguistics have faced this issue as well. Researchers investigating data structures for representing intra-sentential structure, for instance, generally fall into two groups. One group tries to formulate principles that allow representing some aspects of structure in the tree itself and other aspects in some augmentation formalism (e.g. Chomsky (1973);

[10] Notice that the structure for (36) might also be represented as
"elaboration (contrast (0, 1), contrast (2, 3))"
However, such a representation would not, for example, distinguish between the structure for (36) and the structure for an example where discourse segments 2 and 3 are reversed.

Marcus et al. (1994)). Another group argues that it is more parsimonious to assume a unified dependency-based representation that drops the tree constraints of allowing no crossed dependencies (e.g. Brants et al. (2002); Skut et al. (1997); König & Lezius (2000)). Our approach falls into the latter group. As we will point out, there does not seem to be a well-defined set of constraints on crossed dependencies in discourse structures. Without such constraints, it does not seem viable to represent discourse structures as augmented tree structures.

An important question is how many different kinds of crossed dependencies occur in naturally occurring discourse. If there are only a very limited number of different structures with crossed dependencies in natural texts, one could make special provisions to account for these structures and otherwise assume tree structures. (36), for instance, has a list-like structure. It is possible that list-like examples are exceptional in natural texts. However, there are many other naturally occurring non-list-like structures that contain crossed dependencies. As an example of a non-list-like structure with a crossed dependency (between {3, 1} and {2, 0-1}), consider the constructed example (37):

(37)

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

(constructed)

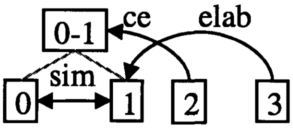


Figure 12. Coherence graph for (37). Abbreviations used: *sim* = *similarity*; ce = *cause-effect*; *elab* = *elaboration*.

The coherence structure for (37), shown in Figure 12, can be derived as follows:

- *Similarity* relation between 0 and 1: 0 and 1 both describe shopping for grocery items.
- *Cause-effect* relation between 2 and 0-1: 2 describes the cause for the shopping described by 0 and 1.
- *Elaboration* relation between 3 and 1: 3 provides details about the basil in 1.

(38) from the AP Newswire 1989 corpus is an example with a similar structure:

(38)

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

(from ap890109-0012; AP Newswire 1989 corpus; Harman & Liberman (1993))

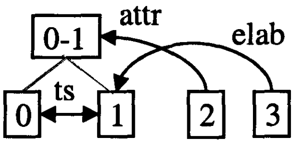


Figure 13. Coherence graph for (38). Abbreviations used: *ts = temporal sequence*; *attr = attribution*; *elab = elaboration.*

The coherence structure for (38) can be derived as follows:

- *Temporal sequence* relation between 0 and 1: 0 describes the takeoff that happens before the engine fire described by 1 occurs.
- *Attribution* relation between 2 and 0-1: 2 mentions the source of what is said in 0-1.
- *Elaboration* relation between 3 and 1: 3 provides more detail about the engine fire in 1.

The resulting coherence structure, shown in Figure 13, contains a crossed dependency between {3, 1} and {2, 0-1}.

(39)

0. [Mr. Baker's assistant for inter-American affairs,]0a [Bernard Aronson,]0b
1. while maintaining
2. that the Sandinistas had also broken the cease-fire,
3. acknowledged:
4. "It's never very clear who starts what."

(from wsj_0655; Wall Street Journal 1989 corpus; Harman & Liberman (1993))

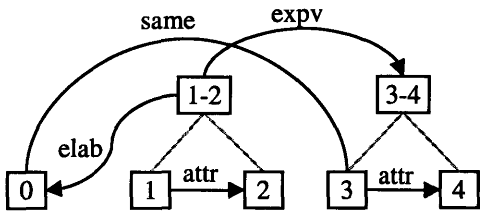


Figure 14. Coherence graph for (39). Additional abbreviation used: *expv = violated expectation.*

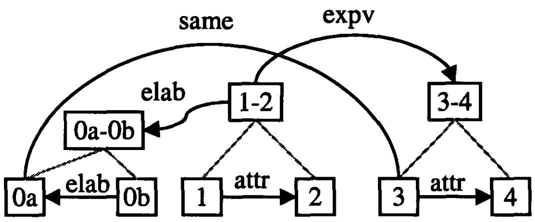


Figure 15. Coherence graph for (39) with discourse segment 0 split up into two segments.

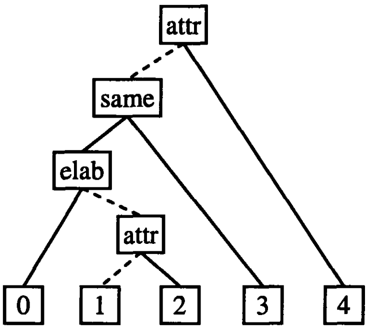


Figure 16. Tree-based RST-annotation for (39) from Carlson et al. (2002). Abbreviations used: attr = attribution; elab = elaboration. Dashed lines represent the start of asymmetric coherence relations; continuous lines represent the end of asymmetric coherence relations; symmetric coherence relations have two continuous lines (cf. Section 2.3.3).

Consider (39) from the Wall Street Journal 1989 corpus (Harman & Liberman (1993)). The annotations based on our annotation scheme are presented with the discourse segmentation based on the segmentation guidelines in Carlson et al. (2002)

(Figure 14), and based on our segmentation guidelines from Section 2.3.1 (Figure 15). Figure 16 shows a tree-based RST annotation for (39) from Carlson et al. (2002). The only difference between Carlson et al. (2002) and our approach with respect to how (39) is segmented is that Carlson et al. (2002) assume discourse segment 0 to be one single segment. By contrast, based on our segmentation guidelines, discourse segment 0 would be segmented into two segments (because of the comma that does not separate a complex NP or VP), 0a and 0b, as indicated by the angle brackets below[11]:

(40) [Mr. Baker's assistant for inter-American affairs, $]_{0a}$ [Bernard Aronson, $]_{0b}$

We then derived the coherence structure for (39) as follows:

- If discourse segment 0 is segmented into 0a and 0b (following our discourse segmentation guidelines): *elaboration* between 0a and 0b: 0b provides additional detail (a name) about what is stated in 0a (Mr. Baker's assistant).
- *Same* relation between 0 (or 0a) and 3: the subject NP in 0 ("Mr. Baker's assistant...") is separated from its predicate in 3 ("acknowledged") by intervening discourse segments 1 and 2 (and 0b in our discourse segmentation).
- *Attribution* relation between 1 and 2: 1 states the source of what is stated in 2 (the source in 1 is the elided "Mr. Baker").
- *Elaboration* relation between the group of discourse segments 1 and 2 and discourse segment 0 (or the group of discourse segments 0a and 0b in our discourse segmentation): 1 and 2 state additional detail (a statement about a political process) about what is stated in 0, or 0a and 0b (Mr. Baker's assistant).
- *Attribution* relation between 3 (and by virtue of the *same* relation also 0 or 0a) and 4: 3 states the source (Mr. Baker's assistant) of what is stated in 4.

[11] Based on our segmentation guidelines, the complementizer "that" in discourse segment 2 would be part of discourse segment 1 instead (cf. (26) in Section 2.3.3). However, since that would not make a difference in terms of the resulting discourse structure, we do not provide alternative analyses with "that" part of discourse segment 1 instead of discourse segment 2.

- *Violated expectation* relation between the group of discourse segments 1 and 2 and the group of discourse segments 3 and 4: although Mr. Baker's assistant acknowledged cease fire violations by one side (discourse segments 1 and 2), he acknowledges that it is in fact difficult to clearly blame one side for cease fire violations (discourse segments 3 and 4).

The resulting coherence structure, shown in Figure 15 (discourse segmentation from Carlson et al. (2002)) and Figure 16 (our discourse segmentation), contains a crossed dependency: the *same* relation between discourse segment 0 and discourse segment 3 crosses the *violated expectation* relation between the group of discourse segments 1 and 2 and the group of discourse segments 3 and 4.

Figure 16 represents a tree-based RST annotation for (39) from Carlson et al. (2002); in Figure 16, dashed lines represent the start of asymmetric coherence relations, and continuous lines mean the end of asymmetric coherence relations; symmetric coherence relations have two continuous lines (cf. Section 2.3.3 for the distinction between symmetric and asymmetric coherence relations, and for the directions of asymmetric coherence relations). We do not have a description of how Carlson et al. (2002) derived a tree-based RST structure for (39). Therefore, instead of discussing how the tree-based RST structure for (39) was derived, we show a comparison of the RST structure and our chain-graph-based structure in Table 5. Notice in particular that the RST structure does not represent the *violated expectation* relation between 1-2 and 3-4; that relation could not be annotated without violating the tree constraint of not allowing crossed dependencies.

Tree-based RST structure from Carlson et al. (2002)	Our chain-graph-based structure
(0a and 0b are one discourse segment)	*Elaboration* between 0a and 0b
Same between 0-2 and 3	*Same* between 0 (or 0a) and 3
Attribution between 0 and 1	*Attribution* between 0 and 1
Elaboration between 1-2 and 0	*Elaboration* between 1-2 and 0 (or 0a-0b)
Attribution between 0-3 and 4	*Attribution* between 3 and 4
(no relation)	*Violated expectation* between 1-2 and 3-4

Table 5. Comparison of tree-based RST structure from Carlson et al. (2002) and our chain-graph-based structure for (39).

2.4.2 Nodes with multiple parents

In addition to including crossed dependencies, many coherence structures of natural texts include nodes with multiple parents. Such nodes cannot be represented in tree structures. Consider (41) from the AP Newswire 1989 (Harman & Liberman (1993)). The coherence structure for (41) can be derived as follows:

- *Attribution* relation between 1 and 0 and 1 and 3 respectively: 1 states the source of what is stated in 0 and in 3.
- *Elaboration* relation between 2 and 1: 2 provides additional detail (the name) about the BMW driver in 1.
- *Condition* relation between 3 and 0: 3 states the BMW driver's condition for being polite, stated in 0[12]. This *condition* relation is also indicated by the phrase "as long as".

In the resultant coherence structure for (41), node 1 has two parents – one *attribution* and one *condition* ingoing arc (cf. Figure 17).

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

(41)

0. "Sure I'll be polite,"
1. promised one BMW driver
2. who gave his name only as Rudolf.
3. "As long as the trucks and the timid stay out of the left lane."

(from ap890103-0014; AP Newswire 1989 corpus; Harman & Liberman (1993))

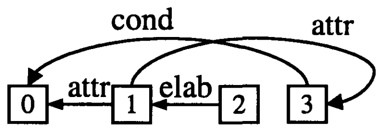


Figure 17. Coherence graph for (41). Additional abbreviation used: *cond* = *condition*.

As another example of a discourse structure that contains nodes with multiple parents, consider the structure of (42) from the AP Newswire 1989 corpus (Harman & Liberman (1993)). Our annotations are shown in Figure 18 (discourse segmentation from Carlson et al. (2002)) and Figure 19 (our discourse segmentation). The only difference between our annotation and Carlson et al. (2002)'s is that we do not assume two separate discourse segments for 0 and 1; 0 and 1 are one discourse segment in our annotation (represented by the node "0+1" in Figure 19). Notice also that in (42), discourse segment 2, "that" is not in a separate discourse segment; it is unclear why in (42) "that" is in a separate discourse segment (discourse segment 1) and not part of discourse segment 2[13].

The discourse structure for (42) can be derived as follows:

- According to our discourse segmentation guidelines (cf. Section 2.3.1), 0 and 1 should be one single discourse segment: therefore either *same* relation between 0 and 1 (cf. Figure 18), or merge 0 and 1 into one single discourse segment, 0+1 (cf. Figure 19).

[13] Based on our segmentation guidelines, the complementizer "that" in discourse segment 2 would be part of discourse segment 1 instead (cf. (26) in Section 2.3.3). However, since that would not make a difference in terms of the resulting discourse structure, we do not provide alternative analyses with "that" part of discourse segment 1 instead of discourse segment 2.

- *Attribution* relation between 0 or 0+1 and 2-3: 0 or 0+1 state the source (the administration) of what is stated in 2-3.
- *Condition* relation between 2 and 3: 2 states the condition for what is stated in 3 (the *condition* relation is also signaled by the cue phrase "if" in 2).
- *Attribution* relation between 4 and 0-3: 4 states the source of what is stated in 0-3.
- *Attribution* relation between 4 and 5: 4 states the source of what is stated in 5.
- *Evaluation-s*[14] relation between 5 and 2-3: 2-3 state what is evaluated by 5 – the Contra supporters should call for military aid, and if the February election is voided (group of discourse segments 2-3), the Contra supporters might win (discourse segment 5). Notice that in our annotation scheme, the *evaluation-s* relation would be an *elaboration* relation (5 provides additional detail about 2-3: Elliott Abrams' opinion on the Contras' chances to win).

In the resultant coherence structure for (42), node 2-3 has multiple parents or ingoing arcs – one *attribution* ingoing arc and one *evaluation-s* ingoing arc (cf. Figure 18 and Figure 19).

(42)

0. "The administration should now state
1. that
2. if the February election is voided by the Sandinistas
3. they should call for military aid,"
4. said former Assistant Secretary of State Elliott Abrams.
5. "In these circumstances, I think they'd win."

["they" in 3 and 5 = "Contra supporters"; this is clear from the whole text wsj_0655]

(from wsj_0655; Wall Street Journal 1989 corpus; Harman & Liberman (1993))

[14] The relation *evaluation-s* is part of the annotation scheme in Carlson et al. (2002) but not part of our annotation scheme. In an *evaluation-s* relation, the situation presented in the satellite assesses the situation presented in the nucleus (Carlson et al. (2002)). An *evaluation-s* relation would be an *elaboration* relation in our annotation scheme.

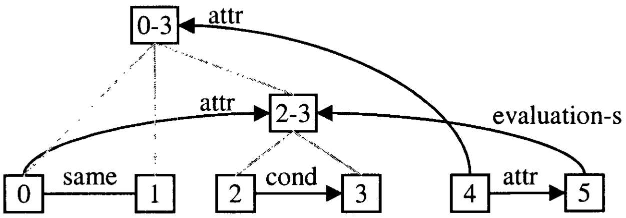


Figure 18. Coherence graph for (42). Additional coherence relation used (from Carlson et al. (2002)): *evaluation-s = the situation presented in the satellite assesses the situation presented in the nucleus* (*evaluation-s* would be *elaboration* in our annotation scheme).

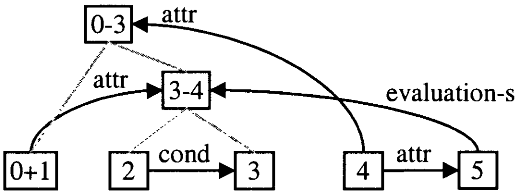


Figure 19. Coherence graph for (42) with discourse segments 0 and 1 merged into one single discourse segment.

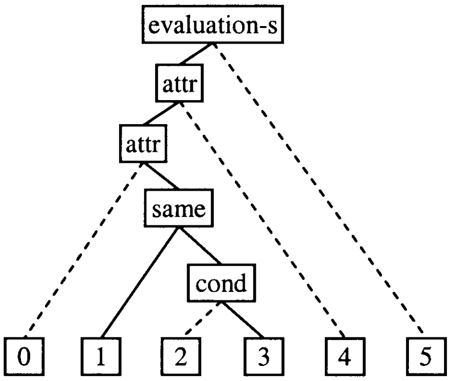


Figure 20. Tree-based RST annotation for (42) from Carlson et al. (2002). Dashed lines represent the start of asymmetric coherence relations; continuous lines represent the end of asymmetric coherence relations; symmetric coherence relations have two continuous lines (cf. Section 2.3.3).

As for (41), we do not have a description available to us of how Carlson et al. (2002) derived their tree-based RST annotation (Figure 20). Therefore, as for (41), instead of presenting how the tree-based RST annotation in Figure 20 was derived, we present a comparison of the RST annotation and our chain-graph-based annotation in Table 6. Notice in particular that the *attribution* relation between 4 and 5 cannot be represented in the RST tree structure. Notice furthermore that the RST tree contains an *evaluation-s* relation between 5 and 0-4. However, this *evaluation-s* relation seems to hold rather between 5 and 2-3: what is being evaluated is a chance for the Contras to win a military conflict under certain circumstances. But annotating a coherence relation between 5 and 2-3 could not have been accommodated in a tree structure.

Tree-based RST structure from Carlson et al. (2002)	Our chain-graph-based structure
Same between 1 and 2-3	*Same* between 0 and 1, or merging of 0 and 1 to 0+1
Attribution between 0 and 1-3	*Attribution* between 0 or 0+1 and 2-3
Condition between 2 and 3	*Condition* between 2 and 3
Attribution between 4 and 0-3	*Attribution* between 4 and 0-3
(no relation)	*Attribution* between 4 and 5
Evaluation-s between 5 and 0-4	*Evaluation-s* between 5 and 2-3

Table 6. Comparison of tree-based RST structure from Carlson et al. (2002) and our chain-graph-based structure for (42).

2.5 Statistics

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

An important question is how frequent the phenomena discussed in the previous sections are. The more frequent they are, the more urgent the need for a data structure that can adequately represent them. The following sections report statistical results on crossed dependencies (Section 2.5.1) and nodes with multiple parents (Section 2.5.2).

2.5.1 Crossed dependencies

The following sections report counts on crossed dependencies in the annotated database of 135 texts (cf. Section 2.1). Section 2.5.1.1 reports results on the frequency of crossed dependencies, Section 2.5.1.2 reports results concerning the question of what types of coherence relations tend to be involved in crossed dependencies, and Section 2.5.1.3 reports results on the arc lengths of coherence relations involved in crossed dependencies. Section 2.5.1.4 provides a short summary of the statistical results on crossed dependencies.

2.5.1.1 Frequency of crossed dependencies

In order to track the frequency of crossed dependencies for the coherence structure graph of each text, we counted the minimum number of arcs that would have to be deleted in order to make the coherence structure graph free of crossed dependencies. The example graph in Figure 21 illustrates this process. This graph contains the following crossed dependencies: {0, 1} crosses with {1, 3}, {2, 4} with {1, 3}, and {4, 6} with {5, 7}. By deleting {1, 3}, two crossed dependencies can be eliminated: the crossing of {0, 1} and {1, 3}, and the crossing of {2, 4} with {1, 3}. By deleting either {4, 6} or {5, 7} the remaining crossed dependency between {4, 6} and {5, 7} can be eliminated. Therefore two edges would have to be deleted from the graph in Figure 21 in order to make it free of crossed dependencies.

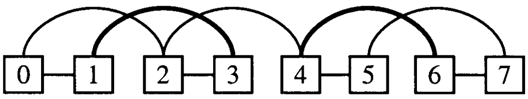


Figure 21. Example graph with crossed dependencies.

Table 7 shows the results of the counts. On average for the 135 annotated texts, 12.5% of arcs in a coherence graph have to be deleted in order to make the graph free of crossed dependencies. Seven texts out of 135 had no crossed dependencies. The mean number of arcs for the coherence graphs of these texts was 36.9 (minimum: 8; maximum: 69; median: 35). The mean number of arcs for the other 128 coherence graphs (those with crossed dependencies) was 125.7 (minimum: 20; maximum: 293; median: 115.5). Thus, the graphs with no crossed dependencies have significantly fewer arcs than those graphs that have crossed dependencies ($\chi^2(1)$ = 15330.35 (Yates' correction for continuity applied); $p < 10^{-6}$). This is a likely explanation for why these seven texts had no crossed dependencies.

mean	12.5
minimum	0
maximum	44.4
median	10.9

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

More generally, linear regressions show a correlation between the number of arcs in a coherence graph and the number of crossed dependencies. The more arcs a graph has, the higher the number of crossed dependencies ($R^2 = 0.39$; $p < 10^{-4}$; cf. Figure 22). The same linear correlation holds between text length and number of crossed dependencies – the longer a text, the more crossed dependencies are in its coherence structure graph (for text length in discourse segments: $R^2 = .29$, $p < 10^{-4}$; for text length in words: $R^2 = .24$, $p < 10^{-4}$).

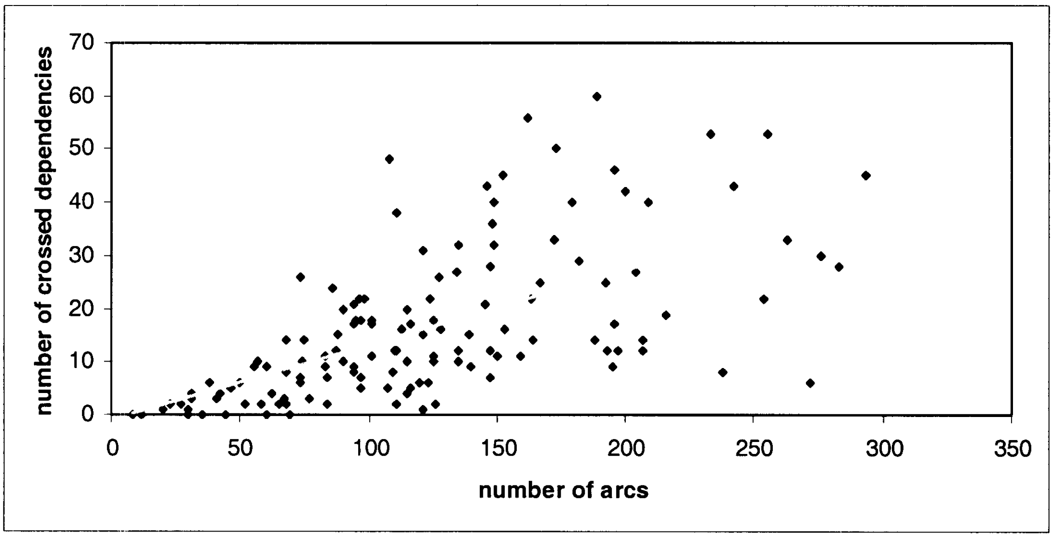


Figure 22. Correlation between number of arcs and number of crossed dependencies.

2.5.1.2 Types of coherence relations involved in crossed dependencies

In addition to the question of how frequent crossed dependencies are, another question is whether there are certain types of coherence relations that participate more or less frequently in crossed dependencies than other types of coherence relations. For an arc to participate in a crossed dependency, it means that the arc is in the set of arcs that would have to be deleted from a coherence graph in order to make that graph free of crossed dependencies (cf. the procedure outlined in the beginning of Section 2.5.1). In other words, the question is whether the frequency distribution over types of coherence relations is different for arcs participating in crossed dependencies compared to the overall frequency distribution over types of coherence relations in the whole database.

Figure 23 shows that the overall distribution over types of coherence relations participating in crossed dependencies is not different from the distribution over types of coherence relations overall. This is confirmed by a linear regression, which shows a significant correlation between the two distributions of percentages ($R^2 = 0.84$; $p < .0001$). Notice that the overall distribution includes only arcs with length greater than one, since arcs of length one could not participate in crossed dependencies.

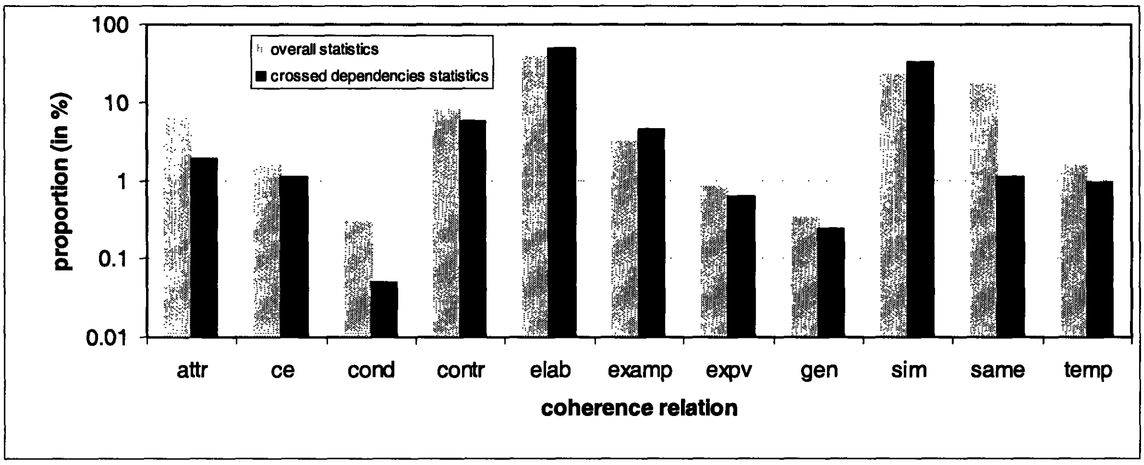


Figure 23. Distributions over types of coherence relations. For each condition ("overall statistics" and "crossed dependencies statistics"), the sum over all coherence relations is 100; each bar in each condition represents a fraction of the total of 100 in that condition. The y-axis uses a $\log_{10}$ scale.

However, there are some differences for individual coherence relations. Some types of coherence relations occur considerably less frequently in crossed dependencies than overall in the database. Table 8 shows the data from Figure 23, ranked by the factor of "proportion of overall coherence relations" by "proportion of coherence relations participating in crossed dependencies". The proportion of *same* relations, for instance, is 15.23 times greater, and the percentage of *condition* relations is 5.59 times greater overall in the database than in crossed dependencies. We do not yet understand the reason for these differences, and plan to address this question in future research.

Coherence relation	**Proportion of coherence relations participating in crossed dependencies (in %)**	**Proportion of overall coherence relations (in %)**	**Factor (= overall / crossed dependencies)**
same	1.13	17.21	15.23
condition	0.05	0.28	5.59
attribution	1.93	6.31	3.27
temporal sequence	0.94	1.56	1.66
generalization	0.24	0.34	1.40
contrast	5.84	7.93	1.36
cause-effect	1.13	1.53	1.35
violated expectation	0.61	0.82	1.40
elaboration	50.52	37.97	0.71
example	4.43	3.15	1.34
similarity	33.18	22.91	0.69

Table 8. Proportions of coherence relations.

Another way of testing whether certain coherence relations contribute more than others to crossed dependencies is to remove coherence relations of a certain type from the database and then count the remaining number of crossed dependencies. For example, it is possible that the number of crossed dependencies is reduced once all *elaboration* relations are removed from the database. Table 9 shows that by removing all *elaboration* relations from the database of 135 annotated texts, the percentage of coherence relations involved in crossed dependencies is reduced from 12.5% to 4.96% of the remaining coherence relations. That percentage is reduced even further, to 0.84%, by removing all *elaboration* and *similarity* relations from the database. These numbers seem to be partial support for Knott (1996)'s hypothesis: Knott (1996) argued that *elaboration* relations are less constrained than other types of coherence relations (cf. the discussion of Knott (1996) in Section 2.4).

	Remaining percentage of coherence relations involved in crossed dependencies			
Coherence relation removed	**mean**	**min**	**max**	**median**
Same	13.08	0	44.44	11.39
Condition	12.63	0	45.28	10.89
Attribution	13.44	0	44.86	11.36
temporal sequence	12.53	0	44.44	10.87
Generalization	12.53	0	44.44	10.84
Contrast	11.88	0	46.15	9.86
cause-effect	12.67	0	49.47	11.03
violated expectation	12.51	0	44.44	10.87
Elaboration	4.96	0	47.47	1.23
Example	12.08	0	44.44	9.89
Similarity	7.32	0	24.56	7.04
elaboration and similarity	0.84	0	10.68	0.00

Table 9. The effect of removing different types of coherence relations on the percentage of coherence relations involved in crossed dependencies.

However, there is a possible alternative hypothesis to Knott (1996). In particular, *elaboration* relations are very frequent (37.97% of all coherence relations, cf. Table 8). It is possible that removing *elaboration* relations from the database only reduces the number of crossed dependencies because a large number of coherence relations, 37.97%, are removed when *elaborations* are removed. In other words, an alternative hypothesis to Knott (1996) is that the lower number of crossed dependencies is just due to less dense coherence graphs (i.e. the less dense coherence graphs are, the lower the chance for crossed dependencies). We tested this hypothesis by correlating the proportion of coherence relations removed with the proportion of crossed dependencies that remain after removing a certain type of coherence relation[15]. Figure 24 shows that the higher the proportion of removed coherence relations, the lower the proportion of coherence

[15] Notice that the proportions of removed coherence relations do not include coherence relations of absolute arc length 1, since removing those coherence relations cannot have any influence on the number of crossed dependencies (coherence relations of absolute arc length 1 cannot be involved in crossed dependencies). Thus, the proportions of coherence relations removed in Figure 24 are from the third column of Table 8.

relations becomes that are involved in crossed dependencies. This correlation is confirmed by a linear regression ($R^2 = 0.7697$; $p < .0005$; after removing the *elaboration* data point: $R^2 = 0.4504$; $p < .05$; these linear regressions do not include the data point *elaboration* + *similarity*). Thus, while removing certain types of coherence relations reduces the number of crossed dependencies, it results in a very impoverished representation of coherence structure (i.e. after removing all *elaboration* and all *similarity* relations, only 39.12% of all coherence relations would still be represented, cf. Table 8; 52.13% based on the distribution over coherence relations including those with absolute arc length 1, cf. Table 11).

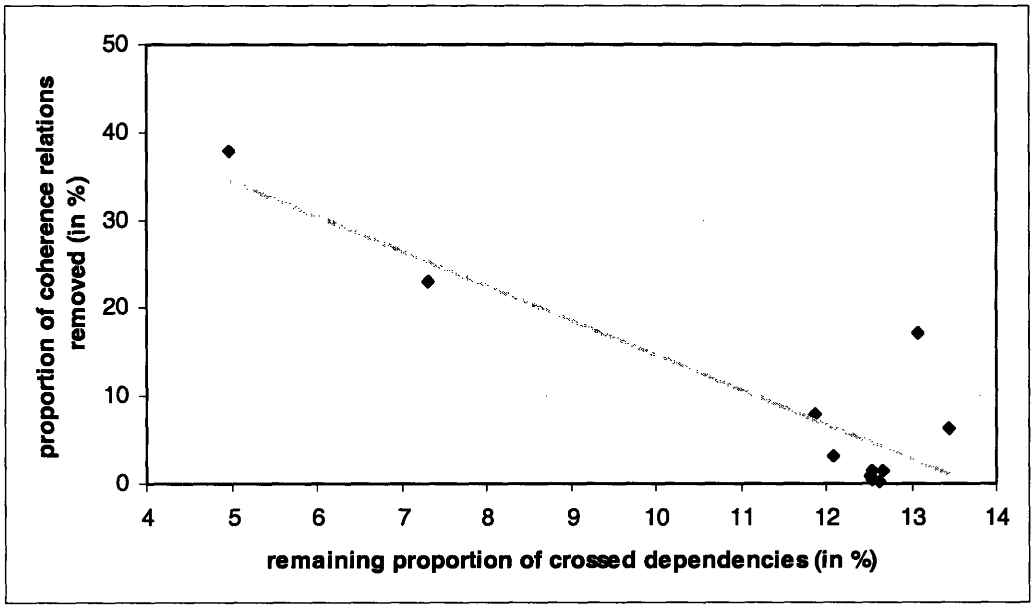


Figure 24. Correlation between removed proportion of overall coherence relations and remaining proportion of crossed dependencies. Notice that the data point for *elaboration* + *similarity* is not included in the graph above. Both axes represent percent values. $R^2 = 0.7699$; $p < .0005$.

With respect to Knott (1996)'s hypothesis, notice that leaving out *elaboration* relations still leaves the proportion of remaining crossed dependencies at 4.96% (cf. Table 9). In order to further reduce the proportion of remaining crossed dependencies, it is necessary to remove *similarity* relations in addition to removing *elaboration* relations (cf. Table 9). This is a pattern of results that is not predicted by any literature that we are aware of (including Knott (1996) among others, although he predicts these results partially). We believe this issue should be addressed in future research.

2.5.1.3 Arc length of coherence relations involved in crossed dependencies

Another question is how great the distance typically is between discourse segments that participate in crossed dependencies, or how great the arc length is for coherence relations that participate in crossed dependencies[16]. It is possible, for instance, that crossed dependencies primarily involve long-distance arcs and that more local crossed dependencies are disfavored. However, Figure 25 shows that the distribution over arc lengths is practically identical for the overall database and for coherence relations participating in crossed dependencies (linear regression: $R^2 = 0.937$, $p < 10^{-4}$), suggesting a strong locality bias for coherence relations overall as well as for those participating in crossed dependencies[17]. The arc lengths are normalized in order to take into account the length of a text. Normalized arc length is calculated by dividing the absolute length of an arc by the maximum length that that arc could have, given its position in a text. For example, if there is a coherence relation between discourse segment 0 and discourse segment 3 in a text, the raw distance would be 3. If these discourse segments are part of a text that has 5 discourse segments total (i.e. 0 to 4), the normalized distance would be 3 / 4 = 0.75 (because 4 would be the maximum possible length of an arc that originates in discourse segment 0 or 3, given that the text has 5 discourse segments in total).

[16] The distance between two discourse segments is measured not in terms of how many coherence links one has to follow from any discourse segment *x* to any discourse segment *y* to which discourse segment *x* is related via a coherence relation. Instead, distance is measured in terms of the number of intervening discourse segments. Thus, distance between nodes reflects linear distance between two discourse segments in a text. For example, the distance between a discourse segment 1 and a discourse segment 4 would be 3.

[17] The arc length distribution for the database overall does not include arcs of (absolute) length 1, since such arcs could not participate in crossed dependencies.

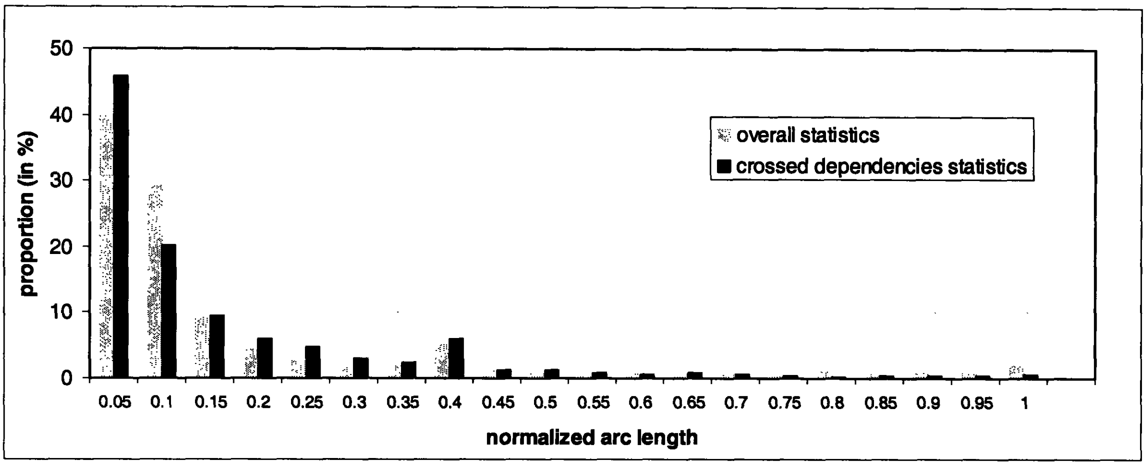


Figure 25. Comparison of normalized arc length distributions. For each condition ("overall statistics" and "crossed dependencies statistics"), the sum over all coherence relations is 100; each bar in each condition represents a fraction of the total of 100 in that condition.

2.5.1.4 Summary on crossed dependencies statistics

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

2.5.2 Nodes with multiple parents

Section 2.4.2 provided examples of coherence structure graphs that contain nodes with multiple parents. In addition to crossed dependencies, nodes with multiple parents are another reason why trees are inadequate for representing natural language coherence structures. The following sections report statistical results from our database on nodes with multiple parents. Similar to the previous section on crossed dependencies, we report results on the frequency of nodes with multiple parents (Section 2.5.2.1), the types of coherence relations ingoing to nodes with multiple parents (Section 2.5.2.2), and the arc length of coherence relations ingoing to nodes with multiple parents (Section 2.5.2.3). Section 2.5.2.4 provides a short summary of the statistical results on nodes with multiple parents.

2.5.2.1 Frequency of nodes with multiple parents

We determined the frequency of nodes with multiple parents by counting the number of nodes with in-degree greater than 1. We assume nodes with in-degree greater than 1 in a graph to be the equivalent of nodes with multiple parents in a tree. The result of our count indicated that 41.22% of all nodes in the database have an in-degree greater than 1. In addition to counting the number of nodes with in-degree greater than 1, we determined the mean in-degree of the nodes in our database. Table 10 shows that the mean in-degree (=mean number of parents) of all nodes in the investigated database of 135 texts is 1.6. Similar as for coherence relations involved in crossed dependencies (cf. Section 2.5.1.1), a linear regression showed a significant correlation between the number of arcs in a coherence graph and the number of nodes with multiple parents (cf. Figure 26; $R^2 = 0.7258$, $p < 10^{-4}$; for text length in discourse segments: $R^2 = .6999$, $p < 10^{-4}$; for text length in words: $R^2 = .6022$, $p < 10^{-4}$). The proportion of nodes with in-degree greater than 1 and the mean in-degree of the nodes in our database suggest that even if a mechanism could be derived for representing crossed dependencies in (augmented) tree graphs, nodes with multiple parents present another significant problem for trees representing coherence structures.

mean	1.60
min	1
max	12
median	1

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

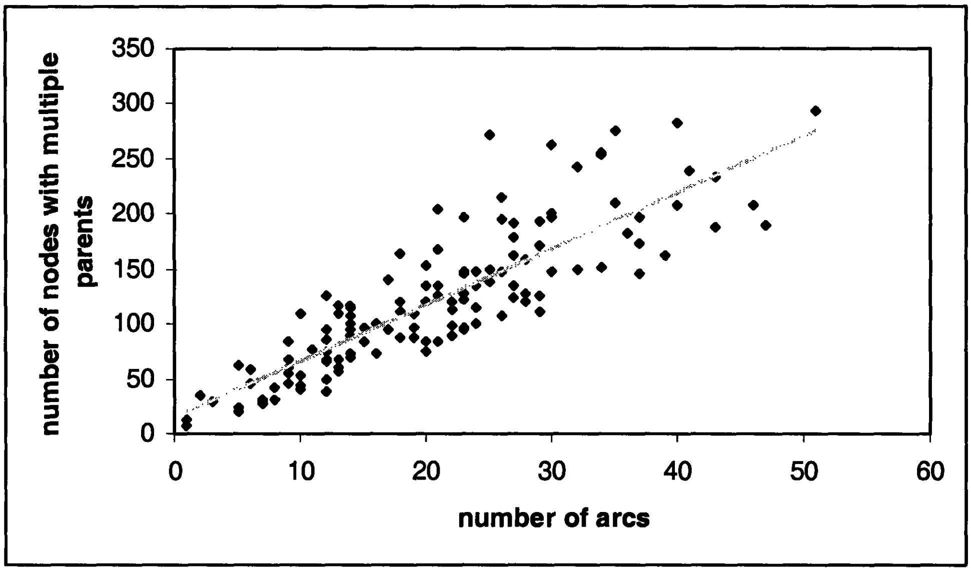


Figure 26. Correlation between number of arcs and number of nodes with multiple parents.

2.5.2.2 Types of coherence relations ingoing to nodes with multiple parents

As with crossed dependencies, an important question is whether there are certain types of coherence relations that are more or less frequently ingoing to nodes with multiple parents than other types of coherence relations. In other words, the question is whether the frequency distribution over types of coherence relations is different for arcs ingoing to nodes with multiple parents compared to the overall frequency distribution over types of coherence relations in the whole database. Figure 27 shows that the overall distribution over types of coherence relations ingoing to nodes with multiple parents is not different from the distribution over types of coherence relations overall[18]. This is confirmed by a linear regression, which shows a significant correlation between the two distributions of percentages ($R^2 = 0.967$ $p < 10^{-4}$).

[18] Notice that, unlike in Section 2.5.1.2, the distribution over coherence relations for all coherence relations includes arcs with length 1, since this time there was no reason to exclude them.

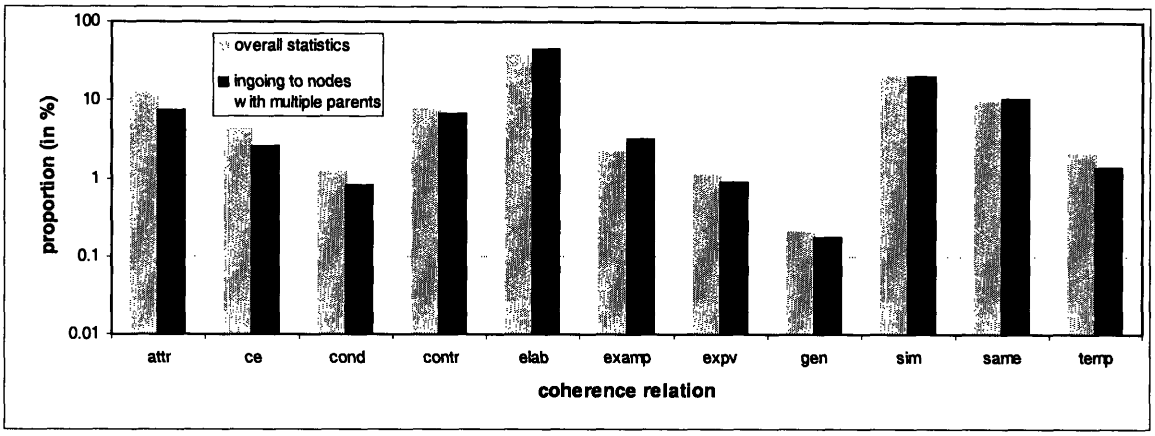


Figure 27. Distributions over types of coherence relations. For each condition ("overall statistics" and "ingoing to nodes with multiple parents"), the sum over all coherence relations is 100; each bar in each condition represents a fraction of the total of 100 in that condition. The y-axis uses a $\log_{10}$ scale.

Unlike for crossed dependencies (cf. Table 8), there are no big differences for individual coherence relations. Table 11 shows the data from Figure 27, ranked by the factor of "proportion of overall coherence relations" by "proportion of coherence relations ingoing to nodes with multiple parents".

Coherence relation	**Proportion of coherence relations ingoing to nodes with multiple parents (in %)**	**Proportion of overall coherence relations (in %)**	**Factor (= overall / ingoing to nodes with multiple parents)**
attribution	7.38	12.68	1.72
cause-effect	2.63	4.19	1.59
temporal sequence	1.38	2.11	1.53
condition	0.83	1.21	1.46
violated expectation	0.90	1.13	1.26
generalization	0.17	0.21	1.22
contrast	6.72	7.62	1.13
same	10.72	9.74	0.91
similarity	20.22	20.79	1.03
elaboration	45.83	38.13	0.83
example	3.20	2.19	0.68

Table 11. Proportion of coherence relations.

As for crossed dependencies, we also tested whether removing certain kinds of coherence relations reduces the mean in-degree (number of parents) and / or the proportion of nodes with in-degree greater than 1 (more than one parent). Table 12 shows that removing all *elaboration* relations from the database reduces the mean in-degree of nodes from 1.60 to 1.238, and the proportion of nodes with in-degree greater than 1 from 41.22% to 20.29%. Removing all *elaboration* as well as all *similarity* relations reduces these numbers further to 1.142 and 11.24% respectively. As Table 12 also shows, removing other types of coherence relations does not lead to as great reduction of the mean in-degree and proportion of nodes with in-degree greater than one.

Coherence relation removed	In-degree of nodes				Proportion of nodes with in-degree > 1 (in %)
	mean	min	max	median	
same	1.519	1	12	1	35.85
condition	1.599	1	12	1	41.01
attribution	1.604	1	12	1	41.18
temporal sequence	1.599	1	12	1	41.12
generalization	1.6	1	12	1	41.16
contrast	1.569	1	12	1	39.45
cause-effect	1.599	1	12	1	41.14
violated expectation	1.598	1	12	1	40.96
elaboration	1.238	1	11	1	20.29
example	1.574	1	11	1	40.37
similarity	1.544	1	12	1	36.25
elaboration and similarity	1.142	1	11	1	11.24

Table 12. The effect of removing different types of coherence relations on the mean in-degree of nodes and on the proportion of nodes with in-degree > 1.

However, as with crossed dependencies (cf. Section 2.5.1.2), we also tested whether the reduction in nodes with multiple parents could simply be due to removing more and more coherence relations (i.e. the less dense a graph is, the smaller the chance that there are nodes with multiple parents). We correlated the proportion of coherence relations removed with the mean in-degree of the nodes after removing different types of coherence relations[19]. Figure 28 shows that the higher the proportion of removed coherence relations, the lower the mean in-degree of the nodes in the database becomes. This correlation is confirmed by a linear regression ($R^2 = 0.9455$; $p < 10^{-4}$; after removing the *elaboration* data point: $R^2 = 0.8310$; $p < .0005$; notice that these linear regressions do not include the data point *elaboration + similarity*). We also correlated the proportion of coherence relations removed with the proportion of nodes with in-degree greater than one

[19] Notice that in the correlations in this chapter, the proportions of removed coherence relations include coherence relations of absolute arc length 1, because removing these coherence relations also has an effect on the mean in-degree of nodes and the proportion of nodes with in-degree greater than 1. Thus, the proportions of coherence relations removed in Figure 28 and in Figure 29 are from the third column of Table 11.

after removing different types of coherence relations. Figure 29 shows that the higher the proportion of removed coherence relations, the lower the proportion of nodes with in-degree greater than one. This correlation is also confirmed by a linear regression ($R^2 = 0.9574$, $p < 10^{-4}$; after removing the *elaboration* data point: $R^2 = 0.8146$, $p < .0005$; notice that these correlations do not include the data point *elaboration + similarity*).

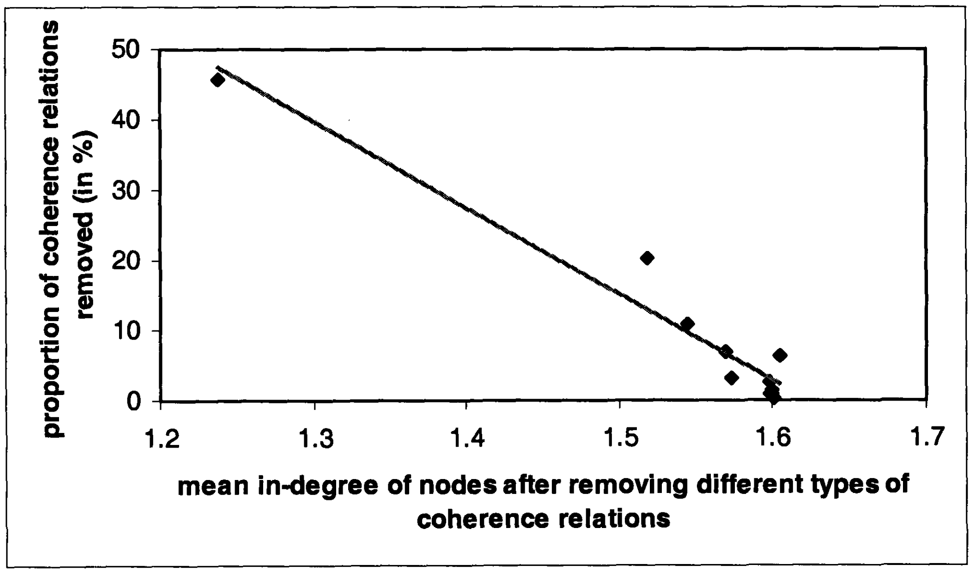


Figure 28. Correlation between proportion of removed coherence relations and mean in-degree of remaining nodes. Notice that the data point for *elaboration + similarity* is not included in the graph above. Both axes represent percent values. $R^2 = 0.9455$; $p < 10^{-4}$.

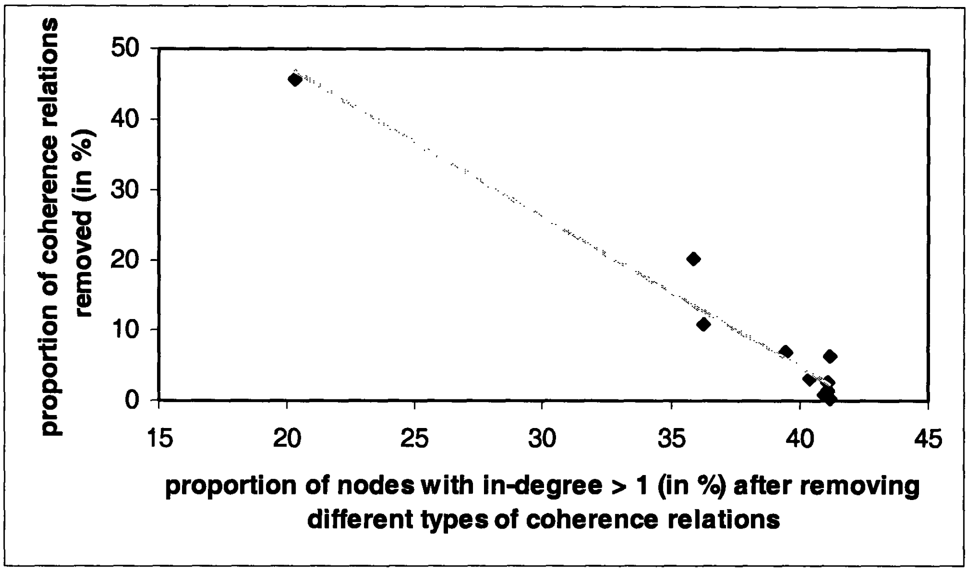


Figure 29. Correlation between proportion of removed coherence relations and proportion of nodes with in-degree > 1. Notice that the data point for *elaboration* + *similarity* is not included in the graph above. Both axes represent percent values. $R^2 = 0.9574$; $p < 10^{-4}$.

Thus, while removing certain types of coherence relations (the same as for crossed dependencies, i.e. *elaboration* and *similarity*, cf. Section 2.5.1.2) can reduce the mean in-degree of nodes and the proportion of nodes with in-degree greater than one, the result is a very impoverished coherence structure. For example, after removing both *elaboration* and *similarity* relations, only 52.13% of all coherence relations would still be represented (cf. Table 11). Furthermore, notice that this pattern of results is not predicted by any literature we are aware of, including Knott (1996), although he predicts the results partially (he predicts that removing *elaboration* relations but not that removing *elaboration* as well as *similarity* relations is necessary in order to remove basically all nodes with multiple parents, cf. the discussion in the last paragraph of Section 2.5.1.2). This issue will have to be investigated in future research.

2.5.2.3 Arc lengths of coherence relations ingoing to nodes with multiple parents

As for crossed dependencies, we also compared arc lengths. Here, we compared the length of arcs that are ingoing to nodes with multiple parents to the overall distribution of arc length. Again, we compared normalized arc lengths (see Section 2.5.1 for the normalization procedure). By contrast to the comparison for crossed dependencies, we included arcs of (absolute) length 1 because such arcs can be ingoing to nodes with either single or multiple parents. Figure 30 shows that the distribution over arc lengths is practically identical for the overall database and for arcs ingoing to nodes with multiple parents (linear regression: $R^2 = 0.993$, $p < 10^{-4}$), suggesting a strong locality bias for coherence relations overall as well as for those participating in crossed dependencies.

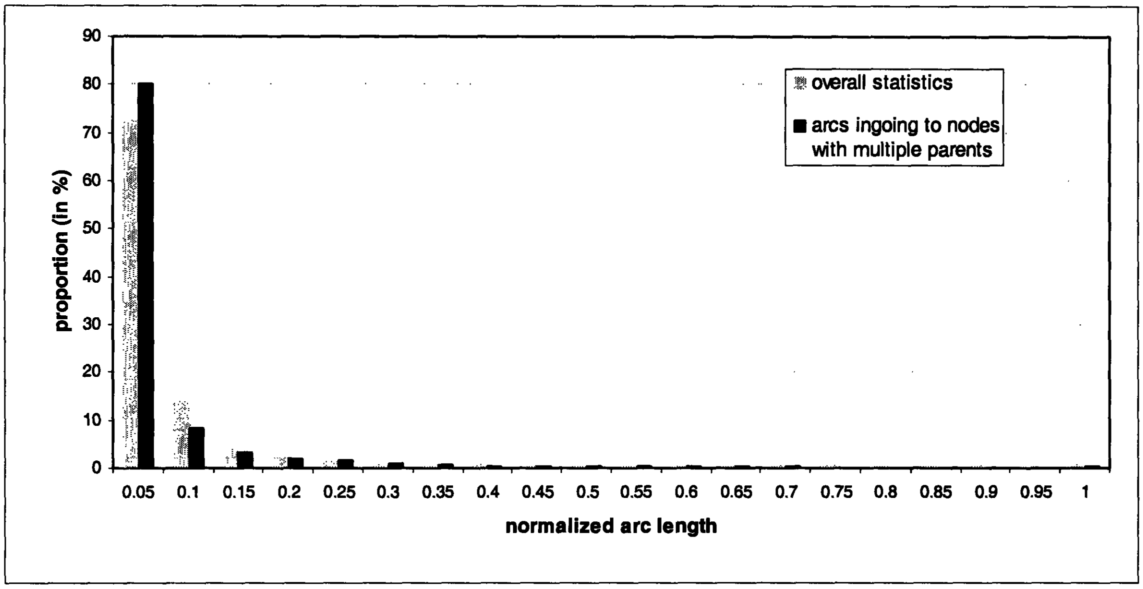


Figure 30. Comparison of normalized arc length distributions. For each condition ("overall statistics" and "arcs ingoing to nodes with multiple parents"), the sum over all coherence relations is 100; each bar in each condition represents a fraction of the total of 100 in that condition.

2.5.2.4 Summary of statistical results on nodes with multiple parents

In sum, statistical results on nodes with multiple parents suggest that they are a frequent phenomenon, and that they are not limited to certain kinds of coherence relations. However, similar to crossed dependencies, removing certain kinds of coherence relations (*elaboration* and *similarity*) can reduce the mean in-degree of nodes and the proportion of nodes with in-degree greater than 1. But, also similar to crossed dependencies, our data at present do not distinguish whether this reduction in nodes with multiple parents is due to a property of the coherence relations removed (*elaboration* and *similarity*), or whether it is just that removing more and more coherence relations simply reduces the chance for nodes to have multiple parents. We plan to address this question in future research. In addition to the results on frequency of nodes with multiple parents and types of coherence relations ingoing to nodes with multiple parents, the statistical results reported here suggest that ingoing arcs to nodes with multiple parents are primarily local.

2.6 Conclusion

The goals of this chapter have been to present a set of coherence relations that are easy to code, and to illustrate the inadequacy of trees as a data structure for representing discourse coherence structures. We have developed a coding scheme with high inter-annotator reliability and used that scheme to annotate 135 texts with coherence relations. An investigation of these annotations has shown that discourse structures of naturally occurring texts contain various kinds of crossed dependencies as well as nodes with multiple parents. Both phenomena cannot be represented using trees. This implies that existing databases of coherence structures that use trees are not descriptively adequate.

Our statistical results suggest that crossed dependencies and nodes with multiple parents are not restricted phenomena that could be ignored or accommodated with a few exception rules. For sentence structures, it has been argued that a principled distinction can be made between phenomena like phrase structure on the one and phenomena like agreement features on the other hand (e.g. Shieber (1986), Chomsky (1973)). For discourse structures, on the other hand, our statistical results suggest no obvious equivalent of that distinction that would allow representing some aspects of coherence structures in trees and some aspects in augmentations to these trees.

Because trees are neither a descriptively adequate data structure for representing coherence structures nor easier to derive, we argue for less constrained graphs as a data structure for representing coherence structures. In particular, we argue for a representation such as chain graphs (cf. last paragraph of Section 2.4). Such less constrained graphs would have the advantage of being able to adequately represent coherence structures in one single data structure (cf. Brants et al. (2002); Skut et al. (1997); König & Lezius (2000)). Furthermore, they are at least not harder to derive than (augmented) tree structures. The greater descriptive adequacy might in fact make them easier to derive. However, this is still an open issue and will have to be addressed in future research.

In Section 2.3.3 we briefly illustrated the possibility of more fine-grained discourse segmentation than in the current project. While such a detailed annotation of coherence relations was beyond the scope of the current project, future research should address this issue. More fine-grained discourse segmentation could then also facilitate integrating discourse-level with sentence-level structural descriptions.

Another issue that should be addressed in future research is empirically viable constraints on inferences for building discourse structures. As pointed out in Section 2.4, we have argued against trees as a data structure for representing discourse structures; however, that does not necessarily mean that discourse structures can be completely arbitrary. Future research should investigate questions such as whether there are structural constraints on coherence graphs (e.g. as proposed by Danlos (2004)), or whether there are systematic structural differences between the coherence graphs of texts that belong to different genres (e.g. as proposed by Bergler (1992)).

3 Discourse coherence and pronoun resolution

This chapter will test whether coherence can influence other linguistic processes, in particular, pronoun processing. For informational coherence relations, this has been claimed in a hypothesis by Kehler (2002) that is based on Hobbs (1979). Other predictions about pronoun processing that will be tested here are either structural and pronoun-specific (Chambers & Smyth (1998)), or part of an account of attentional discourse structure (Grosz & Sidner (1986); Wundt (1911)). The claims will be evaluated in an on-line language comprehension study (Experiment 1) as well as in an off-line language production study (Experiment 2).

3.1 Introduction

An important component of language comprehension in most natural language contexts involves connecting clauses and phrases together in order to establish a coherent discourse. One critical way in which coherence can be established between clauses is by the use of referring expressions, such as pronouns (Garnham (2001); Haliday & Hassan (1976); Johnson-Laird (1983); Kintsch & van Dijk (1978); Sanford & Garrod (1989)). Thus an important part of discourse comprehension involves discovering how antecedents for pronouns are resolved. One well-known account of discourse processing with implications on pronominal resolution is Centering Theory, which, for two-sentence discourses such as the ones investigated in this chapter, predicts that pronouns prefer to have antecedents in subject position (Brennan et al. (1987); Grosz et al. (1995); cf. also Wundt (1911)). That prediction is not based on informational but on attentional structure of discourse (cf. Section 2.2 for this distinction). In support of Centering Theory, Gordon et al. (1993) found that there is a preference to use pronouns to refer to entities in subject position, but not for entities in object position. Consider the sentences in (43):[20]

[20] We follow the processing literature in focusing on the interpretation of unstressed pronouns. See Akmajian & Jackendoff (1970) among others for a discussion of the interpretation of stressed pronouns.

(43)

(43a) Fiona complimented Craig, and she congratulated James.

(43b) Fiona complimented Craig, and he congratulated James.

Intuitively, (43a) is easier to process than (43b). Centering explains this preference because the pronoun "she" in (43a) refers back to the subject of the preceding clause, whereas the pronoun "he" in (43b) refers back to the object of the preceding clause[21]. A problem for Centering Theory is provided by the contrast in (44):

(44)

(44a) Fiona complimented Craig, and James congratulated her.

(44b) Fiona complimented Craig, and James congratulated him.

Contrary to Centering Theory's subject preference prediction, Chambers & Smyth (1998) found in a self-paced reading experiment that sentences like (44b) were read faster than sentences like (44a). This pattern of results motivates the Parallel Preference account (Chambers & Smyth (1998); see Lappin & Leass (1994), for a combination of Centering Theory and Parallel Preference). Under the parallel preference account (Smyth (1994)), pronouns are argued to prefer antecedents in a parallel position when the pronoun- and the antecedent-containing sentence have the following properties: (a) both sentences have the same global constituent structure, (b) the thematic roles of the verbs in both sentences concur. When these conditions are met, subject pronouns should prefer subject antecedents, and object pronouns should prefer object antecedents. This is the case in (43) and (44) above. In (43), people prefer the preceding clause's subject as the

[21] Notice that Brennan et al. (1987)'s version of Centering Theory has a filter that would just solve pronouns in examples like (43) by gender match (there is no gender ambiguity in (43)). However, we used gender-disambiguated examples simply because we think that they better illustrate the intuitions about pronoun resolution preferences that we wanted to illustrate, especially about the non-preferred readings. In order for Centering Theory to make predictions, both possible referents would have to agree with the pronoun in gender, number (and case in some languages).

referent for the subject pronoun, whereas in (44) people prefer the preceding clause's object as the referent for the object pronoun.

Although a parallel preference account can explain the preferences in (43) and (44), it does not explain the preferences in (45), from Winograd (1972):

(45)

(45a) The city council denied the demonstrators the permit because they advocated violence.

(45b) The city council denied the demonstrators the permit because they feared violence.

In sentence (45a) the pronoun "they" refers to "the demonstrators", whereas in sentence (45b) it refers to "the city council". Neither sentence seems particularly difficult to process. Notice, however, that both Centering Theory and Parallel Preference predict a preference for "they" to refer to the subject, "the city council" – Centering Theory because the subject antecedent is the "forward-looking center" in (45), and Parallel Preference because it predicts a preference for an antecedent in a parallel position. Examples like (45) motivate causal-inference-based accounts of pronoun processing (Hobbs (1979); Hobbs et al. (1993); Kehler (2002)). According to such accounts, "they" refers to "the demonstrators" in sentence (45a) because advocating violence is assumed to be a good reason for being denied a permit. In sentence (45b) "they" refers to "the city council" because fearing violence by demonstrators is a good reason for denying a permit to these demonstrators.

Experimental evidence relevant to causal-inference-based accounts of pronominal resolution is provided by Ehrlich (1980), who used an off-line questionnaire to investigate people's preferred pronoun resolution. Ehrlich found that pronoun resolution is only driven by causal inferences (cf. Caramazza et al. (1977); Stewart et al. (2000)) when the clauses containing pronoun and antecedent respectively are in a causal relation. When there is no such causal relation, Ehrlich found that people prefer antecedents in topic or subject position (cf. Centering Theory, Grosz et al. (1995)).

Although causal-based strategies can explain the effects in (45), they do not explain the patterns in (43) and (44), because there is no causal connection between the two clauses in each of these sentences. Furthermore, resorting to a topic-based strategy like Centering Theory as suggested by Ehrlich (1980) makes the right prediction for (43), but not for (44), where the pronoun with an object antecedent is easier to process.

Based on Hobbs (1979), Kehler (2002) provides a hypothesis that aims to explain all of these patterns of pronoun resolution. Instead of arguing for pronoun-specific processing mechanisms, Kehler (2002), like Hobbs (1979), argues that pronoun resolution is a byproduct of establishing coherence. Kehler (2002) extends Hobbs (1979)'s key insight that the establishment of coherence guides pronoun resolution and vice versa, noting that discourse coherence and pronoun resolution mutually constrain each other: Pronoun resolution guides coherence, but coherence also guides pronoun resolution. Thus he hypothesizes that how a pronoun is resolved may depend on the coherence relation between the clauses.

Two classes of coherence relations that are particularly relevant to the examples that have been discussed in the pronoun resolution literature are *cause-effect* and *resemblance.* A *cause-effect* relation holds between two clauses if a plausible causal relation can be inferred to hold between the events described by the two clauses. The event described by one clause is the cause for the event described by the other clause, as in (45a). Because the demonstrators advocated violence, the city council denied them a permit to demonstrate. Kehler (2002) argues that the pronoun is interpreted such that a plausible *cause-effect* relation between the two clauses can be established. Pairing "they" with "the demonstrators" provides a more plausible interpretation for (3a) than pairing "they" with "the city council". A similar analysis applies to the pronoun resolution of "they" in (45b).

The *resemblance* discourse relation is relevant to explaining the pattern of preferences in (43) and (44). A *resemblance* relation holds between two clauses if the events described by the two clauses are in a *similarity* or in a *contrast* relation, as in the following examples from Kehler (2002):

(46) Resemblance, Similarity

(46a) Gephardt organized rallies for Gore,

(46b) and Daschle distributed pamphlets for him.

(47) Resemblance, Contrast

(47a) Gephardt supported Gore,

(47b) but Armey opposed him.

Kehler (2002) hypothesizes that the first step in establishing a *resemblance* relation between clauses is to find parallel corresponding entities and events. Then these entities and events are put into *similarity* or *contrast* relations. For example, in sentence (46a), "organized rallies" is parallel and similar to "distributed pamphlets" in sentence (46b) (both predicates describe actions of supporting a political candidate), and "Dick Gephardt" in sentence (46a) is parallel and similar to "Tom Daschle" in sentence (46b) (both are American politicians that are similar in that they support Al Gore). Then, Kehler (2002) argues, the pronoun "him" in sentence (46a) is paired with its parallel preceding element, "Gore", in sentence (46b). In sentence (47a), "supported" is parallel and in contrast to "opposed" in sentence (47b). "Gephardt" in sentence (47a) is parallel and in contrast to "Armey" in sentence (47b) (both are politicians that are in contrast in that one of them supports Gore and the other one opposes him). Then, as in example (46), the pronoun "him" in sentence (47a) is paired with its parallel preceding element, "Gore", in sentence (47b). Thus, in both examples (46) and (47), the pronoun is bound to its antecedent during the establishment of a *resemblance* coherence relation, when parallel entities are matched[22].

[22] Notice that the pronoun in (46) could also be resolved using general inference mechanisms: someone organized rallies for Gore, suggesting that Gore is a political candidate (and not Gephardt). Daschle distributed pamphlets for someone; usually pamphlets are distributed by someone who is not a political candidate (i.e. Daschle) for someone who is a political candidate. Since Gore is the most plausible political candidate (and not Gephardt or Daschle), and "him" should refer to a political candidate (because pamphlets are usually distributed for political candidates), it is most likely that "him" refers to "Gore". However, such general inferences are not sufficient for resolving the pronoun in (47): it is not clear why the fact that Gephardt supported Gore should make it more plausible for Armey to oppose Gore than to oppose

The *resemblance* relation is the most plausible coherence relation between each of the clauses in the sentences in (43) and (44). In particular, the use of the similar verbs "complimented" and "congratulated" in the absence of any other cues induces a resemblance-similarity relation between each pair of clauses. Kehler (2002)'s theory then predicts that a parallel preference strategy would be in effect under the resemblance relation, which has been observed experimentally in such sentences (Chambers & Smyth (1998)).

A strong prediction of Kehler (2002)'s theory is that pronoun resolution preferences can be altered depending on the coherence relation between clauses. The experiments presented here test this prediction directly.

3.2 Experiment 1

Experiment 1 describes an on-line self-paced reading experiment to test the different predictions of the pronoun processing accounts discussed above.

3.2.1 Method

3.2.1.1 Participants

Forty participants from the MIT community were paid for their participation. All were native speakers of English and were naive as to the purpose of the study.

3.2.1.2 Materials

Twenty sets of sentences were constructed, each with four conditions in a 2x2 design: coherence relation (*resemblance*, *cause-effect*) x parallel reference (parallel, nonparallel). An example item is presented in (48):

Gephardt (perhaps particularly since opposing Gephardt could also be interpreted as indirectly opposing Gore – Gephardt may be opposed because of his support for Gore).

(48)

(48a) Resemblance, Parallel Reference
Fiona complimented Craig and similarly James congratulated him after the match but nobody took any notice.

(48b) Resemblance, Nonparallel Reference
Fiona complimented Craig and similarly James congratulated her after the match but nobody took any notice.

(48c) Cause-Effect, Parallel Reference
Fiona defeated Craig and so James congratulated him after the match but nobody took any notice.

(48d) Cause-Effect, Nonparallel Reference
Fiona defeated Craig and so James congratulated her after the match but nobody took any notice.

Each sentence consisted of three clauses. The second clause was the target clause which consisted of the same words across the coherence manipulation. We manipulated the coherence relation between *resemblance* and *cause-effect* by making two changes to the items: (1) by using different connectives between the clauses ("and similarly" vs. "and so"), (2) by using a different verb in the first clause. For *resemblance*, the verbs in the two clauses were semantically similar according to the WordNet lexical database (Fellbaum (2001)), e.g., "compliment" and "congratulate" in (48). For the *cause-effect* conditions, the verb of the first clause in the cause-effect condition was chosen so that there was a plausible causal relation between the two clauses such that the object pronoun referred to the subject of the first clause, e.g., "defeat" and "congratulate" in (48). The first clause verb in the *cause-effect* conditions always differed from the first clause verb in the *resemblance* conditions. The remainder of the sentences consisted of a prepositional phrase and a third clause. This portion of the items was the same across the four conditions. Overall, the only differences between the *resemblance* and *cause-effect* conditions were the verbs of the first clause and the connectives relating the two clauses.

Notice that this experiment did not explore the relative contribution of different coherence cues to changing pronoun interpretation preferences. This does not diminish the point of the current design, which is simply to show that changing the coherence relation – by using one or more cues – may alter pronoun interpretation preferences.

The target sentences were combined with 76 fillers of various types in four lists balancing all factors in a Latin Square design. Appendix A provides a complete list of the stimuli. The stimuli were pseudo-randomized separately for each participant, so that at least one filler item intervened between two targets.

3.2.1.3 Procedure

The task was self-paced word-by-word reading with a moving window display (Just et al. (1982)) using Linux computers running software developed in our lab. Each trial began with a series of dashes marking the length and position of the words in the sentences, printed approximately a third of the way down the screen. Participants pressed the spacebar to reveal each word of the sentence. As each new word appeared, the preceding word disappeared. The amount of time the participant spent reading each word was recorded as the time between key-presses. After the final word of each item, a question appeared which asked about information contained in the sentence (e.g. "Did James congratulate Fiona?"). Participants pressed one of two keys to respond "yes" or "no." After an incorrect answer, the word "INCORRECT" flashed briefly on the screen. No feedback was given for correct responses. Participants were asked to read sentences at a natural rate and to be sure that they understood what they read. They were told to answer the questions as quickly and accurately as they could and to take wrong answers as an indication to read more carefully.

Before the main experiment, a short list of practice items and questions was presented in order to familiarize the participant with the task. A session averaged 25 minutes.

3.2.2 Predictions

The predictions are made in terms of reading times on the pronoun plus the next word, because in self-paced reading, effects often spill over to the next word (Sanford & Garrod (1989)). Faster reading times are assumed to reflect easier processing of the pronoun.

Centering Theory predicts that pronouns referring to antecedents in subject position should always be read faster. Thus, the pronouns in sentences (48b) and (48d) should be read faster than those in sentences (48a) and (48c).

Parallel Preference makes the opposite prediction for sentences (48a) and (48b). Because the pronouns in the experimental items are in object position, Parallel Preference predicts that pronouns referring to antecedents in (parallel) object position should be read faster. Thus, the pronouns in sentence (48a) should be read faster than those in sentence (48b). Parallel Preference does not apply to sentences (48c) and (48d), because these sentences do not meet Smyth (1994)'s criteria for parallelism.

Causal-inference-based accounts do not apply to sentences (48a) and (48b), because the events described by the clauses are not causally related. Causal-inference accounts predict that the pronoun in sentence (48d) should be read faster than in (48c), because a causal inference to resolve the pronoun is much easier to establish in (48d); in (48c), it is hard to see why James should congratulate Craig, because he lost the match. Ehrlich (1980)'s proposal that a topic-based strategy applies when there is no causal relation predicts that the pronoun in (48b) should be read faster than the pronoun in (48a).

Kehler (2002)'s coherence-based theory predicts that the cues in sentences (48a) and (48b) will indicate a *resemblance* relation between the clauses, so that a parallel preference strategy will be in effect. Thus the pronoun in sentence (48a) should be read faster than the one in sentence (48b). Kehler (2002)'s account furthermore predicts that the cues in sentences (48c) and (48d) will indicate a *cause-effect* relation between the clauses, with the consequence that the pronoun in sentence (48d) should be read faster than the pronoun in sentence (48c) because of the more plausible causal inference for sentence (48d). Thus Kehler (2002)'s account predicts an interaction between the coherence relations and the pronominal reference.

3.2.3 Results

Pronoun Reference	**Coherence Relation**	
	Resemblance	Cause-Effect
Parallel	86	80.5
Nonparallel	79.5	85

Table 13. Question answering performance in percent correct.

Table 13 shows the question answering performance for the experiment. A 2x2 ANOVA, coherence relation (*resemblance*, *cause-effect*) by reference (parallel, nonparallel), revealed an interaction by subject (F1(1,39) = 8.150, MSe = 1210, p < .01; F2(1,19) = 3.385, MSe = 605, p = .08). Pairwise comparisons by subject showed that under resemblance, question answering performance was better under parallel than under nonparallel reference (F1(1,39) = 5.354, MSe = 845, p < .05). There was no significant difference under Cause-Effect (F1(1,39) = 2.395, MSe = 405, p = .13).

Only reading times for items for which the comprehension question was answered correctly were analyzed. Reading times beyond 3 SD from the mean for a given condition and position were excluded from the analysis. This affected 2.79% of the data. Mean word-by-word reading times by subject are shown in Figure 31.

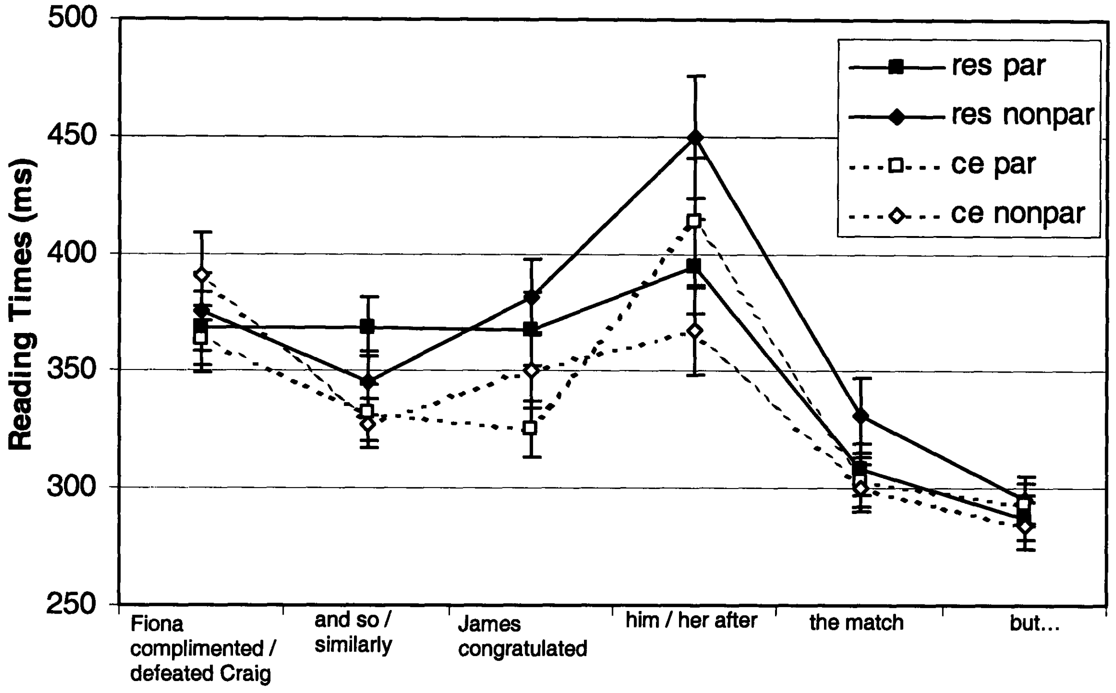


Figure 31. Plot of the reading times.

A 2x2 ANOVA, coherence relation by reference, was computed for a region including the pronoun and the following word. It showed a significant interaction of coherence relation and reference (F1(1,39) = 14.669, MSe = 103997, p < .001; F2(1,19) = 13.398, MSe = 67545, p < .005). There was also a main effect of coherence relation – *cause-effect* items were read faster than *resemblance* items (F1(1,39) = 4.431, MSe = 40563, p < .05; F2(1,19) = 3.898, MSe = 22222, p = .06). For the region containing the pronoun and the region before that region, there was a significant three-way-interaction of coherence relation, reference, and region (F1(1,39) = 12.111, MSe = 64630, p < .005; F2(1,19) = 20.126, MSe = 44344, p < .0005). There were no other significant effects.

Pairwise comparisons showed that under *resemblance*, parallel was read faster than nonparallel (F1(1,39) = 7.849, MSe = 60866, p < .01; F2(1,19) = 5.785, MSe = 40196, p < .05). Under *cause-effect*, nonparallel was read faster than parallel (F1(1,39) = 4.822, MSe = 43829, p < .05; F2(1,19) = 5.785, MSe = 27907, p < .05).

3.2.4 Discussion

The results of Experiment 1 showed that under a *resemblance* discourse relation, pronouns with an antecedent in parallel object position were read faster than pronouns with an antecedent in subject position. This is predicted by the Parallel Preference account as well as by Kehler (2002)'s account. By contrast, Centering Theory and Ehrlich (1980)'s account would have predicted a subject antecedent preference. Causal-inference-based accounts make no prediction for pronoun preferences in the absence of causal relations between the clauses containing pronoun and antecedent.

Under the *cause-effect* discourse relations in our items, pronouns referring to a subject antecedent were read faster. This is predicted by causal-inference-based accounts as well as Kehler (2002)'s account, but not predicted by the Parallel Preference account. Centering Theory does predict this preference, but not as a part of a causal inference process.

To summarize, the only account that makes the correct predictions for all conditions is Kehler (2002)'s. It predicts different preferences in pronoun resolution, depending on the coherence relation between the clauses containing the pronoun and the antecedent.

3.3 Experiment 2

Experiment 2 describes the results of a corpus study that tested the predictions of the pronoun processing accounts described in Section 3.1. Whereas Experiment 1 tested on-line language comprehension preferences, Experiment 2 tested off-line language production preferences.

3.3.1 Method

3.3.1.1 Materials

We extracted materials from the Wall Street Journal and the Brown corpus (Marcus et al. (1994)). The materials we used were clauses containing pronouns and clauses containing antecedents of these pronouns. The next section describes how we collected the materials.

3.3.1.2 Procedure

We used the following procedure to collect materials and determine frequencies of interest:

1. Extract all sentences containing non-reflexive pronouns from both the Wall Street Journal and the Brown corpus
2. From the materials collected in Step 1:
    a. In order to collect Cause-Effect materials: extract examples containing "because" with the pronoun in the second clause (i.e. the clause after the word "because")
    b. In order to collect Resemblance materials:
        i. extract examples containing "and" as a sentential conjunction with a pronoun in the second clause (i.e. the clause after the word "and")
        ii. from these materials, extract those where the clauses conjoined by "and" are in a Resemblance (i.e. Similarity or Contrast) relation. This relation was determined by a human annotator.
3. Determine pronoun antecedents
4. Extract grammatical roles for pronouns and antecedents
5. Keep only those materials that had the structure "S V O <and | because> S V O", and where the pronouns and the antecedents were in either direct subject or direct object position

Using this procedure, we extracted 410 pairs of pronouns and antecedents from the Wall Street Journal corpus, and 470 pairs of pronouns and antecedents from the Brown corpus.

Below are two examples from the Wall Street Journal, one for each type of coherence relation. The pronouns and their antecedents are in boldface. In (49), an object pronoun refers to an antecedent in object position. In (50), a subject pronoun refers to an object antecedent.

(49) Resemblance

The Exchequer Nigel Lawson's resignation slapped **the market**, and Wall Street's rapid selloff knocked **it** down.

(50) Cause-Effect

They shredded **the document** simply because **it** contained financial information.

3.3.2 Predictions

The predictions are made in terms of frequencies of pronouns and antecedents in subject and object position respectively. Higher frequencies of a type are assumed to reflect a preference for producing that type. Thus, while Experiment 1 tested preferences in language comprehension, Experiment 2 tested preferences in language production.

Centering Theory predicts that there should always be more pronoun antecedents in subject position than pronoun antecedents in object position. It does not make predictions about the frequency of pronouns in subject vs. pronouns in object positions.

Parallel Preference predicts that there should be more subject pronouns with subject antecedents than object antecedents. Furthermore, it predicts that there should be more object pronouns with object antecedents than subject antecedents.

Causal-inference-based accounts do not apply to Experiment 2 because in that experiment we tested frequency distributions over structural types (i.e. pronouns or antecedents in subject or object position). Causal-inference-based accounts would predict higher frequencies of types that reflect more plausible causal inferences. However, the question of whether a causal inference is plausible or not is orthogonal to the question of grammatical function (at least, we are not aware of a mapping of plausible causal inferences to grammatical functions).

Kehler (2002)'s coherence-based theory predicts that under Resemblance there should be more subject pronouns with subject antecedents than object antecedents. Furthermore, it predicts that under Resemblance there should be more object pronouns with object antecedents than subject antecedents. Thus, Kehler (2002)'s prediction for Resemblance are the same as the predictions made by Parallel Preference. By contrast, Kehler (2002)'s theory makes no predictions for Cause-Effect, just like causal-inference-based accounts: the question of plausibility in causal inferences is orthogonal to the question of grammatical function.

3.3.3 Results

The results of the counts are shown in Figure 32 through Figure 35.

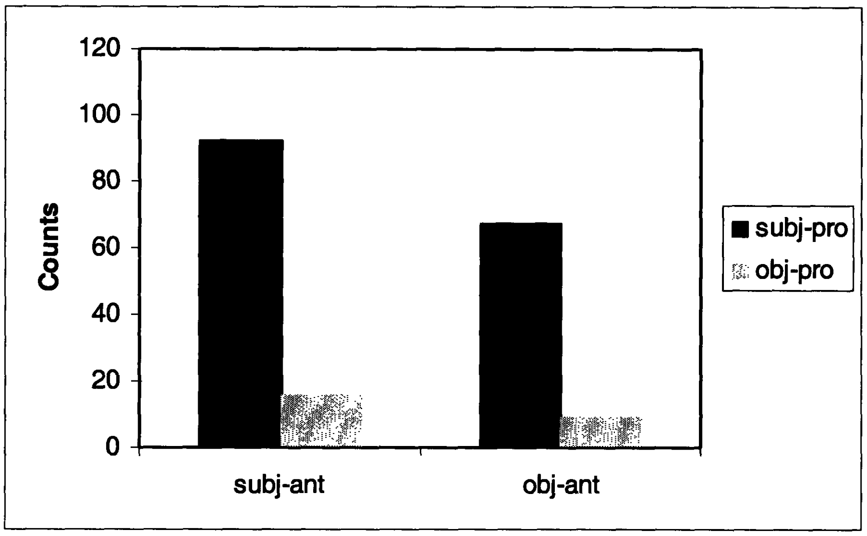


Figure 32. Cause-Effect data from the Wall Street Journal corpus.

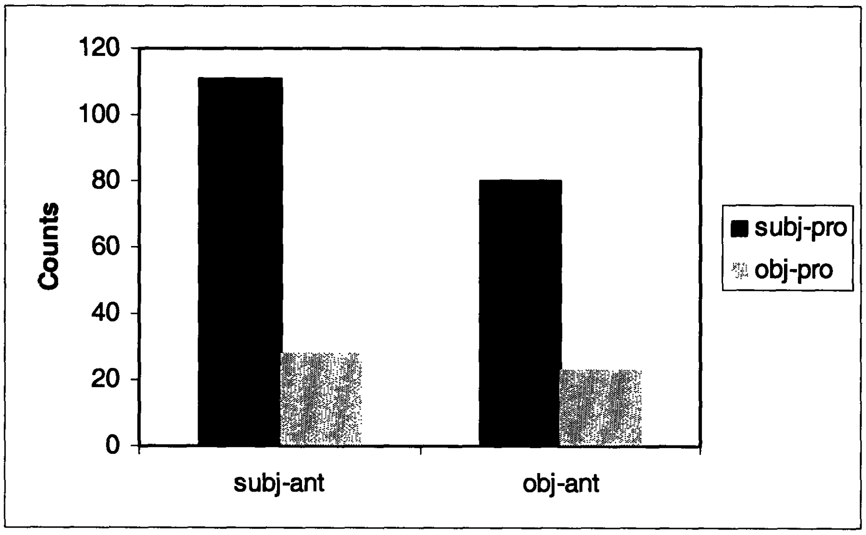


Figure 33. Cause-Effect data from the Brown corpus.

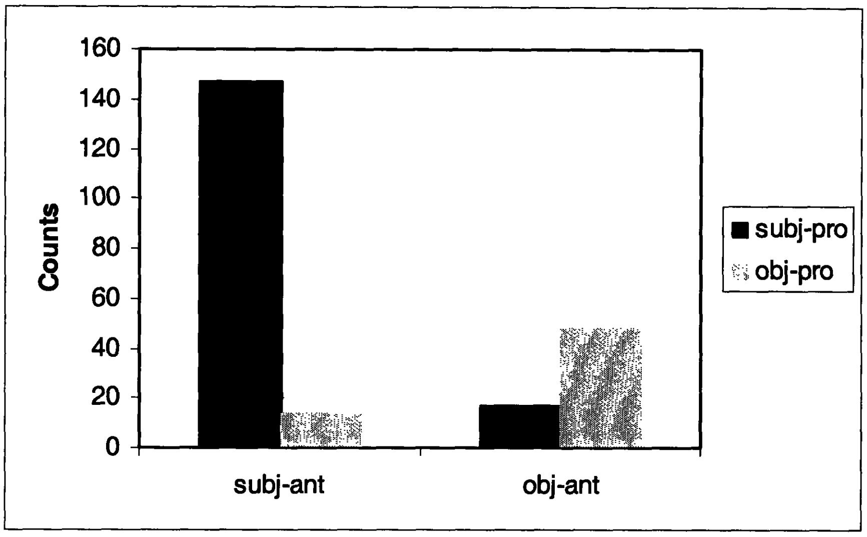


Figure 34. Resemblance data from the Wall Street Journal corpus.

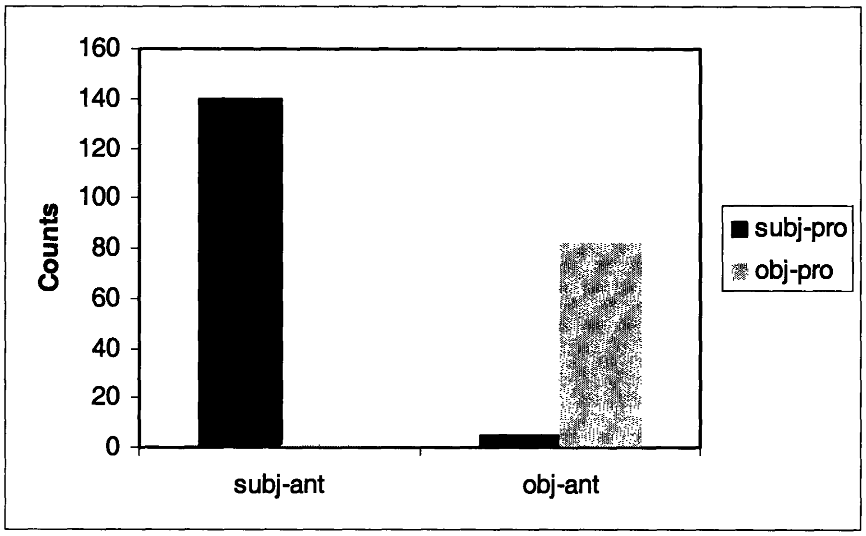


Figure 35. Resemblance data from the Brown corpus.

In order to test differences in the distributions for significance, we conducted a series of Chi-square tests. Below are the results of these tests:

- Wall Street Journal corpus
    - Cause-Effect
        - More pronouns in subject than in object position ($\chi^2(1) = 96.14$, $p = 0.0000001$)
        - More antecedents in subject than in object position ($\chi^2(1) = 5.22$, $p = 0.02$)
        - No significant interaction ($\chi^2(1) < 1$)
    - Resemblance
        - More pronouns in subject than in object position ($\chi^2(1) = 45.14$, $p = 0.0000001$)
        - More antecedents in subject than in object position ($\chi^2(1) = 38.27$, $p = 0.0000001$)
        - Significant interaction ($\chi^2(1) = 99.31$, $p = 0.0000001$)
- Brown corpus
    - Cause-Effect
        - More pronouns in subject than in object position ($\chi^2(1) = 79.84$, $p = 0.0000001$)
        - More antecedents in subject than in object position ($\chi^2(1) = 5.06$, $p = 0.02$)
        - No significant interaction ($\chi^2(1) < 1$)
    - Resemblance
        - More pronouns in subject than in object position ($\chi^2(1) = 16.32$, $p = 0.000053$)
        - More antecedents in subject than in object position ($\chi^2(1) = 12.32$, $p = 0.000448$)
        - Significant interaction ($\chi^2(1) = 199.33$, $p = 0.0000001$)

3.3.4 Discussion

The results of Experiment 2 showed that under a Resemblance discourse relation, there are more subject pronouns with subject than object antecedents, and more object pronouns with object than subject antecedents. This is predicted by the Parallel

Preference account as well as by Kehler (2002)'s account. By contrast, Centering Theory and Ehrlich (1980)'s account would have predicted there to be more subject antecedents than object antecedents, independent of pronoun position. Causal-inference-based accounts make no prediction for pronoun and antecedent grammatical function frequencies in the absence of causal relations between the clauses containing pronoun and antecedent.

Under a Cause-Effect discourse relation, there were overall more subject antecedents and more subject pronouns. This is not predicted by the Parallel Preference account. Causal-inference-based accounts do not make grammatical function frequency predictions; neither does Kehler (2002)'s account under a Cause-Effect relation. Centering Theory does predict more subject antecedents overall, but not selectively for Cause-Effect coherence relations.

In Experiment 2, there were overall more subject pronouns and antecedents than object pronouns and antecedents. We do not yet know the reason for this, and this question remains to be addressed in future research.

In sum, Kehler (2002)'s account is the only one that makes correct predictions (or at least not any wrong predictions) for all conditions. Importantly, it predicts the observed difference in frequency distributions between Resemblance and Cause-Effect coherence relations.

3.4 Conclusions

The results from the experiment reported here support the idea that the preferences observed in pronoun processing depend on the coherence relation between the clause containing the pronoun and the clause containing the antecedent (Kehler (2002)). However, this is not to say that other factors such as focusing attention on specific discourse elements (cf. Grosz et al. (1995); Wundt (1911)) play no role in pronoun processing. In fact, Kehler (2002) points out that in narratives, shifting attention to different discourse entities is an important factor in pronoun processing preferences. He argues that under such circumstances the observed preferences may be more like predicted by accounts such as Centering Theory. Notice, however, that of the accounts considered here, Kehler (2002)'s is the only one that predicts all observed preferences not as a result of the operations of pronoun-specific mechanisms, but as a byproduct of more general cognitive mechanisms and their interaction – establishing coherence and focusing attention.

4 Coherence structure and discourse segment rankings

4.1 Introduction

Automatic generation of text summaries is a natural language engineering application that has received considerable interest, particularly due to the ever-increasing volume of text information available through the internet. The task of a human generating a summary generally involves three subtasks (Brandow et al. (1995); Mitra et al. (1997)): (1) understanding a text; (2) ranking text pieces (sentences, paragraphs, phrases, etc.) for importance; (3) generating a new text (the summary). Like most approaches to summarization, we are concerned with the second subtask (e.g. Carlson et al. (2001); Goldstein et al. (1999); Gong & Liu (2001); Jing et al. (1998); Luhn (1958); Mitra et al. (1997); Sparck-Jones & Sakai (2001); Zechner (1996)). Furthermore, we are concerned with obtaining generic rather than query-relevant importance rankings (cf. Goldstein et al. (1999), Radev et al. (2002) for that distinction).

We evaluated different approaches to sentence ranking against human sentence rankings. To obtain human sentence rankings, we asked people to read 15 texts from the Wall Street Journal on a wide variety of topics (e.g. economics, foreign and domestic affairs, political commentaries). For each of the sentences in the text, they provided a ranking of how important that sentence is with respect to the content of the text, on an integer scale from 1 (not important) to 7 (very important).

The approaches we evaluated are a simple paragraph-based approach that serves as a baseline, two word-based algorithms, and two coherence-based approaches[23]. We furthermore evaluated the MSWord summarizer.

[23] We did not use any machine learning techniques to boost performance of the algorithms we tested. Therefore performance of the algorithms tested here will almost certainly be below the level of performance that could be reached if we had augmented the algorithms with such techniques (e.g. Carlson et al. (2001)). However, we think that a comparison between 'bare-bones' algorithms is viable because it allows to see how performance differs due to different basic approaches to sentence ranking, and not due to

4.2 Approaches to sentence ranking

Sentences at the beginning of a paragraph are usually more important than sentences that are further down in a paragraph, due in part to the way people are instructed to write. Therefore, probably the simplest approach conceivable to sentence ranking is to choose the first sentences of each paragraph as important, and the other sentences as not important. We included this approach merely as a simple baseline.

4.2.1 Word-based approaches

Word-based approaches to summarization are based on the idea that discourse segments are important if they contain "important" words. Different approaches have different definitions of what an important word is. For example, Luhn (1958), in a classic approach to summarization, argues that sentences are more important if they contain many significant words. Significant words are words that are not in some predefined stoplist of words with high overall corpus frequency[24]. Once significant words are marked in a text, clusters of significant words are formed. A cluster has to start and end with a significant word, and fewer than n insignificant words must separate any two significant words (we chose $n = 3$, cf. Luhn (1958)). Then, the weight of each cluster is calculated by dividing the square of the number of significant words in the cluster by the total number of words in the cluster. Sentences can contain multiple clusters. In order to compute the weight of a sentence, the weights of all clusters in that sentence are added. The higher the weight of a sentence, the higher is its ranking.

A more recent and frequently used word-based method used for text piece ranking is *tf.idf* (e.g. Manning & Schuetze (2000); Salton & Buckley (1988); Sparck-Jones & Sakai (2001); Zechner (1996)). The *tf.idf* measure relates the frequency of words in a text piece, in the text, and in a collection of texts respectively. The intuition behind *tf.idf* is to give more weight to sentences that contain terms with high frequency in a document

potentially different effects of different machine learning algorithms on different basic approaches to sentence ranking. In future research we plan to address the impact of machine learning on the algorithms tested here.

[24] Instead of stoplists, *tf.idf* values have also been used to determine significant words (e.g. Buyukkokten et al. (2001)).

but low frequency in a reference corpus. Figure 36 shows a formula for calculating *tf.idf*, where ds_{ij} is the *tf.idf* weight of sentence *i* in document *j*, n_{si} is the number of words in sentence *i*, *k* is the *k*th word in sentence *i*, tf_{jk} is the frequency of word *k* in document *j*, n_d is the number of documents in the reference corpus, and df_k is the number of documents in the reference corpus in which word *k* appears.

$$ds_{ij} = \sum_{k=1}^{n_{si}} tf_{jk} \cdot \log\left(\frac{n_d}{df_k}\right)$$

Figure 36. Formula for calculating *tf.idf* (Salton & Buckley (1988)).

We compared both Luhn (1958)'s measure and *tf.idf* scores to human rankings of sentence importance. We will show that both methods performed remarkably well, although one coherence-based method performed better.

4.2.2 Coherence-based approaches

The sentence ranking methods introduced in the two previous sections are solely based on layout or on properties of word distributions in sentences, texts, and document collections. Other approaches to sentence ranking are based on the informational structure of texts. With informational structure, we mean the set of informational relations that hold between sentences in a text. This set can be represented in a graph, where the nodes represent sentences, and labeled directed arcs represent informational relations that hold between the sentences (cf. Hobbs (1985)). Often, informational structures of texts have been represented as trees (e.g. Carlson et al. (2001), Corston-Oliver (1998), Mann & Thompson (1988), Ono et al. (1994)). We will present one coherence-based approach that assumes trees as a data structure for representing discourse structure, and one approach that assumes less constrained graphs. As we will show, the approach based on less constrained graphs performs better than the tree-based approach when compared to human sentence rankings.

4.3 Coherence-based sentence ranking revisited

This section will discuss in more detail the data structures we used to represent discourse structure, as well as the algorithms used to calculate sentence importance, based on discourse structures.

4.3.1 Representing coherence structures

4.3.1.1 Discourse segments

As pointed out in Section 2.3.1, discourse segments can be defined as non-overlapping spans of prosodic units (Hirschberg & Nakatani (1996)), intentional units (Grosz & Sidner (1986)), phrasal units (Lascarides & Asher (1993)), or sentences (Hobbs (1985)). We adopted a mostly clause unit-based definition of discourse segments for the coherence-based approach that assumes chain graphs (cf. Section 2.3.1). For the coherence-based approach that assumes trees, we used Marcu (2000)'s more fine-grained definition of discourse segments because we used the discourse trees from Carlson et al. (2002)'s database of coherence-annotated texts.

4.3.1.2 Kinds of coherence relations

For the coherence structure-based summarization approaches discussed in this chapter, we assume the set of coherence relations that was introduced in Section 2.3.3. As a reminder, below are examples of each coherence relation.

(51) Cause-Effect

[There was bad weather at the airport]$_a$ [and so our flight got delayed.]$_b$

(52) Violated Expectation

[The weather was nice]$_a$ [but our flight got delayed.]$_b$

(53) Condition

[If the new software works,]$_a$ [everyone will be happy.]$_b$

(54) Similarity

[There is a train on Platform A.]$_a$ [There is another train on Platform B.]$_b$

(55) Contrast

[John supported Bush]$_a$ [but Susan opposed him.]$_b$

(56) Elaboration

[A probe to Mars was launched this week.]$_a$ [The European-built 'Mars Express' is scheduled to reach Mars by late December.]$_b$

(57) Attribution

[John said that]$_a$ [the weather would be nice tomorrow.]$_b$

(58) Temporal Sequence

[Before he went to bed,]$_a$ [John took a shower.]$_b$

As pointed out in Section 2.3.3, *cause-effect*, *violated expectation*, *condition*, *elaboration*, *temporal sequence*, and *attribution* are asymmetrical or directed relations, whereas *similarity*, and *contrast* are symmetrical or undirected relations (Mann & Thompson (1988); Marcu (2000)). As also pointed out in Section 2.3.3, in the chain graph-based approach, the directions of asymmetrical or directed relations are as follows:

- cause → effect for *cause-effect*
- cause → absent effect for *violated expectation*
- condition → consequence for *condition*
- elaborating → elaborated for *elaboration*
- source → attributed for *attribution*
- sooner → later for *temporal sequence*

In the tree-based approach, the asymmetrical or directed relations are between a more important discourse segment, or a Nucleus, and a less important discourse segment, or a

Satellite (Mann & Thompson (1988)). The Nucleus is the equivalent of the arc destination, and the Satellite is the equivalent of the arc origin in the chain graph-based approach. The symmetrical or undirected relations are between two discourse elements of equal importance, or two Nuclei. Below we will explain how the difference between Satellites and Nuclei is considered in tree-based sentence rankings.

4.3.1.3 Data structures for representing discourse coherence

As mentioned above, we used two alternative representations for discourse structure, tree- and chain graph based. In order to illustrate both data structures, consider (59) as an example:

(59)

0. Susan wanted to buy some tomatoes.
1. She also tried to find some basil.
2. The basil would probably be quite expensive at this time of the year.

(constructed)

Figure 37 shows one possible tree representation of the coherence structure of (59)[25]. *Sim* represents a *similarity* relation, and *elab* an *elaboration* relation. Furthermore, nodes with a "Nuc" subscript are Nuclei, and nodes with a "Sat" subscript are Satellites.

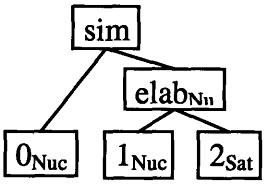


Figure 37 Coherence tree for (59).

[25] Another possible tree structure might be "(elab (sim (0, 1), 2))". What structure is chosen would probably mostly depend on how the *similarity* relation is defined: if it can only hold between what is actually in a *similarity* relation (0 and 1 are, but 2 has no corresponding parallel entities or events anywhere), then "(elab (sim (0, 1), 2))" should be chosen over "(sim (0, elab (2, 1)))".

Figure 38 shows a chain graph representation of the coherence structure of (59). Here, the heads of the arrows represent the directionality of a relation.

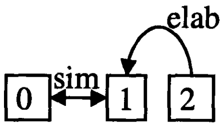


Figure 38. Coherence chain graph for (59).

4.3.2 Coherence-based sentence ranking

This section explains the algorithms for the tree- and the chain graph-based sentence ranking approach.

4.3.2.1 Tree-based approach

We used Marcu (2000)'s algorithm to determine sentence rankings based on tree discourse structures. In this algorithm, sentence salience is determined based on the tree level of a discourse segment in the coherence tree. Figure 39 shows Marcu (2000)'s algorithm, where *r(s,D,d)* is the rank of a sentence *s* in a discourse tree *D* with depth *d.* Every node in a discourse tree *D* has a promotion set *promotion(D)*, which is the union of all Nucleus children of that node. Associated with every node in a discourse tree D is also a set of parenthetical nodes *parentheticals(D)* (for example, in "Mars – half the size of Earth – is red", "half the size of earth" would be a parenthetical node in a discourse tree). Both *promotion(D)* and *parentheticals(D)* can be empty sets. Furthermore, each node has a left subtree, *lc(D)*, and a right subtree, *rc(D)*. Both *lc(D)* and *rc(D)* can also be empty.

$$r(s,D,d)=\begin{cases}0 & \textit{if } D \textit{ is } \textit{NIL},\\ d & \textit{if } s \in \textit{promotion}(D),\\ d-1 & \textit{if } s \in \textit{parentheticals}(D),\\ \max(r(s,lc(D),d-1), & \\ \quad r(s,rc(D),d-1)) & \textit{otherwise}\end{cases}$$

Figure 39. Formula for calculating coherence-tree-based sentence rank (Marcu (2000)).

We illustrate how Marcu (2000)'s algorithm works in an example shown in Figure 40 through Figure 45. The gray numbers on the right side of each figure represent inverse tree levels, which are used for determining discourse segment ranks. Like in the examples in Sections 2.4.1 and 2.4.2, solid lines represent Nucleus relations and dashed lines represent Satellite relations (cf. Mann & Thompson (1988)). Terminal nodes can only be promoted through Nucleus relations. The promotions are illustrated by the arrows in the figures. The rank for each segment or terminal node is determined by the highest level in the tree to which the node can be promoted. The segment ranks are shown in boldface below the terminal nodes. Here is how the discourse segment ranks for the tree in Figure 40 were derived:

- Discourse segment 0: Discourse segment 0 is in a Nucleus relation with a non-terminal node at inverse tree level 2, so discourse segment 0 gets "promoted" to inverse tree level 2. The non-terminal node at inverse tree level 2 is in a Nucleus relation with another non-terminal node at inverse tree level 3, so discourse segment 0 gets "promoted" to inverse tree level 3. The non-terminal node at inverse tree level 3 is in a Nucleus relation with a non-terminal node (the root node) at inverse tree level 4, so discourse segment 0 gets "promoted" to inverse tree level 4. Thus discourse segment 0 gets rank 4 by Marcu (2000)'s algorithm (cf. Figure 41).
- Discourse segment 1: Discourse segment 1 is in a Satellite relation. Therefore discourse segment 1 does not get "promoted" and gets rank 1 by Marcu (2000)'s algorithm (cf. Figure 42).

- Discourse segment 2: Discourse segment 2 is in a Satellite relation. Therefore discourse segment 2 does not get "promoted" and gets rank 2 by Marcu (2000)'s algorithm (cf. Figure 43).
- Discourse segment 3: Discourse segment 3 is in a Nucleus relation with a non-terminal node at inverse tree level 3, so discourse segment 3 gets "promoted" to inverse tree level 3. Because the non-terminal node at inverse tree level 3 is in a Satellite relation with the non-terminal node at the next-higher inverse tree level, discourse segment 3 does not get "promoted" further. Thus discourse segment 3 gets rank 3 by Marcu (2000)'s algorithm (cf. Figure 44).
- Discourse segment 4: Discourse segment 4 is in a Nucleus relation with a non-terminal node at inverse tree level 3, so discourse segment 4 gets "promoted" to inverse tree level 3. Because the non-terminal node at inverse tree level 3 is in a Satellite relation with the non-terminal node at the next-higher inverse tree level, discourse segment 4 does not get "promoted" further. Thus discourse segment 4 gets rank 3 by Marcu (2000)'s algorithm (cf. Figure 45).

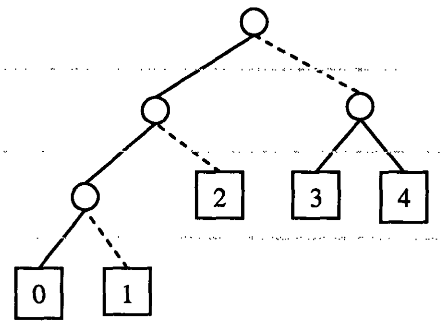


Figure 40. Example tree for illustrating Marcu (2000)'s algorithm.

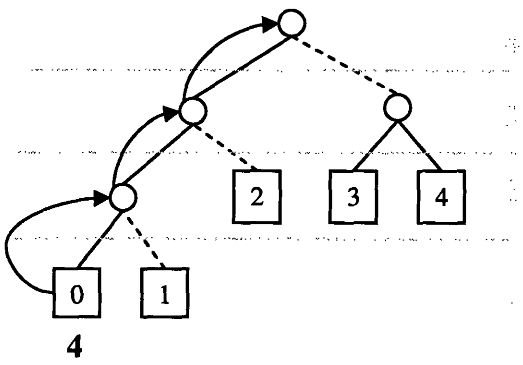


Figure 41. Determining the ranking for segment 0.

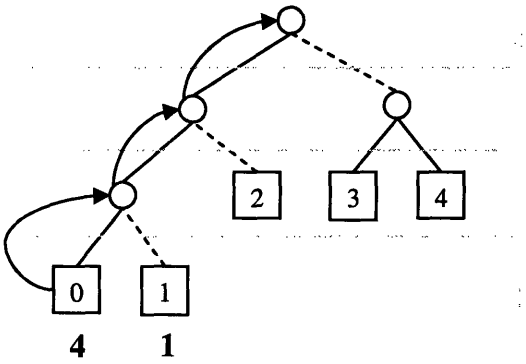


Figure 42. Determining the ranking for segment 1.

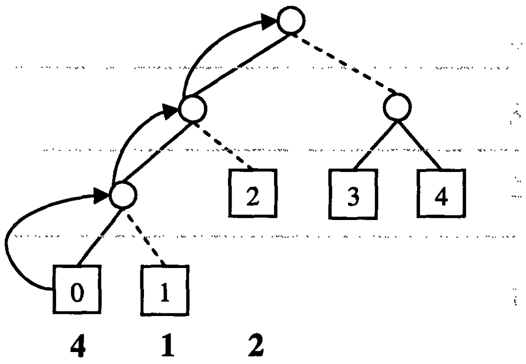


Figure 43. Determining the ranking for segment 2.

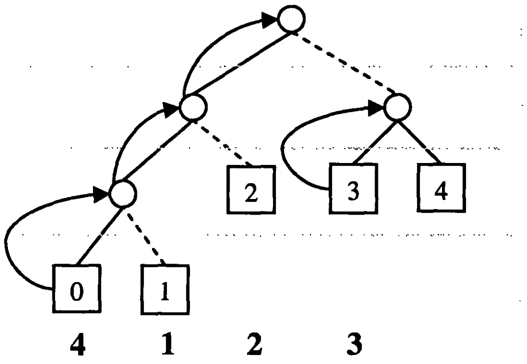


Figure 44. Determining the ranking for segment 3.

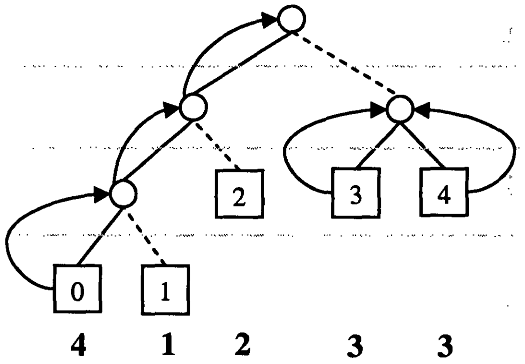


Figure 45. Determining the ranking for segment 4.

The discourse segments in Carlson et al. (2002)'s database are often sub-sentential. Therefore, we had to calculate sentence rankings from the rankings of the discourse segments that form the sentence under consideration. We did this by calculating the average ranking, the minimal ranking, and the maximal ranking of all discourse segments in a sentence. Our results showed that choosing the minimal ranking performed best, followed by the average ranking, followed by the maximal ranking (cf. Section 4.4.4).

4.3.2.2 Chain graph-based approaches

We used three different methods to determine sentence rankings for the coherence chain graphs[26]. All three methods implement the intuition that sentences are more important if other sentences relate to them (Sparck-Jones (1993)).

The first method consists of simply determining the in-degree of each node in the graph. A node represents a sentence, and the in-degree of a node represents the number of sentences that relate to that sentence.

The second method is based on the idea that a discourse segment is more important if other important discourse segments relate to it. In order to implement this idea, we used Page et al. (1998)'s *PageRank* algorithm, which is used, for example, in the Google™ search engine. Unlike just determining the in-degree of a node, *PageRank* takes into account the importance of sentences that relate to a sentence. Figure 46 shows how *PageRank* is calculated. PR_n is the *PageRank* of the current sentence, PR_{n-1} is the *PageRank* of the sentence that relates to sentence n, o_{n-1} is the out-degree of sentence $n-1$, and α is a damping parameter that is set to a value between 0 and 1. We report results for α set to 0.85 because this is a value often used in applications of *PageRank* (e.g. Ding et al. (2002); Page et al. (1998)). We also calculated *PageRanks* for α set to values between 0.05 and 0.95, in increments of 0.05; changing α did not affect performance. For calculating *PageRank*, we treated coherence graphs as directed graphs and replaced undirected arcs (those arcs that represent symmetrical or undirected coherence relations, i.e. *similarity*, *contrast*, *same*) by two directed arcs (one arc in each direction).

[26] Neither of these methods could be implemented for coherence trees since Marcu (2000)'s tree-based algorithm assumes binary branching trees. Thus, the in-degree for all non-terminal nodes is always 2.

$$PR_n = 1 - \alpha + \alpha \frac{PR_{n-1}}{O_{n-1}}$$

Figure 46. Formula for calculating PageRank (Page et al. (1998)).

The third method, *cRank* (for "coherence-rank"), is based on the idea that a discourse segment ds_0 is more important if more other discourse segments relate to it. Furthermore, the more directly these other discourse segments relate to ds_0, the more they boost the importance of ds_0. In order to determine how directly discourse segments relate to ds_0, *cRank* conducts a breadth-first search through the coherence graph, starting at ds_0 (this ranking procedure is done for every node in the coherence graph). Pseudocode for *cRank* is shown in Figure 47 (based on breadth-first search Pseudocode from Goodrich & Tamassia (2001)):

- i: counter that counts the rounds of the breadth-first search algorithm
- L_i: container for graph vertices at round i of the breadth-first search algorithm
- ds: discourse segment vertex
- $rank_{ds}$: rank of discourse segment vertex ds
- $V(G_{discourse\text{-}structure})$: set of discourse segment vertices in the graph representing the discourse structure, $G_{discourse\text{-}structure}$
- e: edge in the coherence graph
- v, w: vertices adjacent to an edge e
- α: parameter set to]0..1]
- $\|L_{i+1}\|$: number of elements in container L_{i+1}
- discovery edge: an edge that leads to a previously unvisited vertex
- cross edge: an edge that leads to a previously visited vertex

```
1.  for each discourse segment vertex $ds$ in $V(G_{discourse\text{-}structure})$ do
2.      $rank_{ds} \leftarrow 0$
3.      initialize container $L_0$ to contain vertex $ds$
4.      $i \leftarrow 0$
5.      while $L_i$ is not empty do
6.          create container $L_{i+1}$ to initially be empty
7.          for each vertex $v$ in $L_i$ do
8.              if edge $e$ is unexplored then
9.                  let $w$ be the other endpoint of $e$
10.                 if vertex $w$ is unexplored then
11.                     label $e$ as a discovery edge
12.                     insert $w$ into $L_{i+1}$
13.                 else
14.                     label $e$ as a cross edge
15.         $rank_{ds} \leftarrow rank_{ds} + \alpha^{i+1} \cdot \|L_{i+1}\|$
16.         $i \leftarrow i + 1$
```

Figure 47. Pseudocode for *cRank*, based on pseudocode for a breadth-first search algorithm in Goodrich & Tamassia (2001).

Notice that *cRank* treats coherence graphs as undirected graphs. Notice furthermore that we set the parameter α to 1, thus effectively reducing Line 15 of the algorithm to a parameter-free form:

15. $rank_{ds} \leftarrow rank_{ds} + (i+1) \cdot \|L_{i+1}\|$

The effect of setting α to a value between 0 and 1 would be that the nodes further up in a breadth-first traversal tree would add more to $rank_{ds}$ than nodes that are further down. This could be done to optimize the performance of *cRank*. However, that was not the goal here.

We illustrate *cRank* with α = 1 for the graph shown in Figure 48. Trees representing breadth-first graph traversals for each node in the graph are shown in Figure 49 through Figure 53. The gray numbers in the figures represent the inverse tree levels. Notice that it does not make a difference whether the inverse tree level for the starting node of each breadth-first graph traversal is included in the calculations of *cRank* because it would affect the rank of all nodes equally.

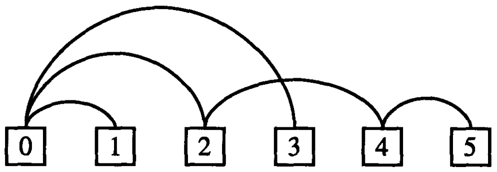


Figure 48. Example graph for illustrating *cRank*.

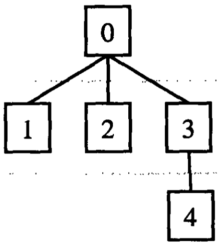


Figure 49. Breadth-first graph traversal starting at node 0.

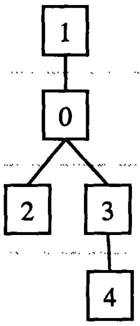


Figure 50. Breadth-first graph traversal starting at node 1.

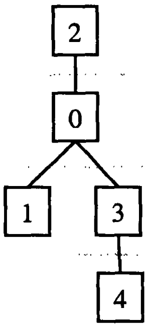


Figure 51 Breadth-first graph traversal starting at node 2.

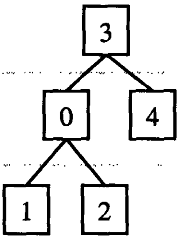


Figure 52 Breadth-first graph traversal starting at node 3.

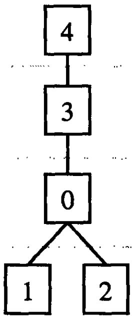


Figure 53 Breadth-first graph traversal starting at node 4.

Based on the breadth-first graph traversals shown in Figure 49 through Figure 53, the rank for each node in the graph in Figure 48 can be determined as follows (with parameter $\alpha = 1$):

Node 0:

$3 \cdot 3$ (nodes 1, 2, and 3)

$+ \quad 1 \cdot 2$ (node 4)

$+ \quad 0 \cdot 1$

$= \quad \mathbf{11;}$

Node 1:

$1 \cdot 3$ (node 0)

$+ \quad 2 \cdot 2$ (nodes 2 and 3)

$+ \quad 1 \cdot 1$ (node 4)

$= \quad \mathbf{8;}$

Node 2:

$1 \cdot 3$ (node 0)

$+ \quad 2 \cdot 2$ (nodes 1 and 3)

$+ \quad 1 \cdot 1$ (node 4)

$= \quad \mathbf{8;}$

Node 3:

$2 \cdot 3$ (nodes 0 and 4)

$+ \quad 2 \cdot 2$ (nodes 1 and 2)

$+ \quad 0 \cdot 1$

$= \quad \mathbf{10;}$

Node 4:

$1 \cdot 3$ (node 3)

$+ \quad 1 \cdot 2$ (node 0)

$+ \quad 2 \cdot 1$ (nodes 1 and 2)

$= \quad \mathbf{7;}$

Thus, by *cRank*, node 0 has rank 11, node 1 has rank 8, node 2 has rank 8, node 3 has rank 10, and node 4 has rank 7.

4.4 Experiments

In order to test algorithm performance, we compared algorithm sentence rankings to human sentence rankings. This section describes the experiments we conducted. In Experiment 1, the texts were presented with paragraph breaks; in Experiment 2, the texts were presented without paragraph breaks. This was done to control for the effect of paragraph information on human sentence rankings.

4.4.1 Materials for the coherence-based approaches

In order to test the tree-based approach, we took coherence trees for 15 texts from a database of 385 texts from the Wall Street Journal that were annotated for coherence (Carlson et al. (2002)). The database was independently annotated by six annotators. Inter-annotator agreement was determined for six pairs of two annotators each, resulting in kappa values (Carletta (1996)) ranging from 0.62 to 0.82 for the whole database (Carlson et al. (2003)). No kappa values for just the 15 texts we used were available.

For the chain graph based approach, we used coherence graphs from a database of 135 texts from the Wall Street Journal and the AP Newswire, annotated for coherence. Each text was independently annotated by two annotators. For the 15 texts we used, kappa was 0.78, for the whole database, kappa was 0.84.

4.4.2 Experiment 1: With paragraph information

15 participants from the MIT community were paid for their participation. All were native speakers of English and were naïve as to the purpose of the study (i.e. none of the subjects was familiar with theories of coherence in natural language, for example).

Participants were asked to read 15 texts from the Wall Street Journal, and, for each sentence in each text, to provide a ranking of how important that sentence is with respect to the content of the text, on an integer scale from 1 to 7 (1 = not important; 7 = very important).

The texts were selected so that there was a coherence tree annotation available in Carlson et al. (2002)'s database. Text lengths for the 15 texts we selected ranged from 130 to 901 words (5 to 47 sentences); average text length was 442 words (20 sentences), median was 368 words (16 sentences). Additionally, texts were selected so that they were about as diverse topics as possible.

The experiment was conducted in front of personal computers. Texts were presented in a web browser as one webpage per text; for some texts, participants had to scroll to see the whole text. Each sentence was presented on a new line. Paragraph breaks were indicated by empty lines; this was pointed out to the participants during the instructions for the experiment.

4.4.3 Experiment 2: Without paragraph information

The method was the same as in Experiment 1, except that texts in Experiment 2 did not include paragraph information. Each sentence was presented on a new line. None of the 15 participants who participated in Experiment 2 had participated in Experiment 1.

4.4.4 Results of the experiments

Human sentence rankings did not differ significantly between Experiment 1 and Experiment 2 for any of the 15 texts (all Fs < 1). This suggests that paragraph information does not have a big effect on human sentence rankings, at least not for the 15 texts that we examined. Figure 54 shows the results from both experiments for one text.

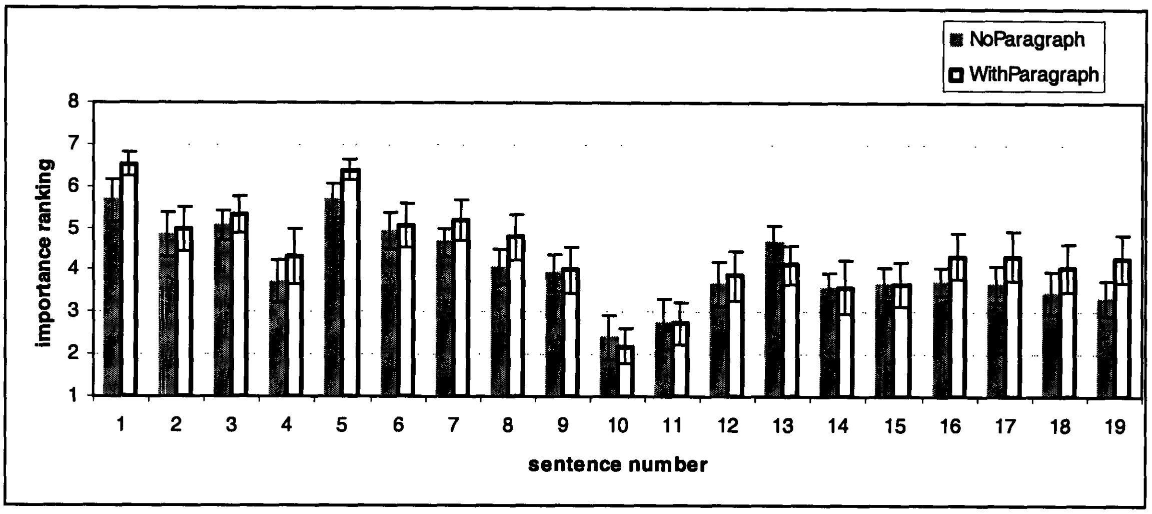


Figure 54. Human ranking results for one text (wsj_1306).

We compared human sentence rankings to different algorithmic approaches. The paragraph-based rankings do not provide scaled importance rankings but only "important" vs. "not important". Therefore, in order to compare human rankings to the paragraph-based baseline approach, we calculated point biserial correlations (cf. Bortz (1999)). We obtained significant correlations between paragraph-based rankings and human rankings only for one of the 15 texts.

All other algorithms provided scaled importance rankings. Many evaluations of scalable sentence ranking algorithms are based on precision/recall/F-scores (e.g. Carlson et al. (2001); Ono et al. (1994)). However, Jing et al. (1998) argue that such measures are inadequate because they only distinguish between hits and misses or false alarms, but do not account for a degree of agreement. For example, imagine a situation where the human ranking for a given sentence is "7" ("very important") on an integer scale ranging from 1 to 7, and Algorithm A gives the same sentence a ranking of "7" on the same scale, Algorithm B gives a ranking of "6", and Algorithm C gives a ranking of "2". Intuitively, Algorithm B, although it does not reach perfect performance, still performs better than Algorithm C. Precision/recall/F-scores do not account for that difference and would rate Algorithm A as "hit" but Algorithm B as well as Algorithm C as "miss". In order to collect performance measures that are more adequate to the evaluation of scaled importance rankings, we computed Spearman's rank correlation coefficients. The rank correlation coefficients were corrected for tied ranks because in our rankings it was possible for more than one sentence to have the same importance rank, i.e. to have tied ranks (Horn (1942); Bortz (1999)).

In addition to evaluating word-based and coherence-based algorithms, we evaluated one commercially available summarizer, the MSWord summarizer, against human sentence rankings. Our reason for including an evaluation of the MSWord summarizer was to have a more useful baseline for scalable sentence rankings than the paragraph-based approach provides.

We used Carlson et al. (2002)'s discourse trees as input to Marcu (2000)'s algorithm. That means that Marcu (2000)'s algorithm operated on sub-sentential discourse segments. By contrast, the human rankings, and all the rankings obtained from the other algorithms, were for whole sentences. Therefore we had to convert the rankings obtained from Marcu (2000)'s algorithm from rankings of sub-sentential segments into rankings of sentences. We did that in three different ways:

- *MarcuAvg*: sentence rankings were calculated as the average of the rankings of all discourse segments in that sentence
- *MarcuMin*: sentence rankings were the minimum of the rankings of all discourse segments in that sentence
- *MarcuMax*: sentence rankings were the maximum of the rankings of all discourse segments in that sentence

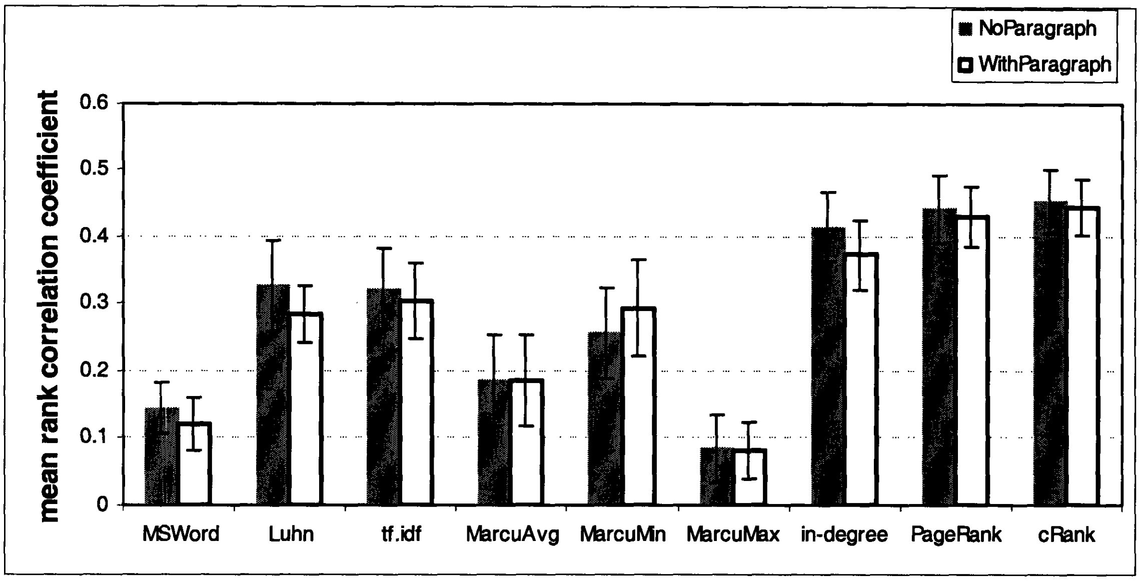


Figure 55. Average rank correlations (ρ_{avg}) of algorithm and human sentence rankings.

Figure 55 shows average rank correlations (ρ_{avg}) of each algorithm and human sentence ranking for the 15 texts. It shows that the MSWord summarizer performed numerically worse than most other algorithms, except *MarcuMax*. Figure 55 also shows that *cRank* performed numerically better than all other algorithms. Performance was significantly better than most other algorithms:

- *MSWord*
  - *NoParagraph*: F(1,28) = 24.779, p = 0.0001
  - *WithParagraph*: F(1,28) = 31.832, p = 0.0001
- *Luhn*
  - *WithParagraph*: F(1,28) = 7.326, p = 0.011
- *MarcuAvg*
  - *NoParagraph*: F(1,28) = 10.382, p = 0.003
  - *WithParagraph*: F(1,28) = 10.821, p = 0.003
- *MarcuMin*
  - *NoParagraph*: F(1,28) = 5.527, p = 0.026
- *MarcuMax*
  - *NoParagraph*: F(1,28) = 27.722, p = 0.0001

  - *WithParagraph*: F(1,28) = 37.778, p = 0.0001

cRank performed marginally better than *tf.idf*, *WithParagraph* (F(1,28) = 4.162, p = 0.051).

cRank performed not significantly better than the following algorithms:

- *Luhn*
  - *NoParagraph*: F(1,28) = 2.298, p = 0.141
- *tf.idf*
  - *NoParagraph*: F(1,28) = 2.858, p = 0.102
- *MarcuMin*
  - *WithParagraph*: F(1,28) = 3.287, p = 0.081

As mentioned above, human sentence rankings did not differ significantly between Experiment 1 and Experiment 2 for any of the 15 texts (all Fs < 1). Therefore, in order to lend more power to our statistical tests, we collapsed the data for each text for the *WithParagraph* and the *NoParagraph* condition, and treated them as one experiment. Figure 56 shows that when the data from Experiments 1 and 2 are collapsed, *cRank* performed significantly better than all other algorithms except *in-degree* and *PageRank*, as shown by two-tailed t-tests:

- *MSWord*: F(1, 58) = 57.881, p = 0.0001
- *Luhn*: F(1,58) = 8.108, p = 0.006
- *tf.idf*: F(1,58) = 7.104, p = 0.010
- *MarcuAvg*: F(1,58) = 21.928, p = 0.0001
- *MarcuMin*: F(1,58) = 8.946, p = 0.004
- *MarcuMax*: F(1,58) = 66.076, p = 0.0001
- *in-degree*: F(1,58) = 1.350, p = 0.250
- *PageRank*: F(1,58) < 1

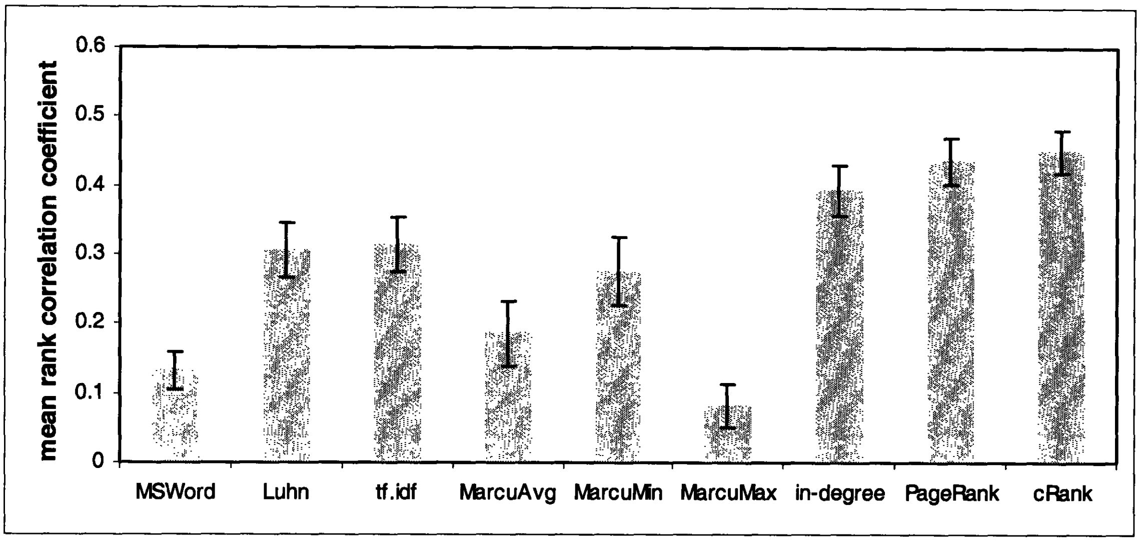


Figure 56. Average rank correlations (ρ_{avg}) of algorithm and human sentence rankings with collapsed data.

4.5 Discussion

4.5.1 *MSWord* and word-based algorithms

The results of our experiments showed that *MSWord* showed the second-worst performance of all algorithms tested; only *MarcuMax* was numerically worse, although the difference was not significant. Unfortunately we cannot determine why *MSWord* did not perform better because it is a proprietary product. It is also difficult to evaluate why the word-based methods (*Luhn* and *tf.idf*) did not perform better. What we can say, however, is that word-based methods provide a very high baseline, but improvements are possible, as the performance of the chain graph-based algorithms (*in-degree*, *PageRank*, *cRank*) showed.

4.5.2 Coherence-based algorithms that operate on trees

Our results showed that coherence-based algorithms that operate on trees (*MarcuMin*, *MarcuMax*, *MarcuAvg*) performed poorly. One possible reason for this could be that trees are not descriptively adequate or introduce biases that hurt performance on sentence ranking. In particular, one problem could be that trees cannot represent nodes with multiple parents. Consider an example of a coherence structure as shown in Figure 57. In that structure, R_1, R_2, and R_3 represent three different kinds of coherence relations. Based on any of the algorithms tested in this chapter that operate on chain graphs (*in-degree*, *PageRank*, *cRank*), Node 0 would obtain the highest rank, whereas Nodes 1, 2, and 3 would all obtain the same rank that is lower than the rank of Node 0.

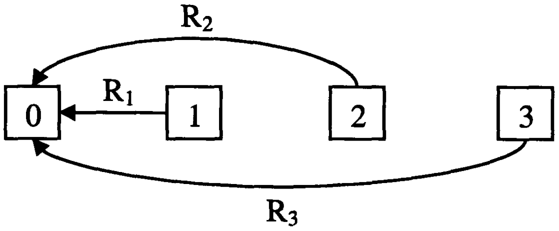


Figure 57. Example of a coherence structure that contains a node with multiple parents (node 0).

However, the nodes would be ranked differently if a tree-based representation for the structure in Figure 57 is used. The problem with a structure such as the one shown in Figure 57 is that it contains nodes with multiple parents. Figure 58 shows a tree representing that structure, where the grey numbers on the right indicate inverse tree level, solid lines in the tree indicate Nucleus relations, and dashed lines indicate Satellite relations[27]. By Marcu (2000)'s algorithm, Node 0 would obtain the highest rank (rank 4 because it gets "promoted" through Nucleus relations up to inverse tree level 4). However, unlike by chain graph-based ranking algorithms, by Marcu (2000)'s algorithm, the other nodes would each get different ranks: Node 1 would get rank 1, Node 2 would get rank 2, and Node 3 would get rank 3 (because they are in Satellite relations, these nodes do not get "promoted"). Notice that these ranks are a byproduct of the tree data structure.

[27] Notice that the point of the example is that R_1, R_2, and R_3 are different kinds of relations. Therefore a structure like "R_1 (0, R_2 (1, 2, 3))" would not work for this example.

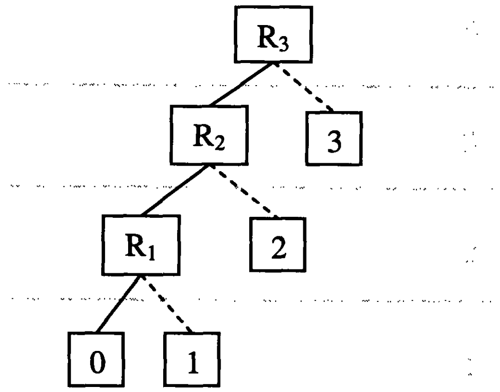


Figure 58. Tree-based representation for the coherence structure shown in Figure 57. The grey numbers on the right indicate inverse tree levels. Solid lines indicate Nucleus relations, and dashed lines indicate Satellite relations.

Our experimental text segment ranking results provide some support for this idea. An example is the discourse structure for text wsj_1148 from the Wall Street Journal corpus (Harman & Liberman (1993)), segmented into sentences (as we used them for the human sentence ranking experiments):

(60)

1. Mobil Corp. is preparing to slash the size of its work force in the U.S., possibly as soon as next month, say individuals familiar with the company's strategy.
2. The size of the cuts isn't known, but they'll be centered in the exploration and production division, which is responsible for locating oil reserves, drilling wells and pumping crude oil and natural gas.
3. Employees haven't yet been notified.
4. Sources said that meetings to discuss the staff reductions have been scheduled for Friday at Mobil offices in New Orleans and Denver.
5. This would be a second round of cuts by Mobil, which along with other oil producers and refiners reduced its work force by 15% to 20% during the mid-1980s as part of an industrywide shakeout.
6. Mobil's latest move could signal the beginning of further reductions by other oil companies in their domestic oil-producing operations.

7. In yesterday's third-quarter earnings report, the company alluded to a $40 million provision for restructuring costs involving U.S. exploration and production operations.
8. The report says that "the restructuring will take place over a two-year period and will principally involve the transfer and termination of employees in our U.S. operations."
9. A company spokesman, reached at his home last night, would only say that there will be a public announcement of the reduction program by the end of the week.
10. Most oil companies, including Mobil, have been reporting lower third-quarter earnings, largely as a result of lower earnings from chemicals as well as refining and marketing businesses.
11. Individuals familiar with Mobil's strategy say that Mobil is reducing its U.S. work force because of declining U.S. output.
12. Yesterday, Mobil said domestic exploration and production operations had a $16 million loss in the third quarter, while comparable foreign operations earned $234 million.
13. Industrywide, oil production in this country fell by 500,000 barrels a day to 7.7 million barrels in the first eight months of this year.
14. Daily output is expected to decline by at least another 500,000 barrels next year.
15. Some Mobil executives were dismayed that a reference to the cutbacks was included in the earnings report before workers were notified.
16. One Mobil executive said that the $40 million charge related to the action indicates "a substantial" number of people will be involved.
17. Some will likely be offered severance packages while others will be transferred to overseas operations.

(example wsj_1148 from the Wall Street Journal corpus, Harman & Liberman (1993))

The discourse structure for the text “wsj_1148”, when represented in a tree, is very similar to the one shown in Figure 58, and it is schematically represented in Figure 59. Figure 59 shows that Node 17, which represents the last text segment in the text, gets rank 11 by Marcu (2000)'s algorithm. This is the second-highest rank; only Node 1 is ranked higher because it is "promoted" to rank 12 through Nucleus relations[28]. However, in our human rankings (both with and without paragraph information), Node 17 is ranked lowest of all nodes.

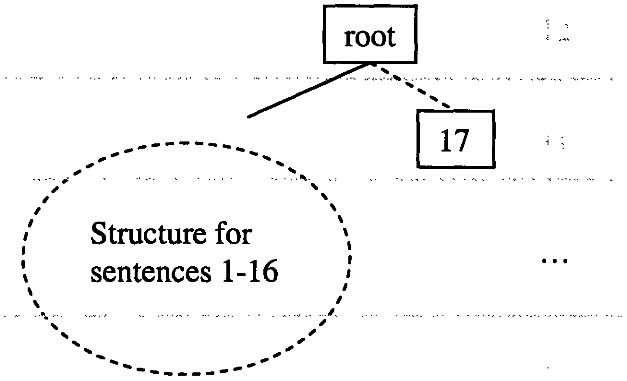


Figure 59. Partial representation of the tree discourse structure for text "wsj_1148" (Carlson et al. (2002)).

It is possible that human rankings of the last sentence of a text might be artificially low. This might be due in part to certain ways people are taught to read and write texts. However, in the discourse structure for the same text, Node 13 is ranked third by Marcu (2000)'s algorithm, but only 12th by human judges (both with and without paragraph information). Again, the high ranking of that node by Marcu (2000)'s algorithm is due to the tree representation which does not allow nodes with multiple parents, but enforces representations as shown in Figure 58 and Figure 59, thus resulting in an artificially high ranking for Node 13.

[28] Notice that the original tree structure for text "wsj_1148" in Carlson et al. (2002)'s database has 43 nodes. However, for the purpose of this section, we collapsed Carlson et al. (2002)'s text segments to sentences, in order to make them comparable to our human rankings. Collapsing Carlson et al. (2002)'s 43 text segments results in 17 sentences.

4.5.3 Coherence-based algorithms that operate on chain graphs

Our results show that coherence-based algorithms that operate on chain graphs (*in-degree*, *PageRank*, *cRank*) performed best. However, we have also identified a systematic problem for these algorithms. Consider the following constructed example:

(61)

0. Susan plans to do the following things tomorrow morning:
1. She needs to go grocery shopping.
2. She also has to pick up clothes from the drycleaner.
3. And she has to get some wine for dinner.

(constructed)

A possible coherence structure for (61) is shown in Figure 60[29]. By any of the chain graph-based ranking algorithms discussed in this chapter, segment 0 would get the highest ranking because of all the ingoing arcs leading to it. However, human raters would probably not rank segment 0 very high because it does not contain a lot of useful information.

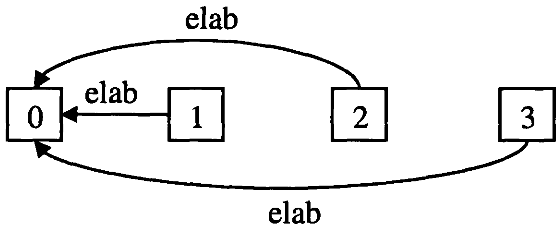


Figure 60. Coherence structure for (61).

An example of this phenomenon from our experimental results is from text "wsj_2354". The first sentence of that text is:

[29] Notice that there might be *similarity* relations between 1, 2, and 3 (all these text segments talk about similar things – things Susan has to do). However, these *similarity* relations have no bearing on the point made here, and are therefore left out in order to keep the graph structure simple.

(62)

Call it the "we're too broke to fight" defense.

(first sentence of text "wsj_2354")

This sentence is followed by a range of statements that elaborate on it, resulting in a structure that is similar to the one shown in Figure 60. Thus, by all chain graph-based ranking algorithms, that sentence is ranked first out of 16. However, the human rankings (with and without paragraph) were only fifth out of 16.

A possible way of dealing with this problem could be to combine coherence chain graph-based ranking algorithms with measures of text segment informational content. Then a text segment that is ranked high by a coherence chain graph-based algorithm would lose some of its rank if it has very low informational content. Possible ways of determining informational content could be measures of word entropy or measures like *tf.idf*. Future research should address this issue.

4.6 Conclusion

The goal of this section was to evaluate the results of three different kinds of sentence ranking algorithms and one commercially available summarizer. In order to evaluate the algorithms, we compared their sentence rankings to human sentence rankings of fifteen texts of varying length from the Wall Street Journal.

Our results indicated that a simple paragraph-based algorithm that was intended as a baseline performed very poorly, and that word-based and some coherence-based algorithms showed the best performance. The only commercially available summarizer that we tested, the MSWord summarizer, showed worse performance than most other algorithms. Furthermore, we found that a coherence-based algorithm that uses *cRank* and takes coherence chain graphs as input performed significantly better than most versions of a coherence-based algorithm that operates on coherence trees. When data from Experiments 1 and 2 were collapsed, the *cRank* algorithm performed significantly better than all other algorithms, except the other coherence-based algorithms that operate on coherence chain graphs.

Our results furthermore suggest that the weak performance of coherence tree-based algorithms might in part be due to the descriptive inadequacy of trees for representing coherence structures. In particular, we found that the inability of trees to represent nodes with multiple parents can systematically inflate text segment ranks.

5 General conclusion

The general topic of this thesis has been coherence in natural language. More specifically, the goals have been (1) to develop a descriptively adequate representation of discourse coherence that is easy to code; (2) to test the influence of coherence on psycholinguistic processes, in particular, pronoun processing; (3) to test the relation between coherence structures and the relative saliency of discourse segments.

In order to address the first goal, a coherence coding scheme was developed and applied to a set of naturally occurring texts. Then, the resultant representations of discourse structure were used to test hypotheses about descriptively adequate data structures for representing coherence. The results suggested that more powerful representations than trees are needed because naturally occurring discourse structures contained many different kinds of crossed dependencies and nodes with multiple parents. It has been argued that requiring more powerful data structures does not necessarily mean that there are no constraints at all on possible discourse structures, but that tree constraints are not the right kinds of constraints.

The second goal, testing the influence of coherence on pronoun processing, was addressed by conducting an on-line comprehension and an off-line production experiment. Results from both experiments suggested that structural, pronoun-specific accounts are unable to account for the full range of experimental results. Instead, the experimental results suggested an account where preferences in pronoun processing and production are a byproduct of more general cognitive processes used in determining discourse coherence.

The third goal of this thesis has been to test the relation between coherence structures and the relative saliency of discourse segments. In order to address this, different discourse segment ranking algorithms were tested and compared to human discourse segment rankings of naturally occurring texts. The results indicated that word-based algorithms provide a relatively high baseline, but that certain coherence-based algorithms can perform better than that baseline. However, the results also suggested that coherence-based algorithms that are based on chain graphs performed better than coherence-based algorithms that are based on trees. Some results from these experiments indicated that the difference in performance might in part be due to the fact that chain graph-based algorithms use a representation of discourse coherence that is more descriptively adequate.

Arguing that trees are not the right kind of constraint on possible coherence structures does not imply that any coherence structure is possible. Future work should address the question of what empirically valid constraints are on possible coherence structures. It should furthermore be explored to what extent these constraints may vary depending on factors such as discourse genre.

Future work should also address open issues in pronoun processing. One such open issue is pronoun processing preferences under other coherence relations than Resemblance or Cause-Effect. Additionally, the time course of pronoun processing should be explored: it is unclear, for example, at what point in processing coherence becomes a factor.

There are also open issues for approaches to discourse segment ranking. Future research should address the problems of coherence-based algorithms that operate on chain graphs (cf. Section 4.5.3). Other open questions are how useful sentence rankings are for tasks such as information extraction, or whether coherence-based algorithms can be used for query-specific sentence rankings.

6 References

Adrian Akmajian, & Ray Jackendoff. 1970. Coreferentiality and stress. *Linguistic Inquiry, 1*, 124-126.

M Ariel. 1990. *Accessing NP antecedents*. London: Routledge.

J Bateman, & KJ Rondhuis. 1994. *Coherence relations: Analysis and specifications*. Tilburg, Edinburgh, Darmstadt, Saarbrücken, Madrid.

Sabine Bergler. 1991. *The semantics of collocational patterns for reporting verbs*. Paper presented at the 5th Conference of the European Chapter of the Association for Computational Linguistics, Berlin, Germany.

Sabine Bergler. 1992. *Evidential analysis of reported speech*. Unpublished PhD thesis, Brandeis University, Waltham, MA, USA.

Lawrence Birnbaum. 1982. *Argument molecules: A functional representation of argument structures*. Paper presented at the Third National Conference on Artificial Intelligence (AAAI-82), Pittsburgh, PA.

Juergen Bortz. 1999. *Statistik fuer Sozialwissenschaftler*. Berlin: Springer Verlag.

Ronald Brandow, Karl Mitze, & Lisa F Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management, 31*(5), 675-685.

Sabine Brants, Sabine Dipper, Silvia Hansen, Wolfgang Lezius, & George Smith. 2002. *The TIGER Treebank*. Paper presented at the Workshop on Treebanks and Linguistic Theories, Sozopol, Bulgaria.

Susan E Brennan, Marilyn W Friedman, & Carl J Pollard. 1987. *A Centering approach to pronouns*. Paper presented at the 25th Meeting of the Association for Computational Linguistics, Stanford, CA.

Bruce K Britton. 1994. Understanding expository text. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 641-674). Madison, USA: Academic Press.

Orkut Buyukkokten, Hector Garcia-Molina, & Andreas Paepcke. 2001. *Seeing the whole in parts: Text summarization for web browsing on handheld devices*. Paper presented at the 10th International WWW Conference, Hong Kong, China.

Alfonso Caramazza, Ellen Grober, Catherine Garvey, & Jack Yates. 1977. Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behavior, 16*, 601-609.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249-254.

Lynn Carlson, John M Conroy, Daniel Marcu, Dianne P O'Leary, Mary E Okurowski, Anthony Taylor, et al. 2001. *An empirical study on the relation between abstracts, extracts, and the discourse structure of texts*. Paper presented at the DUC-2001, New Orleans, LA, USA.

Lynn Carlson, Daniel Marcu, & Mary E Okurowski. 2002. *RST Discourse Treebank*. Philadelphia, PA: Linguistic Data Consortium.

Lynn Carlson, Daniel Marcu, & Mary E Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. van Kuppevelt & R. Smith (Eds.), *Current directions in discourse and dialogue*. New York: Kluwer Academic Publishers.

Craig C Chambers, & Ron Smyth. 1998. Structural parallelism and discourse coherence: A test of Centering Theory. *Journal of Memory and Language, 39*, 593-608.

Noam Chomsky. 1973. Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 232-286). New York: Holt, Rinehart and Winston.

Simon Corston-Oliver. 1998. *Computing representations of the structure of written discourse*. Redmont, WA.

Laurence Danlos. 2004. *Discourse dependency structures as DAGs*. Paper presented at the SigDIAL2004, Cambridge, MA, USA.

Reinhard Diestel. 2000. *Graph theory*. New York: Springer Verlag.

Chris Ding, Xiaofeng He, Perry Husbands, Hongyuan Zha, & Horst Simon. 2002. *PageRank, HITS, and a unified framework for link analysis*. (No. 49372). Berkeley, CA, USA.

David Dowty. 1986. The effects of aspectual class on the temporal structure of discourse: Semantics or pragmatics? *Linguistics and Philosophy, 9*, 37-61.

Karen Ehrlich. 1980. Comprehension of pronouns. *Quarterly Journal of Experimental Psychology, 32*, 247-255.

Christiane Fellbaum (Ed.). 2001. *WordNet - An electronic lexical database*. Cambridge, MA: MIT Press.

Morten Frydenberg. 1989. The chain graph Markov property. *Scandinavian Journal of Statistics, 17*, 333-353.

Alan Garnham. 2001. *Mental models and the interpretation of anaphora*. Hove, East Sussex: Psychology Press.

Jade Goldstein, Mark Kantrowitz, Vibhu O Mittal, & Jamie O Carbonell. 1999. *Summarizing text documents: Sentence selection and evaluation metrics*. Paper presented at the SIGIR-99, Melbourne, Australia.

Yihong Gong, & Xin Liu. 2001. *Generic text summarization using relevance measure and latent semantic analysis*. Paper presented at the Annual ACM Conference on Research and Development in Information Retrieval, New Orleans, LA, USA.

Michael T Goodrich, & Robert Tamassia. 2001. *Data structures and algorithms in Java*. New York: John Wiley & Sons.

Peter C Gordon, Barbara J Grosz, & Laura A Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science, 17*, 311-347.

Herbert P Grice. 1975. Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 41-58). New York: Academic Press.

Barbara J Grosz, Arvind K Joshi, & Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics, 21*(2), 203-225.

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

Michael AK Haliday, & Ruqaiya Hassan. 1976. *Cohesion in English*. London: Longman.

MAK Halliday. 1985. *An introduction to functional grammar*. London: Edward Arnold.

Donna Harman, & Mark Liberman. 1993. *TIPSTER complete*. Philadelphia, PA: Linguistic Data Consortium.

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

Julia Hirschberg, & Christine H Nakatani. 1996. *A prosodic analysis of discourse segments in direction-giving monologues*. Paper presented at the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive Science, 3*, 67-90.

Jerry R Hobbs. 1985. *On the coherence and structure of discourse*. Stanford, CA.

Jerry R Hobbs, Martin E Stickel, Douglas E Appelt, & Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence, 63*, 69-142.

D Horn. 1942. A correction for the effect of tied ranks on the value of the rank difference correlation coefficient. *Journal of Educational Psychology, 33*, 686-690.

Eduard Hovy, & Elisabeth Maier. 1995. *Parsimonious or profligate: How many and which discourse relations?*

David Hume. 1748. *An enquiry concerning human understanding*, from http://www.infidels.org/library/historical/david_hume/human_understanding.html

Hongyan Jing, Kathleen R McKeown, Regina Barzilay, & Michael Elhadad. 1998. *Summarization evaluation methods: Experiments and analysis*. Paper presented at the AAAI-98 Spring Symposium on Intelligent Text Summarization, Stanford, CA, USA.

Christopher R Johnson, Miriam R L Petruck, Collin F Baker, Michael Ellsworth, Josef Ruppenhofer, & Charles J Fillmore. 2003. *FrameNet: Theory and practice*, from http://www.icsi.berkeley.edu/~framenet/book/book.html

Philip Johnson-Laird. 1983. *Mental models*. Cambridge, MA: Harvard University Press.

Aravind K Joshi, & Steve Kuhn. 1979. *Centered logic: The role of entity centered sentence representation in natural language inferencing*. Paper presented at the 6th International Joint Conference on Artificial Intelligence, Tokio, Japan.

Aravind K Joshi, & Yves Schabes. 1997. Tree-adjoining grammars. In G. Rozenberg & A. Salomaa (Eds.), *Handbook of formal languages* (pp. 69-123). New York: Springer.

Marcel A Just, Patricia A Carpenter, & Jacqueline D Woolley. 1982. Paradigms and processing in reading comprehension. *Journal of Experimental Psychology: General, 111*, 228-238.

Hans Kamp, & Christian Rohrer. 1983. Tense in texts. In R. Baeuerle, C. Schwarze & A. von Stechow (Eds.), *Meaning, use, and interpretation of language* (pp. 250-269). Berlin: de Gruyter.

Andrew Kehler. 2002. *Coherence, reference, and the theory of grammar*. Stanford, CA.

Walter Kintsch, & Teun A van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.

Alistair Knott. 1996. *A data-driven methodology for motivating a set of coherence relations*. Unpublished PhD thesis, University of Edinburgh, Edinburgh.

Esther König, & Wolfgang Lezius. 2000. *A description language for syntactically annotated corpora*. Paper presented at the Computational Linguistics Conference (COLING), Saarbrücken, Germany.

William Labov. 1972. *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania Press.

Shalom Lappin, & Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics, 20*(4), 535-561.

Alex Lascarides, & Nicholas Asher. 1991. *Discourse relations and defeasible knowledge*. Paper presented at the 9th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA.

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

Steffen Lauritzen, & Nanny Wermuth. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics, 17*, 31-57.

Robert E Longacre. 1983. *The grammar of discourse*. New York: Plenum Press.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159-165.

William C Mann, & Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243-281.

Christopher D Manning, & Hinrich Schuetze. 2000. *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. Cambridge, MA: MIT Press.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, et al. 1994. *The Penn Treebank: Annotating predicate argument structure*. Paper presented at the ARPA Human Language Technology Workshop, San Francisco, CA.

JR Martin. 1992. *English text: Systems and structure*. Amsterdam: Benjamins.

Kathleen R McKeown. 1985. *Text generation: Using discourse strategies and focus constraints to generate natural language text*. Cambridge, UK: Cambridge University Press.

Eleni Miltsakaki, Rashmi Prasad, Aravind K Joshi, & Bonnie L Webber. 2004. *The Penn Discourse TreeBank*. Paper presented at the Language and Resources and Evaluation Conference, Lisbon, Portugal.

Mandar Mitra, Amit Singhal, & Chris Buckley. 1997. *Automatic text summarization by paragraph extraction*. Paper presented at the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain.

Johanna D Moore, & Martha E Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics, 18*(4), 537-544.

Megan Moser, & Johanna D Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics, 22*(3), 409-419.

Wilhelm Nestle (Ed.). 1977. *Aristoteles: Hauptwerke*. Berlin: Kroener Verlag.

Kenji Ono, Kazuo Sumita, & Seiji Miike. 1994. *Abstract generation based on rhetorical structure extraction*. Paper presented at the COLING-94, Kyoto, Japan.

Lawrence Page, Sergey Brin, Rajeev Motwani, & Terry Winograd. 1998. *The PageRank citation ranking: Bringing order to the web*. Stanford, CA.

Carolyn Penstein Rose, Barbara Di Eugenio, Lori S Levin, & Carol Van Ess-Dykema. 1995. *Discourse processing of dialogues with multiple threads*. Paper presented at the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA.

Livia Polanyi. 1996. *The linguistic structure of discourse*. Stanford, CA.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, & David Ahn. 2004. *A rule based approach to discourse parsing*. Paper presented at the SigDIAL 2004, Cambridge, MA, USA.

Livia Polanyi, & Remko Scha. 1984. *A syntactic approach to discourse semantics*. Paper presented at the 10th International Conference on Computational Linguistics, Stanford, CA, USA.

Dragomir R Radev, Eduard Hovy, & Kathleen R McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics, 28*(4), 399-408.

Rachel Reichman. 1985. *Getting computers to talk like you and me*. Cambridge, MA, USA: MIT Press.

S Russell, & P Norvig. 1995. *Artificial intelligence: A modern approach*. New Jersey: Prentice-Hall.

Gerard Salton, & Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513-523.

TJM Sanders, WPM Spooren, & LGM Noordman. 1992. Towards a taxonomy of coherence relations. *Discourse Processes, 15*(1), 1-36.

Anthony J Sanford, & Simon C Garrod. 1989. What, when, and how?: Questions of immediacy in anaphoric reference resolution. *Language and Cognitive Processes, 4*(3/4), 235-262.

H Schauer. 2000. *Referential structure and coherence structure*. Paper presented at the TALN 2000, Lausanne, Switzerland.

Stuart M Shieber. 1986. *An introduction to unification-based approaches to grammar*. Stanford, CA.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, & Hans Uszkoreit. 1997. *An annotation scheme for free word order languages*. Paper presented at the Fifth Conference on Applied Natural Language Processing (ANLP-97), Washington, D.C.

Ron H Smyth. 1994. Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research, 23*, 197-229.

Karen Sparck-Jones. 1993. What might be in a summary? In G. Knorz, J. Krause & C. Womser-Hacker (Eds.), *Information retrieval 93: Von der Modellierung zur Anwendung* (pp. 9-26). Konstanz: Universitaetsverlag.

Karen Sparck-Jones, & Tetsuya Sakai. 2001, September 2001. *Generic summaries for indexing in IR*. Paper presented at the ACM SIGIR-2001, New Orleans, LA, USA.

Andrew J Stewart, Martin J Pickering, & Anthony J Sanford. 2000. The role of implicit causality in language comprehension: Focus versus integration accounts. *Journal of Memory and Language, 42*, 423-443.

Teun A van Dijk, & Walter Kintsch. 1983. *Strategies of discourse comprehension*. New York: Academic Press.

Carl Vogel, Ulrike Hahn, & Holly Branigan. 1996. *Cross-serial dependencies are not hard to process*. Paper presented at the 16th International Conference on Computational Linguistics, Copenhagen, Denmark.

Marilyn A Walker. 1998. Centering, anaphora resolution, and discourse structure. In E. Prince, A. Joshi & M. Walker (Eds.), *Centering Theory in discourse.* Oxford: Oxford University Press.

Bonnie L Webber, Alistair Knott, & Aravind K Joshi. 1999a. *Multiple discourse connectives in a lexicalized grammar for discourse.* Paper presented at the 3rd International Workshop on Computational Semantics, Tilburg, Netherlands.

Bonnie L Webber, Alistair Knott, Matthew Stone, & Aravind K Joshi. 1999b. *Discourse relations: A structural and presuppositional account using lexicalized TAG.* Paper presented at the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), College Park, MD, USA.

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

Terry Winograd. 1972. *Understanding natural language.* New York: Academic Press.

Florian Wolf, & Edward Gibson. 2004a. *Paragraph-, word-, and coherence-based approaches to sentence ranking: A comparison of algorithm and human performance.* Paper presented at the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain.

Florian Wolf, & Edward Gibson. 2004b. *Representing discourse coherence: A corpus-based analysis.* Paper presented at the 20th International Conference on Computational Linguistics, Geneva, Switzerland.

Florian Wolf, Edward Gibson, & Timothy Desmet. in press. Coherence and pronoun processing. *Language and Cognitive Processes.*

Florian Wolf, Edward Gibson, Amy Fisher, & Meredith Knight. 2003. *A procedure for collecting a database of texts annotated with coherence relations.*Unpublished manuscript, Cambridge, MA.

Wilhelm Wundt. 1911. *Völkerpsychologie.* Leipzig: Engelmann Verlag.

Klaus Zechner. 1996. *Fast generation of abstracts from general domain text corpora by extracting relevant sentences.* Paper presented at the COLING-96, Copenhagen, Denmark.

Ingrid Zukerman, & Richard McConachy. 1995. *Generating discourse across several user modules: Maximizing belief while avoiding boredom and overload.* Paper presented at the Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada.

Appendix A: Items for Experiment 1 in Chapter 3

1. Charles commended/saved Harriet and similarly/so Richard praised her/him in the newspaper but everything was just a big show.
2. David reprimanded/betrayed Sarah and similarly/so Helen chastised her/him after the holidays but all the criticism showed very little effect.
3. Michael disciplined/attacked Shirley and similarly/so Leonard punished her/him two days ago but in the end they reached an agreement.
4. Peter questioned/assaulted Julie and similarly/so Carol interrogated her/him for an hour but a few moments later the police arrived at the scene.
5. Stuart honored/liberated Martha and similarly/so Joseph admired her/him a great deal but unfortunately the feeling was not mutual.
6. Nathan disliked/abandoned Allyssa and similarly/so Nicole hated her/him for a while and in the end they all avoided each other.
7. Ryan safeguarded/feared Emma and similarly/so Adam protected her/him in the evening but all their caution would probably not have been necessary.
8. Kevin rebuked/kicked Claire and similarly/so Grace scolded her/him in the house but nobody else cared about all these quarrels.
9. Erik embraced/rescued Lisa and similarly/so Liam hugged her/him with great enthusiasm and everybody was a little bit relieved.
10. Brian scolded/harassed Cathy and similarly/so Scott lectured her/him after the meeting and everybody ended up hating each other.
11. Fiona complimented/defeated Craig and similarly/so James congratulated him/her after the match but nobody took any notice.
12. Christina lectured/pestered Christopher and similarly/so Stephanie reprimanded him/her for one hour although nobody thought it would have any effect.
13. Jonathan despised/denounced Madeline and similarly/so Patricia scorned her using harsh language but after a while everybody was reconciled again.

14. Rebecca interrogated/punched Anthony and similarly/so Suzanne cross-examined him/her for a while but nothing interesting was said.
15. Melissa suspected/deceived William and similarly/so Natalie distrusted him/her in the end and the whole working atmosphere was spoiled.
16. Tina thanked/supported Robert and similarly/so Fred acknowledged him/her at the conference but nobody seemed to be sincere.
17. Sophia admired/outdid Joshua and similarly/so Gloria respected him/her in the beginning but very soon things changed.
18. Melanie hired/impressed Bradley and similarly/so Malcolm recruited him/her after the interview but not all of the co-workers were satisfied with the situation.
19. Heather hit/insulted Aaron and similarly/so Caitlin punched him/her in the nose and the result was a big fight.
20. Hannah appointed/outperformed Michael and similarly/so George nominated him/her for the job although some people were not happy with the decision.