Integrating Prosody into an Account of Discourse Structure

Sofia Gustafson-Čapková

A Dissertation submitted to
Stockholm University
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

May 2005

Computational Linguistics
Department of Linguistics
Stockholm University
S–106 91 Stockholm
Sweden

National Graduate School
of Language Technology
Faculty of Humanities
Gothenburg University
S–430 00 Göteborg
Sweden

ISBN91–7155–050–X

ISBN91–7155–050–X

Typeset by the author using LaTeX

Printed by Printcenter, Stockholm University, Stockholm 2005

MOD! FRAMÅT!

Abstract

In this thesis a study of discourse segmenting is carried out, which investigates both segment boundaries and segment content. The results are related to discourse theory. We study the questions of how the prosody and the text structure influence subjects' annotations of discourse boundaries and discourse prominence. The hypothesis was that the annotations would be influenced by the discourse type.

Two studies were carried out. 1) a study of boundary annotation, 2) a study of prominence annotation. All studies were made on four different discourse types, scripted and spontaneous monologue and scripted and spontaneous dialogue. In addition the annotations were carried out under two different conditions 1) based on transcripts alone and 2) based on transcripts together with access to the speech signal.

The results indicate that the boundary annotations were less dependent on the speech signal than the prominence annotations. It seems that subjects have segmented on the basis of the text structure, while prominence to a great extent was annotated on the basis of the prosody. In the case of boundary markings the boundary context in terms of parts of speech differs across speaking styles, which is not the case for the prominences. A separate study of segment intentions was also made, and it was found that the interpretation of a specific intention, questions, seems to be arrived at primarily on the basis of the text structure. However, in some cases also the prosody affects the annotations.

The picture that emerges indicates a distribution of labour between text structure and prosody, governed by the principle of economy. In cases where the boundaries were less well definied, as in e.g. spontaneous monologue, the pattern of the prominences was clearer. In cases where the boundaries were more clearly indicated, as in read aloud text, the prominences were less clearly communicated.

The findings were interpreted within Grosz and Sidner's (1986) discourse theory. It is suggested that differences in the segmenting strategy originating from the interaction of text structure and prosody can be expressed as differences in the contributions from the different components of discourse suggested in the framework of Grosz and Sidner (1986).

Descriptors

Discourse Structure, Swedish Prosody, Speaking Styles, Spoken Language, Written Language, Dialogue, Monologue.

Sammanfattning

I denna avhandling studeras diskurssegment och diskurssegmentering i olika talstilar. Segmenten har undersökts både utifrån gränser – vad som avgränsar dem från varandra – och utifrån prominens – segmentens innehåll. Vidare studeras hur prosodi och textstruktur påverkar försökspersoners annoteringar av gränser och prominens, och resultaten har sedan relaterats till en teoretisk beskrivning av diskurssegment.

Två delstudier utfördes: 1) en studie av gränser och 2) en studie av prominens. Båda studierna gjorder på material från fyra olika talstilar, textbaserad och spontan monolog samt texbaserad och spontan dialog. Vidare gjordes varje delstudie i två olika villkor. I det första villkoret annoterade försökspersoner gränser eller prominenser i transkriptioner av texter. Också i det andra villkoret gjorde försökspersonerna annoteringar i transkriptioner av samma texter, men i detta villkor hade de dessutom tillgång till de ljudfiler som transkriptionerna baserades på.

Resultaten visar att gränsannoteringarna påverkades mindre av tillgång till talsignalen än prominensannoteringarna. Det tycks vara så att försökspersonerna har angvit gränser till stor del på grundval av textstrukturen, medan prominens till stor del tycks vara annoterad på grundval av prosodin. Vad gäller gränsannoteringarna så skiljer sig kontexten, i termer av ordklass, mellan talstilarna. Detta är dock inte fallet vad gäller prominensannoteringarna. En separat studie av en typ av segmentintention, frågor, utfördes också. Denna studie antydde att frågor i de flesta fall annoterades på grundval av textstrukturen, men i vissa fall påverkas de också av prosodin.

Helhetsbilden som tonar fram för samspelet mellan textstruktur och prosodi i utformandet av diskurssegment pekar mot en komplementär relation som styrs av en balans mellan ekonomi och tydlighet. I de fall där gränserna är vagare, som i t. ex. spontantal, är de framhävda delarna klarare. I de fall där gränserna är klarare, som i t. ex. uppläst tal, är de framhävda partierna mindre klart markerade.

Resultaten tolkades i Grosz och Sidners (1986) diskursteori. Skillnader i segmenteringsstrategierna relaterade till samspelet mellan textstruktur och prosodi, tolkas som skillnader i bidragen från de olika diskurkomponenter som förslås i Grosz och Sidners Grosz and Sidner (1986) diskursteori.

Acknowledgements

Many people have been helpful to me in the work on this thesis. First of all I want to thank my extremely patient and encouraging supervisors, David House and Elisabet Engdahl. They have followed this work through manuscripts consisting of the most confusing jungles of words sharing just a slight resemblance to natural language. Thank you again, and again and again Elisabet and David! Talking about supervisors I would also take the chance to thank earlier supervisors who all contributed, each in their own way, to the continuation of my work. Going back even further, many thanks go to Christina Hellman, and Benny Brodda, and of course I would like to do my best to send a thank you also to Gunnel Källgren.

Even though the supervisors were many, other people have also been most helpful in my work. I will try to recapitulate the chronological order in which they kicked my work a bit ahead, however, I am a little bit unsure about the order. First then, many thanks go to Pétur Helgason for letting me use his recordings of Map Task dialogues. Many, many thanks go also to Beáta Megyesi for her fruitful cooperation in the work on pausing in different speaking styles, for the tagging and parsing of my data and for being a supportive friend in general. Thank you Bea! I want to thank Björn Gambäck for letting me use his LaTex template for the thesis. Without this help I guess I would have been even later... I also owe many thanks to Hassan Djamshidpey who helped me to carry out the recordings of the monologues and also assisted me in converting data from extremely strange sound file formats. Thanks also to Swedish radio who generously let me use a recording of a radio theatre piece. Sara Rydin made my life easier in letting me use her forms for computing κ statistics. They were most useful. Thank you Sara! And, I just have to send many thanks to speakers A, B and C as well as to all my anonymous subjects!

I would like to thank my collegues in the Department of Linguistics for encouraging me to continue on days when I would rather have gone home to sleep, in particular I want to thank my office mate of many years, Jennifer Spenader. I would also like to thank the whole section of Computational Linguistics for being understanding about my absent-mindedness during the first few months of 2005. I guess that collegues asking me over and over again when I would finish my thesis also are part of the reason that it was finished. So I thank everybody for their attention and interest in my work, even if the help was not always appreciated.

Moreover, I would also like to express my gratitude for having had the chance to be part of the National Graduate School of Language Technology (GSLT). Many of my new acquaintances have been most inspiring in providing brave plans and interesting discussions. Above all, I have to say a little extra thank you to Stina Ericson!

Thus, many people in the academic environment have helped me, and all have in their own way pushed me on a bit further. However, I think their efforts would have been completely fruitless if I had not had the deep and constant support from my parents

Gisela and Björn Gustafson. Without their calm and loving home open to me and my daughter Marta I would have stopped trying before I even started.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

IN the study of meaning, the principle of compositionality plays an important role. It claims that the meaning of the whole is determined by the meaning of the parts and the way they are put together (Frege, 1882). Thus, the meaning of a sentence is determined by the meaning of the words and how the words are combined according to the grammar. It is possible to transfer this view on meaning to the study of stretches of language above the sentence level too: in order to understand a longer message, it is not enough to understand the separate parts of the message. In addition we must understand how these parts are connected to each other, and how together they form the full message. Such a description of the structure of a longer message is what is offered in different kinds of discourse theories, see e.g. Grosz and Sidner (1986) and Mann and Thompson (1988), where the longer message, i.e. the discourse, is constructed from discourse segments, the building blocks of the discourse.

If we adopt the compositional approach and suppose that the impression of coherence of a longer spoken or written message depends on the understanding of its parts and the way those parts are combined, an obvious question is: what constitutes these parts in a discourse, and how are these segments combined? One aim of this thesis is to investigate some characteristics of those discourse segments in different types of spoken Swedish, and thus to investigate how characteristics of discourse segments may differ across speaking styles.

Of course, what constitutes the units above the word and phrase level of a specific message is not a complete mystery. For instance, we have the clause, the sentence and the paragraph, as signalled by the syntactic composition. In written language these units are in most cases shown by punctuation: comma, full stop and paragraph marking. Furthermore, we can also use typographical means to denote higher level groupings like sections and chapters etc. All these means to mark structure in written language are transparent, and we become aware of their use when we learn to read and write at school.

However, in spoken language, the signals indicating the parts of discourse are not as transparent as in the case of punctuation. In general, speakers are not actively aware what signals they use to indicate segments in spoken language in the same way as they are aware of it in the written one. In spoken language these signals include e.g. melody, rhythm and pausing pattern, i.e. speech prosody. The importance of these signals for the shaping of the message structure is stressed by many researchers, see e.g. Grosz and Hirschberg (1992), Swerts (1994), Passonneau and Litman (1997), Bruce (1998), van Donzel (1999), Wennerstrom (2001), Horne (2001) and Hansson (2003).

Spoken and written language differ not only with regard to punctuation and prosody, but also with regard to the lexicogrammatical structure, see e.g. Biber (1988) for English and Einarsson (1978) for Swedish. This also becomes evident when we examine representations of written and spoken language. Examples 1.1 and 1.2 render excerpts from written and spoken language respectively. Example 1.1 is an excerpt from a written representation of a news article and example 1.2 shows a passage from the spontaneous retelling of the same article. Both examples are from the materials used in our studies for this thesis and are described in more detail in chapter 3.

(1.1) Arten människa skaffar sig en religion. Det är en naturlig följd av att vi till skillnad från övriga arter har ett språk.

English transliteration:
Species-DEF-SG human gets self a religion. This is a natural consequence of that we to difference from other species have a language.

English translation:
The human species acquires a religion. This is a natural consequence of the fact that unlike other species we have a language.

(1.2) ...det gör det gjorde att människan kunde ehm kommunicera med varandra och kunna samordna sina krafter...

English transliteration:
...this does this did that mank-DEF-SG could-FIN ehm communicate with one-another-PL and be-able-to-INF coordinate their forces...

English translation:
...this does this did mean that man could ehm communicate with one another and be able to coordinate their forces...

When inspecting the representations in 1.1 and 1.2 we see examples of different structural properties in written and spoken language. In the written language (1.1) the clauses

are constructed according to the grammar book, and sentence boundaries are neatly signalled with a full stop and a capital letter. In contrast to the written language the spoken language example includes features such as a restart (...det gör det gjorde), an agreement error between "människan" (man-DEF-SG) and "varandra" (one-another-PL), and a coordination of one clause with a finite verb and one clause with an infinite verb sharing the same subject (...människan kunde ehm kommunicera med varandra och kunna samordna...). The differences between example 1.1 and 1.2 are distinct, even though the transcript of the speech is very close to written language orthography.

Considering the structural differences between example 1.1 and 1.2, are there corresponding differences in the prosody? In other words, are there specific differences in the use of prosody between read aloud speech (scripted speech) and spontaneous speech (non-scripted speech) mirroring the structural differences in scripted and non-scripted language samples? This question, i.e. the question of how the structure of the string of words and the use of prosody interact in signalling discourse segments in different speaking styles is the main question for this thesis.

Specifically we investigate how on one hand structure and on the other some prosodic variables influence listeners' segmenting of the message, and whether the distribution of labour between the lexicogrammatical structure and the prosody differ between different speaking styles. For example, if the organisation of the lexicogrammatical structure were less regulated, would then the use of some prosodic features become more rigid?

To investigate this question we have performed a series of experiments. These experiments are carried out as comparative studies between four specific speaking styles, an approach that allows us to see if, and in that case how, the use of speech prosody varies under different lexicogrammatical structures. The results are then related to a framework of discourse theory, and we suggest how prosody can be integrated more closely into such a theory.

The unit we investigate is the discourse segment. Therefore, let us elaborate a little on the concept of a segment. Turning to pictures, we can distinguish between different strategies to express a motif: i) we can focus on the contours, ii) we can focus on what is between the contours, or iii) we can make a synthesis of the contours and what is between them.

In figure 1.1 three methods of expressing a motif are exemplified in three sketches of a cat. The top picture shows the contours of a cat. In the middle there is a cat shaped by what we see between the contours, and at the bottom the sketch captures both the contours of the cat and what is between the contours. In other words, in the bottom picture the contours and the spaces in-between interact in shaping the motif. Thus, the first method is to define the contours and induce the content, the second one is to define the content and induce the contours and, lastly, the third method is to work in both directions and make use of both contours and content.

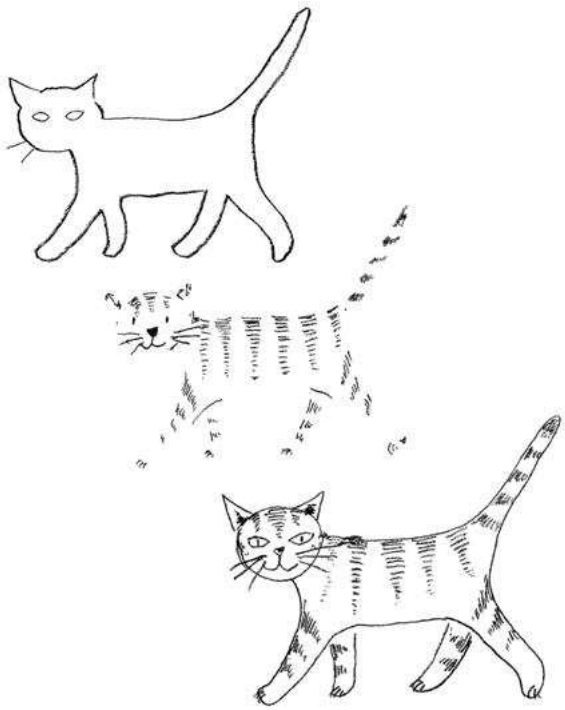

Figure 1.1: Three different ways to sketch a cat

Do different motifs profit more from one or another strategy? Are there motifs which are made clearer from a description of their contours and are there other motifs which are standing out better with a description of their content? For instance, a house might be easier to sketch by contours than a sky. This question is the background to the work in this thesis. If we see the discourse segment as our motif and assume that we can define it either from the contours, i.e. the boundaries, or from the content, i.e. the prominent information carried by the segment, then we can also think about two aspects of describing a discourse segment. Is thus one approach more advantageous in one type of discourse than in another?

In this thesis our first question concerns the relationship between the lexicogrammatical structure and the prosody in the shaping of discourse segments. What does the structure contribute to the segments and what does the prosody contribute? If the structure changes, would then also the prosody change together with the structural change? Moreover, if the structure changes, would the emphasis of the prosodic aspect of the boundaries or of the prominence become more important in the segmenting? For instance, with clearer syntactic boundaries, would also the prosodic boundary marking be clearer in relation to this? In addition, with less clear syntactic boundaries, would the content of the segments become clearer?

We look for answers to these questions in an empirical study of our motif; the discourse segment. In order to change the structure we vary the speaking style, and in order to control what segmenting is due to structure and what is due to prosody we vary between a non-prosody condition and a prosody condition.

We examine four different forms of spoken Swedish, scripted and non-scripted monologue and dialogue. Since the degree of spontaneity in a speaker is difficult to determine, we instead distinguish between scripted and non-scripted speech. With scripted speech we mean any kind of discourse where the content is explicitly present in advance in a script, as in e.g. read aloud speech, and with non-scripted we mean any kind of discourse that is executed without a manuscript.

The structure can be assumed to differ between scripted and non-scripted speech in a similar way to what is shown in example 1.2. In other words, in the case of scriptedness there are clearer syntactic boundaries, while in the non-scripted speaking styles the boundaries are more fuzzy in terms of hesitations and restarts. In addition, the interaction in the dialogues adds yet another kind of boundary to what is present in the monologues, i.e. speaker change. The prosodic feature connected to the boundaries on which we focus are the silent pauses. To study how structure on one hand and prosody on the other affect discourse segmenting we had subjects annotate discourse boundaries, either using transcripts alone or having access to both transcripts and sound recordings.

In spoken discourse, above all two things make elements prominent: the information structure and the focal accent. Since the information structure is partly signalled by the string of words and partly by the prosody, we can assume differences regarding how prominence is expressed in scripted speech compared to non-scripted speech. To investigate the impact of structure on one hand and prosody on the other, also in this task we had subjects annotate discourse prominence, either using transcripts alone or having access to both transcripts and sound recordings.

In our studies we had inexperienced subjects annotate materials from four speaking styles. Two groups worked with the annotation of boundaries; one group did this using transcripts alone, and another group did this using transcripts and sound files. Two other groups worked with the annotation of prominence, and also here one group annotated on the basis of transcripts alone while the other group annotated on the basis of both transcripts and sound files. This experiment setup allowed us to see how the annotations differed between a condition where subjects had access to the prosody and the condition where they had access only to the transcripted string of words in four different speaking styles. In other words, in each of the four speaking styles we could single out the contribution to boundaries and prominence from the structure on one hand and the prosody on the other.

Our next question was how to connect possible variations between the speaking styles regarding the discourse segments to the discourse theory. Many discourse theories, e.g. Grosz and Sidner's (1986) and Mann and Thompson's (1988), have a very loose, if any, theoretical connection to prosody and thus a weak link to a very important part of

the structure of spoken language. However, since discourse theories in general contain some concept of discourse segments and their internal relationships, prosodic phenomena such as prosodic phrasing and prominence can be assumed to be integral parts of the structuring of spoken discourse. Therefore, it is important to pursue the question of a closer relation between discourse theory, prosody and style.

Since the theory of Grosz and Sidner (1986) is one of the discourse theories that claims to hold for both monologue and dialogue, we have chosen this framework for the interpretation of our results. Briefly, Grosz and Sidner's (1986) view on discourse includes three aspects of each discourse segment: the linguistic structure, the attentional state and the intentional structure. The linguistic structure concerns the verbal realisation of a discourse, i.e. the surface form of the discourse segments. The attentional state is a record of concepts in focus of attention in a segment, and the intentional structure concerns the intention behind a specific segment. We have designed our study with regard to these three aspects of discourse, relating the linguistic structure to the boundary annotation, the attentional state to the prominence annotation, and in our study the intentional structure is mirrored by a smaller study of a subset of the segments, the questions.

Thus, the aim of this thesis is to investigate segmental structure as signalled by the string of words on one hand and the prosody on the other in different types of spoken Swedish. Specifically we study the distribution of labour between the structure (the string of words) and the prosody in the different speaking styles. Based on our findings we suggest a way to integrate an account of prosody into the discourse theory.

The structure of the thesis is as following: Chapter 2 gives the background to the work in this thesis. This includes a brief description of phenomena related to discourse structure and how prosody shapes discourse. This is followed by an overview of different conceptions of the discourse segment as well as a description of Grosz and Sidner's (1986) discourse theory. In addition there is a brief survey of lexical and prosodic features contributing to discourse segments and discourse structure, as well as of differences and similarities in written and spoken language. The chapter ends with an account of our research questions.

Chapter 3 describes the design of our studies. We elaborate on how each study relates to Grosz and Sidner's (1986) discourse theory, and give a comprehensive account of the speech materials we use.

In chapter 4 we describe the preprocessing and preanalysis of the materials. This includes the linguistic mark-up, the work with the validation of the differences between the speaking styles, the work with the development of the database, and an overview of the subjects' boundary annotations. In addition we show how the annotations were normalized and also account for the κ statistics which were used in the analyses.

The study of the boundary annotation is presented in chapter 5, and the study of the prominence annotation in chapter 6. Each of these two chapters ends with a discussion

relating to the study in question. These discussions are primarily concerned with the results relating to each of these studies in isolation, and a general discussion follows later in chapter 8.

In chapter 7 we report on the study of a specific type of segments – or a specific segment intention – questions. Also in this case a discussion is included at the end of the chapter, but more general comments are also found in chapter 8.

Chapter 8 contains the general discussion where findings from all three studies are brought together, and based on these findings we suggest a way to integrate prosodic variables into a theory of discourse.

Some concluding remarks of the work are rendered in chapter 9

Before proceeding to the background, let us briefly sum up the main questions in this thesis. First we study discourse segmenting divided into two tasks, the annotation of boundaries (the segment contours) and the annotation of prominence (the segment content). In addition we study one specific type of segment intention: questions. Thus, the first question is: How do i) the string of words and ii) the prosody affect the three different aspects of a discourse segment: i) boundaries, ii) prominence and iii) intention. Are there systematic differences between the four speaking styles in the distribution of labour between the string of words and the use of prosody? The second question concerns a way to integrate prosody into existing discourse theory; are there systematic differences between the speaking styles with regard to the interaction between structure and prosody of a kind that could be expressed in a theory of discourse?

Background

Chapter 2

Discourse Segments and their Relationship to Natural Language Processing

WHAT role do discourse segments, discourse structure and discourse processing play in Natural Language Processing (NLP)? Many researchers have stressed the importance of a discourse level analysis for solving problems like anaphora resolution, see e.g. Webber (1991), Walker (1998), Eckert and Strube (2000), Webber *et al.* (2003), for tracing information status through elements in the discourse, see e.g. Grosz *et al.* (1995), Walker (1996), van Donzel (1999), and for tagging sentence or utterance functions, see e.g. Hirschberg and Nakatani (1996), Marcu (1997), Shriberg *et al.* (1998). Hobbs (1985), who has pointed out that the complexity in the discourse level analysis also raises practical problems with using the analysis in computational systems. Nevertheless, even though there might be difficulties with the applicability of the analyses in computational systems, it is clear that some features in natural language depend on a context wider than a single sentence, thus, they have to be accounted for in a discourse perspective. With regard to this, the most feasible strategy is perhaps not to develop extremely detailed NLP systems making a deep discourse analysis in order to handle these features. However, a deeper understanding of discourse in general is needed in order to understand what could be simplified in a computational approach.

In most approaches to analyses of discourse structure the notion of *discourse segments* is central. A discourse segment means a portion of language, not necessarily overlapping with a clause or a sentence, and it represents the building block on which the discourse is constructed. Discourse segments are described in various shapes and by various names by e.g. Grosjean *et al.* (1979), Grosz and Sidner (1986), Mann and Thompson (1988), Carletta *et al.* (1997) and Allen and Core (1997). Since these different approaches to discourse and discourse segments mirror different ways of looking at discourse, e.g. from

a syntactic, a pragmatic or an acoustic point of view, there are many ways to define the segments.

In this chapter we give the framework for the research related to discourse and discourse segmenting in NLP. In the previous chapter, a picture of a segment as consisting of both the boundaries and the prominent content was outlined. Therefore, a range of phenomena communicating boundaries as well as prominence in language including prosodic as well as lexicogrammatical features is surveyed. In addition, different notions of discourse segments are described and briefly related to different approaches to discourse in general.

Many researchers have stressed the important role that prosody plays in structuring the spoken language message, including different aspects of discourse segmenting, see e.g. Strangert (1990), Grosz and Hirschberg (1992), Swerts and Geluykens (1994), van Donzel (1999) and Hansson (2003). Thus, in an investigation of discourse structure in spoken language an account of the correlations between lexical and prosodic signals is important. The need for an integration of prosody into discourse theory is pointed out by many researchers, in many more extensive apporaches to discourse it is, however, less pursued, see e.g. Grosz and Sidner (1986), Mann and Thompson (1988), Grosz *et al.* (1995). The aim in this thesis is to make such a relationship between a discourse model and Swedish prosody more visible.

One way to study the relationship between prosody and discourse structure is to make a comparative study between lexicogrammatical and prosodic features and their relationship to discourse structure. In the introduction we sketched a hypothesis that the way of expressing an object might differ according to its motif, i.e. one motif would communicate the best picture through the contours whereas another motif would communicate the best picture through the content. The general motif studied in this thesis is the discourse segment, and different speaking styles are used in order to vary the motif. Thus, in this study different speaking styles supply the means of varying the discourse structure, and in these different styles selected lexicogrammatical and prosodic features are studied. In order to throw some light upon differences across speaking styles, both lexicogrammatical and prosodic, a brief survey of research on similarities and differences between written and spoken language is included.

At the end of the chapter we recapitulate and elaborate on the research questions in this thesis.

This thesis focuses on how linguistic and prosodic properties interact with stylistic properties. Thus, we study the realization of discourse segments with regard to linguistic and prosodic cues to boundaries and prominence, and how these features coincide in different speaking styles. Before we proceed any further, it should be stated that this means that there are other aspects of discourse which are left out entirely. It is most important to mention that we do not investigate any part of discourse meaning as studied in e.g. approaches to discourse semantics represented by Asher (1993) or Blackburn and Bos (forthcoming). The fact that this perspective is not present in this

thesis does not imply that we find it superfluous or unconnected, it means only that we have chosen to focus on another aspect.

The chapter is structured as follows: section 2.1 gives a general introduction to the concept of discourse and discourse structure, and also establishes some of the terminology used in the rest of this thesis. In addition the relationship between discourse structure and some linguistic phenomena is highlighted. Then, in section 2.2 some background to speech prosody and an introduction to the Lund model of intonation as presented by Bruce (1998) is given. In section 2.3 different notions of discourse segments are given, and in section 2.4 Grosz and Sidner's (1986) discourse theory is described in more detail. In section 2.6 differences between written and spoken Swedish are surveyed. In the last section, section 2.7 the research questions in this thesis are recapitulated and elaborated upon.

2.1 What Is Discourse and What Phenomena Does It Account for?

So far, we have stressed the discourse segments and the two aspects of these segments, boundaries and prominences. However, what is a complete discourse, and what kind of analysis is made on the discourse level? In this section some terms used in this thesis are defined, and a very brief introductory description of different views on discourse structure is given. A more detailed description follows in sections 2.3 and 2.4. In addition we clarify the relationship between information structure and discourse segments, and examine some examples of phenomena often accounted for from a discourse perspective.

The term *discourse* has many different meanings, but in this thesis it is defined as *a long stretch of coherent spoken or written language.* This is a fairly wide use of the term since it includes both spoken and written language as well as both monologue and dialogue, including multi-party dialogue. However, in the studies carried out in this thesis such a wide definition is needed in order to treat the data in a unifying way.

In the study in this thesis four types of discourses are used: read aloud monologue and dialogue as well as spontaneous monologue and dialogue. However, as previously mentioned, the terms "written" and "spontaneous" are not used since there is a problem with assessing the degree of spontaneity. Instead a distinction is made between *scripted* and *non-scripted* language. Moreover, we need to distinguish between different types of spoken discourse, thus, the term *speaking styles* is used in our studies for the different varieties of spoken Swedish discourse. The term style is in this thesis taken to mean a distinct use of spoken language, for instance scripted or non-scripted speech, and the use of the term is not restricted to categorories such as fiction or non-fiction. To enable a distinction between the verbal content of a discourse (i.e. the words without the sound) and the prosody of the same discourse (i.e. without the verbal content) we shall also distinguish between the *string of words* and the *prosody.* To clarify this: a speaking style

is any type of spoken discourse. A scripted speaking style refers to speech based on a written manuscript, while a non-scripted style is speech executed without a manuscript. Furthermore, we can distinguish between the string of words and the prosody in any speaking style.

A discourse can be described as consisting of discourse segments which contribute to the discourse structure (Grosz and Sidner 1986; Mann and Thompson 1988; Polanyi 1988; Webber 1991; Passonneau and Litman 1997). Some researchers describe the discourse structure as hierarchical, see e.g. (Grosz and Sidner 1986, Mann and Thompson 1988, Polanyi 1988, Grosz and Hirschberg 1992, Carletta *et al.* 1997) whereas others see it as linear, see e.g. (Walker 1996, Passonneau 1993, Passonneau and Litman 1997). Nevertheless, in both approaches the description of the discourse structure aims to account for dependencies across sentences such as e.g. inter-sentential anaphora, inter-sentential coherence and structural relationships between segments such as the sequential form in recipes and question – answer pairs.

Since most approaches to discourse include some segmenting procedure, often based on surface cues such as syntax, the discourse boundaries are implicitly present in nearly all types of discourse analysis. However, the discourse prominences have a less distinct status in discourse analysis, and are more often dealt with within research on information structure. In this thesis the specific study of information structure does not have a prominent position. However, the information structure and the discourse structure can be assumed to be closely interrelated. The information structure within a single sentence is not autonomous, but is included in – and dependent on – a more extensive discourse (Enkvist, 1974:190).

According to Ahrenberg (1995) the concept of discourse structure includes arranging the parts of discourse into a structure that reflects focus. Such an arrangement mirrors both how different segments relate to each other and which concepts are activated and prominent at a certain point in the discourse. This view thus includes both features above the sentence level (relations between segments) and features from the information structure (focus) into the concept of discourse.

To clarify how the idea of prominence is used in this thesis we offer a brief overview of concepts such as information structure, activation, accessibility and focus, and their relationship to prominence.

The information structure can be described as the arrangement of the words within a sentence in order to communicate the pragmatic structure of an underlying proposition (Lambrecht, 1994). The information structure can be signalled through lexicon, prosody, syntactic constituents (in particular nominal) and their order within a sentence (Lambrecht, 1994).

Lambrecht (1994) mentions three main categories of information structure: i) *presupposition and assertion*, ii)*identifiability and activation* and iii)*topic and focus.*

The categories of presupposition and assertion capture the speaker's and the hearer's idea of elements agreed on as introduced into the discourse (old or given elements) and elements not yet agreed on as introduced (new elements). It is thus a classification of specific pieces of propositional information in the discourse. The presupposition relates to what the speaker assumes that the hearer already knows, old (given) information, while the assertion relates to information that the speaker assumes is new to the hearer (new information) (Lambrecht, 1994:52).

Both these statuses of information, given and new, are related to the form of the utterance. New information is in general said to come later, at the end of the sentence and to be less reduced, both lexically and acoustically, while given information in general comes closer to the beginning of the sentence and can be both lexically and acoustically reduced, i.e. it often takes the form of pronouns and acoustically unaccented expressions. Thus, one aspect of prominence is related to assertions and new information.

The category of identifiability and activation contains the speaker's assessment of whether a discourse representation of a referent is already stored in the hearer's mind or not (identifiability) and the degree of accessibility of an identified referent in the hearer's mind (activation) (Lambrecht, 1994:76). Thus, this category concerns the discourse participants' mental representations of discourse referents.

Specifically, activation is related to the notion of prominence since few representations which the speaker assumes to be less activated by the hearer are often uttered in less reduced forms such as longer phrases and acoustic accent. More activated representation instead has a more reduced form. Thus, the notion of prominence is related to less activated representations.

According to Lambrecht (1994:6) the category of topic and focus contains the relative predictability and unpredictability of the relationships between propositions and their elements in given discourse situations. Topic is described as what a sentence is about and focus as what is predicated about the topic. The topic – focus relationship thus concerns the relationship between the elements within a sentence.

Topic and focus are related to prominence since the topic is often more reduced than the focus, whereas focus often has a less reduced form and is often accented and expressed with longer phrases. Thus, focus is often more prominent than topic.

In addition to sentence topic as described above, the notion of discourse topic is also used. This covers the topic of a longer stretch of sentences or utterances, i.e. what a stretch of discourse is about rather than what a sentence is about.

The term focus also has slightly different uses than those described above; they are, however, still relevant for the notion of prominence. Gundel (1999) mentions three kinds of focus: *psychological focus*, *semantic focus* and *contrastive focus*. In addition, the term focus is used to mean *acoustic focus*. Psychological focus refers to an entity which is currently the focus of attention in the speaker's and/or listener's representation of the discourse information (i.e. related to presupposed, activated and topic). Semantic focus

refers to new information predicated about the topic, and it is often accompanied by syntactic or acoustic prominence (i.e. related to assertion, unidentifiable/inactive and focus). Semantic focus often, but not necessarily always, puts an entity into psychological focus. Thus entities in psychological focus *are* already highly activated, whereas entities in semantic focus might *become* highly activated (and thus in psychological focus). It is unclear whether a discourse entity within one discourse segment can first be in semantic focus and later in psychological focus, or if such a change of status of focus implies a segment boundary. Contrastive focus means that an element is acquiring prominence by being contrasted to another element in the discourse. Also contrastive focus is often acoustically prominent. Lastly, acoustic focus correlates to focal accent, i.e. acoustic prominence, which is described more closely in section 2.2.

All the different categories of information structure described above depend on a combination of lexical, syntactic and acoustic features, and communicate different aspects of information status through elements in the discourse. In this thesis a finegrained distinction between these categories is not made, instead a distinction between prominent and not prominent elements is made. This distinction is based on the subjects' annotations of prominence and is thus a perceptual category containing elements from the information structure categories above. Our use of the notion *prominence* thus relates to the notion of the prominent parts of the discourse segments.

The close relationship between focality, accessibility and discourse structure is mirrored in a number of phenomena accounted for in a discourse perspective. One such phenomenon is anaphora resolution. So, let us take a closer look at how anaphora resolution is said to be related to discourse structure and focus. Anaphora resolution concerns the finding of the referent of a certain anaphoric expression. A successful anaphora resolution depends on the accessibility of the referent via the antecedent at the point of the anaphoric expression, see e.g. Webber (1991), Gundel (1999), Walker (1998). This dependency has been related to discourse structure in that the anaphoric expression has to be accessible through the discourse structure, i.e. located in a discourse segment from which it can access the discourse segment containing the antecedent directly (Webber, 1991). If the discourse structure is described as a tree structure, the mother node is thus accessible, however not the sister node. This accessibility is moreover said to mirror the focality (or activation) by the referent, thus indicating that anaphora resolution implies an analysis of both the relationships between the discourse segments and the activation status of entities in the discourse. The discourse structure thus accounts for the structural arrangements of the discourse as well as for the accessibility of possible discourse entities at each point in the discourse. In this way anaphora resolution together with focality can be related to discourse structure. Attempts to model anaphora resolution related to focus have been made by e.g. Sidner (1983), Webber (1991) and Grosz *et al.* (1995), and e.g. Grosz and Sidner (1986) points out the relevance of discourse structure in general for anaphora resolution.

The discourse structure can also mirror coherence patterns, i.e. in which way two discourse segments are related to each other. This type of discourse analysis contains two

aspects of coherence: i) coherence in terms of structural patterns such as tree structure relationships and ii) coherence in terms of meaningful discourse relationships (such as e.g. "contrast" or "elaboration") between two segments.

An example of structural relationships is that some sections of discourse can be understood to be sequential while other ones can be understood to be hierarchically dependent. For instance, instructions in recipes are examples of discourses with a sequential structure, i.e. the ingredients have to be added in strict order and the instructions are sequentially mirroring this order. An example of a hierarchically dependent structure is a statement and a number of motivations for this statement. In the discourse theory of Grosz and Sidner (1986) there is a focus on such structural relationships, and we offer a more elaborate description of this particular theory in section 2.4.

Coherence relationships focusing on meaning are in slightly different varieties present in the work of e.g. Hobbs (1985), Mann and Thompson (1988) and Polanyi (1988). Briefly, all these approaches include the labelling of a sequential or hierarchical relationship with a description of the nature of the relationship. The nature of this description, however, differs between researchers in the degree of formality.

There is no consensus as to whether the discourse structure should be modelled linearly or hierarchically, and as already mentioned there are proponents for both approaches. Very generally, the proponents for a hierarchical discourse structure emphasize the relationships between discourse segments while proponents for a linear structure emphasize the degree of activation by elements, i.e. focality issues, in the discourse.

Phenomena to be contained within the field of discourse are on one hand related to focality, activation and accessibility and on the other hand related to the coherence relationships between sections of discourse. In turn, both are related to segmenting in terms of boundaries which delimit the segments and content which stands out as important.

2.2 Prosodic Features that Shape the Discourse

We have already pointed out the importance of prosody for the discourse structure in spoken language. Prosody divides the flow of speech into portions sharing similarities with units such as discourse segments. In spoken language, the discourse structure in terms of clauses, sentences and paragraphs is signalled by the prosody (Godman-Eisler 1972, Bruce 1982, Grosz and Hirschberg 1992, Strangert 1992, Swets 1997). Prosody is often described as the suprasegmental properties of spoken language, i.e. the systematic acoustic features above the level of a single phoneme, and the most obvious role for prosody is the structuring of the discourse (Bruce, 1998). Prosody shapes spoken language into e.g. phrases, clauses and larger portions, thus facilitating the communication process for both speaker and listener (ibid).

In an investigation of discourse structure in spoken language prosody is central, specifically the concepts of prosodic phrases and utterances, in many ways resembling a discourse segment. In this section we provide a brief background to Swedish prosody with a focus on prosodic phrasing, as described by Bruce (1998).

The primary properties of prosody are rhythmic, dynamic and melodic, and they are closely related to the acoustic correlates duration, intensity and pitch (F0) respectively (Bruce, 1998). The primary functions of prosody are, according to Bruce (1998:15) i) to give prominence, ii) to group and iii) to communicate discourse functions such as e.g. turn taking and to signal attitude.

According to Bruce (1998:14) the two prosodic functions prominence and grouping encompass two complementary aspects. Prominence includes emphasizing ("bringing out") and deemphasizing ("withholding"), while grouping includes both to signal boundaries between groups and coherence within groups (ibid:14). Thus, in the field of prosody the dual perspective on the parts of discourse which we call discourse segments is clearly formulated.

Bruce (1998:15) points out that many of the communicative functions related to prominence and grouping which can be expressed through prosody can also be signalled by other means. In table 2.1 we show some alternative or coinciding means to express prominence, grouping and discourse functions as presented by Bruce (1998:16). Thus, what we find in the column "Syntactic and lexical" can be described as the boundaries and prominences as communicated by the string of words in the discourse segments.

Table 2.1 shows that prominence can be signalled both by prosodic means, e.g. by differences in stress, and by verbal means, e.g. by differences in the syntactic construction. The aspect of bringing out (emphasizing) makes elements more notable through e.g. pitch accent and/or topicalization, while the aspect of withholding (deemphasizing) gives elements in the discourse a more reduced form by e.g. deaccenting and/or pronominalization. In a similar way, grouping may be signalled in different ways. Prosodically it can be signalled by prosodic phrasing, but grouping can also be signalled syntactically. The aspect of boundaries has a prosodic correlate in e.g. pausing and syntactic correlates in the syntactic boundaries. The aspect of coherence has a prosodic correlate in prosodic connection, i.e. when the F0 contour signals continuation. Syntactically the coherence can be signalled by syntactic continuation. The discourse function, exemplified by turn-taking and attitude signalling is not more closely specified, but on this level there are both prosodic and syntactic–lexical signals.

Table 2.1 reflects quite clearly the idea behind the studies in this thesis. We will investigate how the segmenting of discourse is made from the aspects of both prominence and grouping, and also from the aspect of prosodic means and lexical/syntactic means. Regarding prominence we focus on how elements are emphasised, i.e. the realization of pitch accent in terms of focal accent. Regarding the grouping we focus on the boundary signalling in terms of pauses. We mean to change the lexcial/syntactic structure by va-

Alternative and coinciding forms of expression			
	PROSODIC	**SYNTACTIC AND LEXICAL**	**NON-VERBAL**
PROMINENCE	**Differences in stress**	**Syntactic construction**	**Accompanying gestures**
Bringing out, emphasizing	Pitch accent	Emphatic paraphrasing, topicalization	
Withholding, deemphasizing	Deaccenting	Pronominalization, ellipse	
GROUPING	**Prosodic phrasing**	**Phrase structure**	
Boundary	Pausing, etc.	Syntactic boundary	
Coherence	Prosodic connection	Phrase connection, conjunctions	
DISCOURSE	**Intonation, intensity etc.**	**Word order, choice of lexem**	**Body language, gestures, facial expressions**

Table 2.1: Alternative ways of expressing prominence, grouping and discourse functions, adapted from Bruce (1998:16).

rying the speaking styles between monologue and dialogue as well as between scripted and non-scripted speech.

2.2.1 The Lund Model of Intonation for Swedish

In the Lund model of intonation, prominence has a correlate in focal accent, while the boundary marking has a correlate in prosodic phrasing. In the next two sections we give the characteristics of focal accent and prosodic phrasing in Swedish, following the Lund model of intonation given by Bruce (1998). The model of focal accent holds for central Swedish, meaning the variety of Swedish spoken in and around Stockholm (ibid).

Prominence and Accent

Acoustic prominence is usually shown by accenting. In Swedish three levels of prominence are distinguished: i) stress, ii)(word) accent and iii) focal (nuclear or sentence) accent (Bruce 1998:80). The stress is connected to the syllable and has the foot as domain, the accent is connected to the foot and has the word as domain and the focal accent is connected to the word and has the phrase as domain.

In this work we leave out the first level, stress. The reason for doing this is that there is a clear difference between on one hand stress and on the other hand accent and focal accent. Stress has a complex phonetic correlate (duration, intensity and spectral contrast), while accent and focal accent have a primary correlate in tonal gestures (Bruce 1998:81). In addition, stress contributes to a basic pattern of speech rhythm and establishes relationships between syllables, while accent and focal accent represent a reinforcement of the basic pattern and the domain is above the foot (ibid). Moreover, since the two levels of accent and focal accent are the stronger levels, they can also be assumed to be more relevant to prominence on the discourse level, in particular the focal accent.

There are two accents in central Swedish, accent I and accent II, which are assigned on a lexical basis. Accent I is characterized by a rising tonal gesture to a high tone in the pretonic syllable, and a fall to a low tone in the stressed vowel. The word accent may in Swedish also have a distinctive function. In focal position the gesture continues with a rise from the low tone in the stressed vowel to a high tone later in the word. Accent II is characterized by a tonal gesture rising to a high tone in the beginning of the stressed vowel and then a fall through the syllable to a low tone. In focal accent the gesture continues with a rise which gives accent II a distinct contour with two peaks (Bruce 1998:104). Both accent I and accent II are thus characterized by a tonal rise in focal position, however with a difference in the timing: the rise comes later in accent II than in accent I (ibid). A schematic picture of the tonal gesture in accent I and accent II is shown in figure 2.1, picture from Bruce (1977).

The typical position for focal accent is on the last constituent in an utterance (Bruce 1998:178). Thus, in a sentence focal accent theoretically correlates with the default position of new information.

The most important acoustic correlate for withholding, i.e. the feature of reduction which is parallel to given information, is the absence of accent gesture. However, in the studies we are not investigating this part specifically, instead we are focusing on the aspect of acoustic prominence in terms of focal accent.

Prosodic Phrasing

Prosodic phrasing is the prosodic means of dividing the discourse into segments. There are two aspects of this: boundary marking and continuation. In the study in this thesis

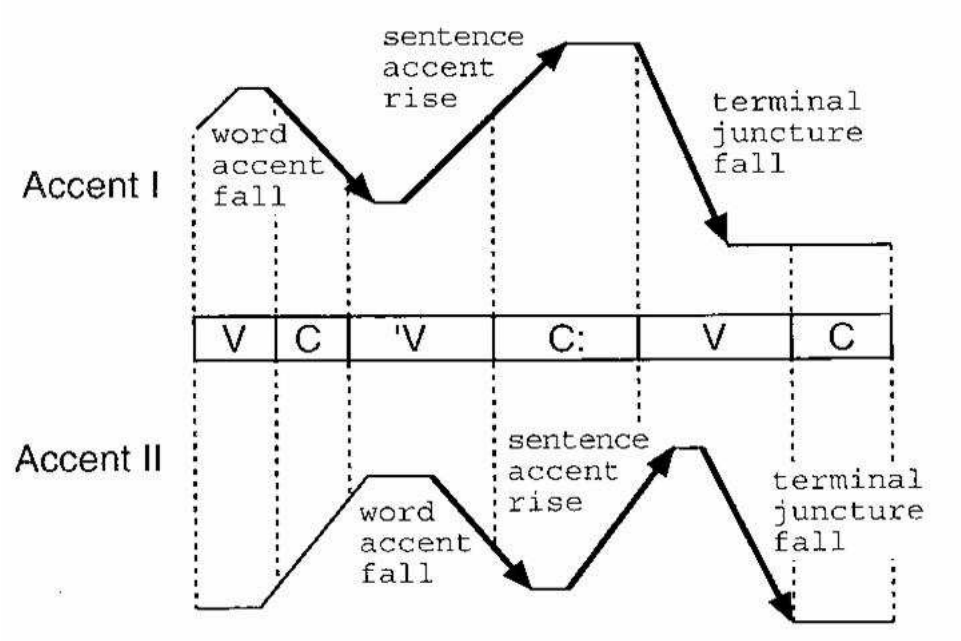

Figure 2.1: The F0- contributions of word accent. Picture from Bruce (1977), reprinted with permission.

there is a focus on the boundary marking, which is why we devote more space to describing the aspect of segmenting than to the aspect of continuation. Prosody groups the speech on multiple levels, e.g. on the word level, on the phrase level and on the utterance level. In this description we focus on the phrase and utterance levels.

Prosodic phrasing involves aspects of both phrasing and accent, but we start by accounting for some correlates to phrasing and then elaborate on the relationship to accent. Phonetic correlates to prosodic phrasing primarily include pausing, lengthening of phrase final segments, F0 fall through the phrase or utterance (although a terminal rise is also possible), lower intensity and a change in voice quality to e.g. a creaky voice (Bruce 1998:133). In the acoustic signalling of prominence, one acoustic cue is not always enough (Bruce, 1998:134), but a combination of cues is needed.

Phrasing is more than just boundaries. The spans signalling coherence between the boundaries are also important in signalling a group or a prosodic phrase. Coherence is signalled by acoustic features opposite to those signalling boundaries, e.g. like higher tempo, higher intensity etc. (Bruce, 1998). Bruce (1998:131)notes that there is interaction between prosodic phrasing and accenting in the signalling of a prosodic phrase or utterance. If two subsequent words both carry accent, they are most probably interpreted as belonging to different phrases. However, if the first of the words is deaccentend

while the second one is accented, it is likely that both words are interpreted as belonging to the same phrase.

Another indication of the importance of the accent for the segmenting of the speech signal is given by Bruce (1998:125) in relation to word recognition. He poses the question as to how a listener can separate words when no "blank spaces" are present in the speech signalling the word boundaries. The answer given by Bruce is that listeners probably direct their attention to the salient peaks of the words, and not to the boundaries in between. Guided by the relevant content the listener then interprets and segments the words. The segmenting might thus also be carried from the centers out to the boundaries instead of from the boundaries alone.

2.3 Different Conceptions of the Discourse Segments

One of the aims of this thesis is to integrate features from speech prosody more closely into a theoretical approach to discourse structure and thus to relate features from prosodic prominence and prosodic phrasing to an existing theoretical concept of a discourse segment. In this section we survey a number of approaches to discourse segments, and at the end of the section is stated the type of discourse segment to which the study in this thesis is related.

Discourse structure can be described as linear or hierarchical, as covering structural or semantic-pragmatic relationships, or as focusing. In the same way as discourse structure can be described in different ways, so can discourse segments. In this section we examine more closely the properties of the discourse segment as described in different approaches to discourse. The units of analysis are not termed "discourse segments" by all researchers, however, we include all types of segments sharing similarities with discourse segments in order to supply a unifying view on discourse segments.

Inspecting some of the more extensive approaches to discourse, i.e. Hobbs (1985), Grosz and Sidner (1986), Polanyi (1988), Mann and Thompson (1988) and Carletta *et al.* (1997) we find a number of different basic motivations for the discourse segments including segments based on speaker intention, on semantic meaning, on grammatical form and on turn taking. In addition some researchers, e.g. Hobbs (1985), argue that discourse segments are artifacts.

In our survey of different conceptions of discourse segments we distinguish between:

- Intention based segments
- Semantically based segments
- Textually based segments
- Interactionally based segments

The approaches to dialogue coding in NLP have developed from focusing on the isolated speech act to include an account of interaction. The later extensions have much in common with features from conversation analysis. Nevertheless, the dialogue coding is kept under the heading of intention based segmenting since the focus in dialogue act tagging in NLP is on the function behind an utterance and not on how the linguistic behaviour of an individual should be interpreted.

2.3.1 Intention-Based Discourse Segments

The view of discourse segments as being intention based relates to the idea about speech acts formulated by Austin (1962 1975). In this view the base is a communicative intention expressed in a linguistic form fulfilling a communicative function. In other words, a conversation serves to realize and communicate the communicative intention(s). Thus, in this view a discourse segment is isomorphic with the linguistic form which corresponds to one intention.

Austin's 1955 Harvard lectures, first published in 1962, form the starting point of speech act theory. Austin (1962 1975) argues that to speak is to perform an action, hence the term speech act for the unit. Prototypical examples of this kind of action are expressions containing performative verbs such as "I promise" or "I bet".

A speech act cannot be true or false, only felicitous or infelicitous. For a speech act to be felicitous means that it makes the intended action behind the speech act come true, i.e. it has to be uttered under certain specified felicitous conditions of success. This means, for instance, that the person uttering a certain promise must be in a position to make that promise.

A speech act is a complex unit. Austin offered an analysis of the concept of speech acts, which distinguishes between three aspects of a speech act:

- Locutionary act
- Illocutionary act
- Perlocutionary act

These three aspects of the speech act are in their turn complex. The *locutionary* act includes the phonetic act (producing noises), the phatic act (conforming the phonetic noises to a certain vocabulary and grammar), and the rhetic act (the use of the phatic act with a special sense and reference) (Austin, 1962 1975). The locutionary aspect deals with saying something which makes sense in a certain language, and can thus be seen as connected to the traditional semantic meaning of language.

The *illocutionary* act is the aspect of the act performed in saying something, i.e. asking or answering a question, giving information etc. The illocutionary act is viewed as composed

of illocutionary force, specifying the type of action (question, answer etc), together with the propositional content which specifies the action in more detail. This aspect can be said to mirror the speaker's or producer's intention behind an utterance.

The third aspect, the *perlocutionary* act, has the characteristics of being connected to the effects of the utterance, i.e. what effect a certain utterance provokes in a certain context. Examples are e.g. persuasion and surprise. This aspect can be seen as mirroring the listener's perception of the intentional content of a certain utterance.

In addition to the speech acts themselves, Austin also suggested a taxonomy for the classification of the different types of speech acts. This taxonomy was further extended and refined by Searle (Searle, 1969).

Austin's and Searle's work on speech acts, e.g.Austin (1962 1975) and Searle (1969), became integrated into a wider account of communication. Most prominently, speech acts were integrated into different notions of dialogue plans or plan-based approaches to dialogue (see e.g. Bruce (1975), Cohen and Perrault (1979)). In these approaches, the dialogue is defined as a plan which is based on the state of affairs and the speaker's communicative intent. The speech act units are parts of the dialogue plan, and they are defined in terms of the goals, beliefs and wishes etc. of interacting agents. Consequently the speech acts become part of a more structured account of dialogue, including e.g. the dialogue participants' plans and beliefs. Further development towards a more formal account for connected dialogues was also made by e.g. Allen and Perrault (1980), Allen (1983), Cohen and Levesque (1991) and Litman and Allen (1990).

All these approaches have in common that they aim to target macro coherence in dialogue instead of the isolated functions of single speech acts. This interest in dialogue as a whole is mirrored in multi-level approaches to dialogue description by e.g. Traum and Hinkelman (1992), who introduced conversation acts like "turn-taking" and "repair", i.e. acts related to dialogue management rather than to utterance function.

The analyses of dialogues contain several levels or strata (Traum, 1999), following analyses of class room conversations made by Sinclair and Coulthardt (1975). In this tradition the dialogue exchange is described as consisting of dialogue moves (utterances), dialogue games (plans) and dialogue transactions (goals), levels which have come to be used in dialogue tag sets such as the one reported by e.g. Carletta *et al.* (1997).

Even though the initial speech act approach has undergone great changes, the speech act unit can still be considered to be the basic unit of NLP approaches to dialogue. The reason for this is that basic segmenting is done on this level. In addition, Carletta *et al.* (1997) also report the highest level of agreement between annotators on this level, indicating a better understanding of these units by the annotators.

There are close similarities between the intention-based discourse segments suggested by Grosz and Sidner (1986) and the dialogue act in dialogue coding as represented by Carletta *et al.* (1997), meaning that both units are based on an intention underlying the actual linguistic behavior. Grosz and Sidner's (1986) discourse theory is also goal-

oriented and was primarily developed for the analysis of dialogue but does also apply to monologue. However, one clear difference between on one hand Grosz and Sidner (1986) and Carletta *et al.* (1997) is the labelling of the discourse segments. In Grosz and Sidner's (1986) approach there is no finite set of intentions for labelling the discourse segments, only structural relationships are labelled. The dialogue act tagset tested by Carletta *et al.* (1997) has, however, a finite set of intention labels such as e.g. "Request" or "Inform", and each segment is assigned one such label.

A discourse segment in the notion of Grosz and Sidner (1986) is said to be characterized by one discourse relevant intention. A discourse relevant intention means that the intention related to the discourse segment has to play a central role in the actual discourse. For example, after a request for something we expect an answer. The question requires the answer and thus both contribute to the development of the discourse as well as give rise to a structural connection between the two utterances which in this case is equivalent to two discourse segments. In this case the intention to request is said to be discourse relevant.

However, not all intentions are discourse relevant, for instance an intention to impress does not influence the discourse structure in the same way as a request. Grosz and Sidner (1986) also point out that one single segment might have multiple underlying intentions, for instance an intention to request something may at the same time be intended to impress. In such cases Grosz and Sidner (1986) argue that it is always possible to isolate one single intention as (most) relevant to the discourse, and this one is the discourse segment intention.

If the intention is taken to be the underlying motivation for discourse segments, the linguistic form can be seen as showing different "symptoms" of the underlying intention. For instance, a question form can indicate an intention to request something. In dialogue coding one line of research has tried to correlate linguistic features with a somewhat wider set of dialogue acts, see e.g. Carletta *et al.* (1997), Allen and Core (1997), Samuel *et al.* (1998) and Stolcke *et al.* (2000). Systems making use of the (Grosz and Sidner, 1986) approach, as described by e.g. Ahrenberg *et al.* (1995), instead make use of a minimal set of intention labels – e.g. only "Initiation" and "Response", i.e. labels describing the meaning of the function of the structure rather than describing the meaning of the function in relation to an interpretation of the meaning.

The intention based discourse segments are quite freely defined, which means that they might be able to capture all our types of data, both monologue and dialogue, both scripted and non-scripted speech.

2.3.2 Semantically Based Units of Discourse

Some approaches to discourse structure take a semantic approach to the parts of discourse. One example of this approach is Livia Polanyi's Linguistic Discourse Model

presented in Polanyi (1988, 1996, 2001). In this view the propositional content of an *elementary discourse constituent unit* is the basic unit of discourse which combines with an *extra propositional discourse operator* to form a discourse, these two parts reflecting the distinction between linguistic content and form (Polanyi, 2001). Polanyi, (1988, 1996, 2001) makes it clear that the intention as motivation for the basic units of discourse is rejected because of the vagueness in the definition of discourse intentions and discourse purposes.

Polanyi argues for a segmenting into clauses, since a clause usually contains one proposition (Polanyi, 1996). The combination of the clauses is determined by their semantic meaning, and the relationship between the two clauses is computed on the basis of their meaning. Polanyi (1996) addresses the parsing of the discourse, while the focus in Polanyi (2001) is on the discourse interpretation which is performed within the framework of Discourse Representation Theory (DRT) (Kamp and Reyle, 1993).

A semantic approach to discourse is also taken by Hobbs (1985). However, in the definition of the units it is more the relationships between them, the *coherence relationships*, which are defined while the discourse segments are said to be artifacts originating from the coherence relationships. In a similar way to Polanyi (1988, 1996, 2001) the segments in practice correlate to clauses. An explanation of this might be that the semantics of the clauses are included in the definitions of the discourse relationships, i.e. the definition of the relationships is based on the content the clauses to which they relate. In Hobbs (1985) the focus lies on the formal definition of specific coherence relationships, such as elaboration and explanation. In addition Hobbs (1985) stresses the knowledge base and inference machinery needed to compute the relationships.

The semantic approach as exemplified by both Polanyi and Hobbs does not imply that discourse intentions in the vein of e.g. Grosz and Sidner (1986) do not exist, but the discourse is viewed from a different perspective than the one used by Grosz and Sidner. The semantic perspective implies a further consequence: the clause as a unit is primary since one clause usually carries one proposition, and semantic interpretation works from the propositions. In the intention-based view the discourse segment is not as clearly defined syntactically, instead it is said that a discourse segment covers a span of discourse, ranging from a phrase to a number of sentences, which is defined on the basis of an underlying intention.

Since this semantically based approach of using propositions is closely related to syntactic units on the surface representation of the segments, it is not entirely suitable for the study in this thesis. We would rather need a unit which is not as syntactically fixed as this, but which offers the possibility of including units not traditionally said to carry semantic content such as e.g. hesitations.

2.3.3 Textually Based Discourse Segments

The textually based segments characterizes a view where neither intentions nor semantics are said to constitute the base of the segments. This view is present in e.g Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In RST the discourse segment is a unit of arbitrary size (Mann and Thompson, 1988), no underlying base of the segment is defined. However, it is suggested that the division of the text units should be based on some theory-neutral classification. RST is said to capture relationships between clauses in a text (Mann and Thompson, 1988), thus both single and complex clauses seem to be candidates for discourse segments in RST.

Based on the segment content, the discourse segments are related to each other in a bottom-up fashion thus shaping a hierarchical discourse tree. Each relationship between two segments is given a label from a set of predefined relationships. Typically one of the segments is classified as the *nucleus* and the other one is classified as the *satellite*, the nucleus being rather independent of the satellite, but not the other way around. There are thus some similarities between nucleus and satellite and the information structure categories topic and focus, but on a different level of analysis. An RST-analysis captures an interpretation of the text structure, but it is possible that different analysts produce different analyses.

A RST relationship is in (Mann and Thompson, 1988) defined as having four aspects: 1) restrictions on the nucleus, 2) restrictions on the satellite, 3) restrictions on the combination of the nucleus and the satellite, 4) the effect. In addition the locus of effect is defined, i.e. whether the effect is communicated through the nucleus, the satellite or both. Number four, the definition of effect is comparable to the definition of the perlocutionary act in speech act theory.

RST is a theory primarily concerned with the text structure, and not with the intentions behind the text (Mann and Thompson, 1988). However, Moser and Moore (1996) argue that the intentional base is implicit also in the RST approach because of the distinction between nucleus and satellite, which requires that the analyst interprets the idea behind the two segments and is able to label one segment as more central than the other one.

RST shows similarities to both Grosz and Sidner (1986) and Polanyi (1988, 1996, 2001) in the way in which it suggests that a hierarchical structure is the result of an analysis of the discourse. In addition, both (Grosz and Sidner, 1986) and (Mann and Thompson, 1988) have clear instructions that the discourse segments structurally relate to each other in a specific dominance order. In Grosz and Sidner (1986) the idea is expressed in terms of dominance and satisfaction – precedence relationships, whereas in RST it is expressed through the relationship between nucleus and satellite. Moreover, in RST the idea of connecting segments with relationships carrying a specific meaning, such as "elaboration" or "contrast" is closely related to the discourse relationships by Polanyi (1988, 1996) and also to Hobbs (1985), although to a lesser extent.

The textually based approach bases the notion of a segment on a syntactic unit such as the sentence or the clause. It is developed in relation to written text, and it is also more suitable for scripted data than for non-scripted data. Since the rhetorical relationships primarily describe the argumentative structure, it is more difficult to apply them to interactive units such as feedback expressions. This makes the framework of RST less suitable for the data in the study in this thesis.

2.3.4 Conversationally Defined Discourse Segments

The fourth perspective on discourse segments, the conversationally defined base, is not commonly used in relation to NLP. However, as we saw, dialogue tagging has points in common with this perspective, and therefore we nevertheless briefly touch upon it. In conversation analysis, (CA), segments are defined through interaction in terms of turns which can consist of one or more utterances. The utterances can be described as carrying meaning in the same way as speech acts, but the turns are primarily mirroring the interaction between the dialogue participants. A segment in CA is called a turn constructing unit (TCU), and possible segment boundaries are classified as transition relevant points (TRP) (Norrby, 2004). A TRP can be a turn boundary, but it does not have to be.

A CA transcription of a dialogue carefully records a range of aspects relevant to the interaction between the dialogue participants, including e.g. the precise form of the tokens, the speech prosody, gazes and nods. The goal is to account for a specific piece of interaction as precisely as possible.

The difference between CA and dialogue coding is primarily that while dialogue tagging aims to offer a framework for computational tagging and processing of dialogues in general, CA aims to analyse and explain specific dialogues. Thus, the component of analysis which aims to generate a dialogue computationally is not present in CA. Much of the understanding of linguistic interaction in general, which over a period of time has become included in dialogue coding such as dialogue management features, has been inspired by research in the field of CA. The differences are thus not so much differences in the tools of analysis as differences in the aims of the analyses.

The conversationally based approach requires interaction, i.e. a dialogue, and since we want to study both monologue and dialogue this approach is less suitable for our study.

2.3.5 Comparison of the Different Approaches to Discourse Segments

The above examples show that discourse segments are described in different terms in different approaches. One of these is the approach to discourse focusing on the intention behind what is said, and within this approach there are both theories focusing on

structural relationships and relationships connected to meaning. In the semantic approaches the focus is on the semantic meaning of the segments, and thus clauses are an important unit. In RST no specific unit is given, but since the material often is written, sentences make a good candidate. In the CA approach the turn, defined from an interactive point of view, is the basic unit. This survey shows that the representation and range of the discourse segments vary depending on which of their aspects we focus on.

To sum up, we have described discourse segments from an intentional, a semantic, a textual and a conversational perspective. The differences can be captured in an example. Let us imagine a situation where a boy, Peter, has to do his homework. His brother Carl can see that Peter is playing instead, and tells their mother, uttering "Peter is playing". In 2.2 the different aspects of this utterance with regard to our four categories are presented.

Type	**Description**
Intention based	$(Intend_{Carl}(Believe_{Mother}(Playing_{Peter})))$
Semantically based	Play(peter).
Textually based	Peter is playing.
Conversationally based	eee (0.2) mum (0.3) peter is playing.

Table 2.2: Different aspects of a discourse segment containing "Peter is playing".

In the first row we find the intention based segment which stresses the discourse relevant goal of the utterance. In the second row the notion of the proposition is found, and in the third row a written sentence capturing the utterance. In the last row a CA notion of the utterance is given, focusing on the precise surface form of the utterance.

The type of discourse segment which best suits our study is the intention based segment, since this allows more variation in the surface form and is applicable to both monologue and dialogue. Based on this it was decided to use Grosz and Sidner's (1986) discourse theory as a theoretical framework for how discourse segments are constituted. The next section contains a more detailed description of the different aspects of the discourse segments in the view of Grosz and Sidner (1986).

2.4 A Closer Look at One Theoretical Approach to Discourse Structure

For us, one of the clear advantages with Grosz and Sidner's (1986) theory is that it is applicable to both monologue and dialogue. In addition, it has an elaborate description of a number of aspects of the discourse segment. These are the main reasons why we

have decided to use Grosz and Sidner's (1986) discourse theory. In this section a more detailed description of the theory follows.

2.4.1 The Grosz and Sidner 1986 Discourse Theory

Grosz and Sidner (1986) provide a framework for describing the processing of utterances in a discourse. In their view a discourse consists of three distinct but interrelated components. The theory is primarily intended for dialogue, but holds also for monologue, including both scripted and non-scripted language. It is intention-based, the intention or purpose behind a discourse segment playing an important role for how one segment is related to other segments, and how the discourse evolves. The components they suggest as constituting the discourse are:

- The linguistic structure.
- The intentional structure.
- The attentional state.

The *linguistic structure* is described as being the structure of the sequence of utterances. It consists of segments of the discourse into which utterances naturally aggregate. The *intentional structure* is the structure of purposes capturing the discourse-relevant purposes expressed in each of the linguistic segments as well as in their internal relationshipships, i.e. the structure of intentions underlying and motivating each discourse. Each discourse segment (containing a portion of the linguistic structure and thus a specific string of words) is paired with an intention underlying and motivating the specific string of words. The *attentional state* consists of a list of all concepts in focus of attention in a specific discourse segment. We illustrate the idea with the three components in example 2.1 and table 2.3 below.

The speaker initiating the discourse is called the Initiating Conversational Participant (ICP), while any other speaker is called Other Discourse Participant (OCP). Example 2.1 shows a possible *linguistic structure*, i.e. the segments of the discourse into which the utterances naturally aggregate, for a short discourse between two speakers:

(2.1) **The linguistic structure of a short discourse**

(ICP) "Excuse me, do you know what time it is?"

(OCP) "Yes, three o'clock."

(ICP) "Thank you!"

As mentioned, the linguistic structure has an accompanying intentional structure, the structure of the underlying intentions. These intentions constitute the motivation for how

to segment the linguistic structure, as well as the basis determining how the discourse segments are arranged and connected to each other. In addition, each discourse has a discourse purpose, i.e. a kind of initial state which holds for the entire discourse and not just for a single discourse segment. An intention is open as long as it is not satisfied, i.e. if the intention is to ask a question, this intention is open until an answer to the question is provided. Thus, the intentions can be viewed as goals which have to be reached in order to let the discourse evolve. If a question is not answered, but instead is met with a counter-question (with its own motivating intention), the second intention is pushed onto the first, thus forming a stack. As an intention is satisfied, e.g. a question is answered, it is removed from the stack. The relationship between the segments of the linguistic structure is based on the relationship between the intentions recorded in each sentence in the discourse. Thus, establishing and satisfaction of the underlying intentions govern how the discourse structure evolves.

Let us continue with the example of the linguistic structure presented in example 2.1 and add intentions to the discourse (intention 0) and also to each segment in the short discourse (intentions 1-3).

Intention number	**Speaker**	**Linguistic structure**	**Formulation of intention**
Intention 0 *(Discourse intention)*	A (ICP)	*Purpose of opening the discourse.*	Find out what time it is through a conversation.
Intention 1 *(Segment intention)*	A (ICP)	"Excuse me, do you know what time it is?"	Get B (OCP) to tell what time it is.
Intention 2 *(Segment intention)*	B (OCP)	"Yes, three o'clock."	Answer A (ICP)
Intention 3 *(Segment intention)*	A (ICP)	"Thank you!"	Acknowledge that B (OCP) answered the question, finish the conversation

Table 2.3: Example of the linguistic structure together with the intentions for each segment.

In table 2.3 we find a record of the *intentions* with their accompanying segments from the linguistic structure. In the left hand column (Intention number) we have enumerated the segments, in the next column we show information about the speaker, in the third column (Linguistic structure) we show the segment of the linguistic structure corresponding to the intention on the same row (column Formulation of intention). In the case of intention 0, no such verbal segment is present, since the intention is the initial state which holds for the discourse as a whole, and not for only a single segment. Lastly, in the right hand column the declared intention is shown. In this example each intention overlaps

one-to-one with the speakers' utterances which constitute the segments of the linguistic structure. Such a one-to-one mapping is not always the case; one intention can also be realized with two or more utterances.

We have now made a segmentation of the linguistic structure into discourse segments and labelled each segment with the intention governing it. The structure in the table looks rather linear, but the Grosz and Sidners' model allows a hierarchical structure created on the basis of how intentions relate to each other. So, let us consider the hierarchical form of the short discourse in table 2.3.

The discourse starts with intention 0 – the intention to find out what time it is. Intention 0 is not satisfied with intention 1 (make B tell what time it is) and 1 is thus pushed onto a stack together with 0. When intention 2 (Answer speaker A) is carried out, intention 1 (make speaker B tell what time it is) is satisfied and intentions 2 and 1 are removed from the stack, but intention 0 is still there. When intention 3 is carried out, intention 0 is satisfied, the whole stack is removed and the conversation/discourse finished. Thus, in intention 3 the goal for the whole discourse is achieved, and both speakers involved understand this; intention 0 is satisfied and the discourse is finished. The intentional structure is a picture of how the underlying intentions are related to each other, and how they structure the discourse in a hierarchical fashion.

We now turn to the third component listed above, the *attentional state*, which is a dynamic record of objects, properties and relationships which are salient at each point of the discourse, i.e. for each discourse segment a specific record of attentional state is present. We can think of the attentional state as a concept present either implicitly or explicitly in the combination of an intention and the accompanying portion of linguistic structure. For example, concepts such as TIME-NOW, WATCH and CLOCK can be assumed to be in the attentional state at intention 1 in the discourse in example 2.3 above ("Excuse me, do you know what time it is?"). This is because speaker A asks for the TIME-NOW and speaker B can use a watch or a clock (or any time-measuring instrument) to answer B.

The object of the analysis in Grosz's and Sidner's discourse theory is the discourse itself, i.e. not the speaker's understanding or interpretation of the discourse. In the example it was shown that the they treat the speaker's contributions in a dialogue as constituting one single and unified discourse, i.e. there is no place for two or more simultaneous interpretations of e.g. the motives behind one utterance or another. In the above discourse, and in most monologues, this simplification poses no problem, but in dialogue it does.

In this study there is a focus on Grosz's and Sidner's three aspects of discourse in relationship to the discourse segments: the linguistic structure, the attentional state and the intentional structure. In our study we relate the linguistic structure to the boundary annotation, the attentional state to the prominence annotation and the intentional structure to the specific question segments. The correspondences here are rather crude, but we consider that they are justified. Our justification for relating the lingu-

istic structure to the boundary annotation is that Grosz and Sidner (1986:176) describe the linguistic structure as "the structure of the actual sequence of utterances in the discourse".

We relate the prominence annotation to the attentional state since the attentional state is said to contain "information about the objects, properties, relationships and discourse intentions that are most salient at any given point." (Grosz and Sidner 1986:177). Admittedly, the attentional state contains a number of more abstract elements than elements prominent on the surface level of the discourse. However, we can assume that elements prominent on the surface level would also be included in the attentional state, which means that we will capture a subset. One might also argue that the attentional state should capture concepts already in focus, i.e. elements in psychological focus, and not prominent elements in general, i.e. elements in semantic focus. However, we consider it justified to relate the attentional state to prominence in general, since it is unclear exactly how the introduction of elements and the segmenting interplay.

The intentional state is said to contain "Intentions of a particular sort and a small number of relationshipships between them" (Grosz and Sidner 1986:177). We relate the intention underlying an utterance to the structure.

2.5 Linguistic and Prosodic Indicators of Discourse Segments

In the study in this thesis no full discourse analysis of all the data is carried out. Instead a number of linguistic, prosodic and acoustic features relevant to boundaries or prominences are selected and examined across the speaking styles. In this section we survey features which have been pointed out as important in previous research and thus constitute the base for the selection of features in the present study.

In the same way as discourse segments can be seen from many perspectives, there are also many different cues as to how a segment could be detected and identified. For instance, the discourse segments can be signalled through the syntactic structure, but also lexically by cue phrases indicating segment boundaries, acoustically through prosody or through term repetition indicating the topic structure.

The sentence, clause and phrase are obvious portions of language, and the same holds for the prosodic phrases. It seems natural to think that the syntactic portions and the prosodic portions are just different ways of signalling the same segments. However, many researchers, e.g. Abney (1992, 1995), Bruce (1982, 1998), Selkirk (2000) have pointed out that prosodic phrasing and syntactic phrasing do not always coincide. Nevertheless, the two kinds of phrasing still show a close relationshipship, especially in read aloud speech (Bruce, 1998). It has also been suggested that the coincidence of prosodic phrasing and

syntactic phrasing increases the higher up in the syntactic tree the comparison is made, see e.g. Selkirk (1984).

The non-isomorphic mapping between what is signalled through the string of words and through the prosody implies that segments can be signalled either lexically, syntactically, prosodically or by a combination of these features. Which single features and which combinations have been identified as relevant for discourse segments and discourse structure?

2.5.1 Boundary Indicators

Boundary indicators in language have been investigated from many aspects. A crude distinction can be made between work aimed at finding lexically based boundary indicators, such as syntactic markers, pragmatic markers, discourse markers, discourse cues or cue phrases, see e.g. Halliday and Hasan (1976), Green (1979), Fraser (1990), Schegloff (1996) and Marcu (1997) and work aimed at finding acoustic correlates to discourse relevant prosodic phrases, see e.g. Lehiste (1979), Bruce (1982), Strangert (1990) Grosz and Hirschberg (1992), van Donzel (1999) Shriberg *et al.* (2000) and Hansson (2003). In addition, there are also combined approaches e.g. aimed at using acoustic cues to disambiguate a lexical cue as discourse relevant or not, see e.g. Hirschberg and Litman (1993), Ferrara (2001) and Horne *et al.* (2001).

Starting with lexical boundary indicators, many researchers have related specific lexical expressions to different types of boundaries, e.g. syntactic boundaries such as phrase, clause, sentence and paragraph boundaries as well as pragmatic boundaries such as (discourse) topic boundary. Different kinds of expressions function as boundary markers on different levels of analysis. Based on experiments with artificial languages Green (1979) has stressed the necessity of function words in sentence syntax. Based on these listeners make an incremental shallow parse based on syntactic markers signalling e.g. "new clause".

Moving to the discourse level of analysis we talk about discourse markers, discourse particles, discourse cues, pragmatic markers, pre-beginnings and cue phrases, all representing lexical markers structuring the discourse. Broadly speaking we can distinguish between markers on a slightly lower level, e.g. indicating RST-like discourse relationships between clauses and markers on a slightly higher level of analysis indicating e.g. topic change. Discourse markers, discourse particles, discourse cues and pragmatic markers are often used for the former category, whereas pre-beginnings and cue phrases can be used for the latter category. The markers in the first category often consist of one word or a shorter multi word unit, while in the latter one they often consist of longer phrases. However, there are no distinct boundaries between these levels. For instance, RST relationships can apply within a complex sentence, but also between sentences. Sometimes a longer cue phrase just signals a local discourse relationship, whereas in other

cases a more reduced expression might signal a topic change. The terms are sometimes interchangeable, but all these types of markers have in common that they are assumed to indicate some type of boundaries.

In table 2.4 examples of markers for different levels of analyses are given, however, please note that the levels are approximate and partly interfere with each other.

Different approaches to discourse markers		
Area	**Researcher**	**Example**
Syntax	Green (1979)	that for – to which
Coherence relationships	Mann & Thompson (1988) Marcu (1997)	but although since
Topic segments	Hirschberg & Litman (1993) Marcu (1997)	as a matter of fact summing up now
Conversational analysis	Fraser (1990)	eh is that right oho

Table 2.4: Examples of discourse markers

A computational approach to discourse segmenting with a focus on lexical boundary indicators was carried out by Marcu (1997) who used cue phrases in order to assign rhetorical relationships to English text in an RST fashion. The idea was to establish relationships between cue phrases and specific RST relationships in order to parse a text with regard to rhetorical structure. The segmenting was carried out on the basis of cue phrases and punctuation. Marco's (1997) results have been used in both (rhetorical) parsing, in summarization and in generation of text.

Many researchers have also tried to find correspondences between cue phrases and specific discourse relationships, see e.g. Marcu (1997), Stede and Umbach (1998), and Moser and Moore (1995).

A discourse parser primarily relying on discourse markers has also been developed by Webber and Joshi (1998), Webber *et al.* (1999). In their framework, discourse markers are interpreted as predicates on the discourse level, i.e. in this view the segments can be seen as the arguments, while the discourse marker functions indirectly as a segment marker in defining the argument it takes. On the surface level, however, the discourse

marker functions as a direct cue to the type of relationship and thus to the segment structure.

A different approach to discourse segmenting was taken by Hearst (1994) who carried out segmentation experiments based on term repetition. This segmentation resulted in larger paragraph-like topical units. Hearst (1994) defined the segments on the basis of a threshold value for the term repetitions within a segment. When the term achieved a frequency lower than the threshold value, a segment boundary was indicated. Thus, in this case the segmentation was performed from the aspect of the segment content and not from the aspect of the segment boundaries. Compared with other discourse segmenting approaches Hearst's (1994) segments are extensive.

In addition to the lexical cues as indicators of discourse boundaries, prosodic phrasing and its acoustic correlates have also been investigated. The relationship between pauses and syntactic boundaries has been pointed out by many researchers, among other e.g. Goldman-Eisler (1972), Lehiste (1979), Bachenko and Fitzpatrick (1990), Strangert (1992), Cutler *et al.* (1997), Bruce (1998), Hansson (2003). For Swedish, Bruce (1998) reports that changes in F0, i.e. a fall or a rise, can indicate a boundary, as well as creaky voice and/or longer segment duration (final lengthening). Cutler (1997:162) reports similar findings to hold also for English and Dutch. Results indicating a relationship between longer pauses and stronger discourse boundaries are achieved by many researchers, e.g. Strangert (1990, 1993) Bruce *et al.* (1993) and Hansson (2003) for Swedish and Grosz and Hirschberg (1992) and Swerts and Geluykens (1994) for English and Dutch respectively.

Strangert (1990) has studied the relationships between prosody and syntax in Swedish and found differences in the acoustic signalling of phrases, clauses, and sentences. The results indicated that stronger boundaries, e.g. sentences, had stronger acoustic correlates. Furthermore, Strangert (1992) reports indications from perceptual experiments that pause length might be an indicator of boundary strength, the longer the pause the stronger the boundary. In addition she argues that a pause in terms of a silent interval seems to be the stronger acoustic cue in cases of conflicts with other cues. Both findings are supported by Hansson (2003).

Grosz and Hirschberg (1992) have studied acoustic correlates to discourse segments in English news broadcasting, and found statistically significant associations between aspects of pitch range, amplitude and timing with features of local and global discourse structure (Grosz and Hirschberg, 1992). For instance phrases at the ends of discourse segments were followed by longer pauses than phrases ending inside discourse segments. Moreover, discourse segment initial phrases were uttered in a wider range than phrases located inside discourse segments (Grosz and Hirschberg, 1992). Thus, a difference was found between phrasing at discourse segment boundaries (global structure) and phrasing inside discourse segments (local structure).

Swerts and Geluykens (1994) have studied spontaneous monologue in Dutch and found that both speakers and listeners use prosodic cues in structuring the discourse. The

ends of larger discourse segments tended to end on a low tone, whereas the segment beginnings tended to start on a higher tone. Similar results were reported for Dutch by van Donzel (1999). The F0 reset as a correlate to a boundary is also mentioned by Bruce (1998).

Interestingly enough Swerts and Geluykens (1994) comment on the clarity of a message and its relationship to the speakers' production strategies. They found that for some subjects, but not for all, a low boundary tone was often present at discourse boundaries. In other cases the pause was a clear boundary signal for some subjects, but not for other ones, so they could not state any single signal as the most important boundary signal. Examining the data closer it was found that the subjects who were less clear in the production of the melodic signal, were instead clearer in the pause signal. In the case where a subject was clear with regard to both signals, the impression was of a very clear boundary marking.

Prosodic phrasing has been used as a correlate for discourse segment boundaries of the Grosz and Sidner (1986) discourse segment style, see e.g. Grosz and Hirschberg (1992), Hirschberg and Nakatani (1996), as well as for segments of dialogue act style, see e.g. Stolcke *et al.* (2000), Shriberg *et al.* (2000) and Levow (2004).

Experiments with automatic discourse segmenting including both boundary features and prominence features have been performed by (Passonneau and Litman, 1997). They tested both algorithms based on either pauses, discourse cues or referential noun phrases, and an algorithm based on all three features. The last algorithm, which used multiple sources of linguistic information, performed best.

To sum up: discourse segment boundaries are reported to correlate with both lexical and acoustic cues, and on analysis both types of cues are present on different levels in the discourse, from phrases to clauses, sentences, paragraphs and longer topical units. Discourse particles have been used successfully as segment indicators as well as specific acoustic correlates such as changes in F0, final lengthening and silent pauses. The last one, pausing, is claimed by many researchers to be the strongest boundary indicator.

2.5.2 Prominence Indicators

Prominence is the aspect of important content which is located in between the boundaries. Regarding lexicogrammatical cues to prominence, the most common one is that prominence is related to content words. Loman (1967) reports on a study of English where the most frequent position for the intonational centre is in common nouns or proper nouns, with finite verbs in the second place. In addition it is mentioned that most intonational centres are found in nominal constructions (60%), and (18%) are found in verb constructions.

In the view on prosodic phrasing the interrelationship between the prominence and the boundaries is clear. According to Bruce (1998) prosodic phrases are signalled not only by

the prosodic boundaries, but by a combination of the boundary signals with a presentation of what is between the boundaries. This view is also supported by Ostendorf (2000) who points out that the accent is an additional cue to the prosodic phrase structure.

There are many indications that discourse segment boundaries have acoustic correlates in prosodic phrasing. Since prosodic phrasing is also related to prominence (Bruce, 1998) it is not far fetched to assume that discourse segmenting also is dependent on both the boundary signals and the prominence signals. Thus, which are the prosodic and acoustic correlates to prominence?

Strangert and Heldner (1995) and Heldner (2001) investigated the relationship between focal accent and subjects' perception of prominence in Swedish. They found a significant positive correlation between focal accent rise and the perceived prominence of a word. Thus, as in the relationship between pauses and boundaries, the relationship between focal accent and prominence is of the type "the more the stronger". However, parallel to the observation that a pause does not have to be a boundary, Heldner (2001) also points out that an F0 rise is neither necessary nor in itself sufficient for subjects' perception of prominence.

The lexico-syntactical and lexico-grammatical parallel to acoustic prominence is captured in the information structure through means such as e.g. topicalization. Models of focality have been suggested by e.g. Sidner (1983) and Grosz *et al.* (1995) in order to predict lexico-syntactical and lexico-grammatical focality.

Bolinger (1972) argues that acoustic prominence in terms of accent mirrors the information focus which he claims is only indirectly related to the syntax, and that it thus is not possible to predict the accent on the basis of syntax. Nevertheless, many researchers have studied the correlation between information structure and prosody, e.g. Swerts and Geluykens (1994), Horne and Philipson (1995), van Donzel (1999), Wennerstrom (2001), and found correlations between information structure prominence, i.e. semantic focus, and focal accent.

Swerts and Geluykens (1994) report higher F0 peaks in topic-introducing noun phrases than in other noun phrases in Dutch. van Donzel (1999), also studying Dutch, investigated the relationship between prosody and new and given information in the taxonomy of Prince (1981). Words with the information status "new" were perceived as more prominent by subjects, and as the status "new" declined, the classification as prominent also declined (van Donzel, 1999). This was also found when the acoustic features of the words were examined. The words classified as new also had a stronger pitch accent, whereas a decrease in the newness corresponded to a decrease in the pitch accent (van Donzel, 1999). van Donzel (1999) concludes that the main cue to information status is the pitch accentuation.

However, there are also exceptions when given information becomes accented. Horne and Philipson (1995) investigated given elements which acquire acoustic prominence in Swedish, and found that an accent on a word representing given information

often corresponded to a change of grammatical role for the referent. In other words, the element is given, but the function is new.

Stifelman (1995) has investigated topic segmenting of English based on Arons emphasis algorithm (Arons, 1994). In this study emphasis on an element was taken to indicate the beginning of a new topic. Stifelman (1995) reports that an emphasis on high precision indicated a topic initial element, but there were also many topic initial elements which were not indicated by an emphasis. This was also pointed out by (Ayers, 1994). In Bruce *et al.* (1993) this phenomenon is discussed for Swedish, indicating that a phrase initial accent seems to have the same function as a prosodic phrasing. Thus, phrasing and accenting are closely interrelated.

Swerts and Geluykens (1994) have in studies on Dutch spontaneous monologue found that at the beginning of a unit a high-pitched accent on a topical noun phrase is often found. Swerts and Geluykens (1994) suggests that this could be a warning signal to the listener that a new information unit has started. In addition they report that both pauses and melodic cues are used as boundary signals, and that segment final cues, such as boundary tone and pausing, seem to be more important to listeners than segment initial, such as F0 reset.

This idea of interaction between the boundaries and what is between the boundaries is supported by Selkirk (2000), who has performed an optimality theory analysis of English aiming to map the prosody to either a more demarcative strategy or a more cohesive strategy. In a more demarcative strategy the relationship between prosody and syntax is more prominent, while in the cohesive strategy the relationship between phonology and the use of information structure is more prominent. This approach stresses the relationship between boundaries and prominence in the segmenting, together with the relationship to either syntax or information structure, and therefore has similarities to our study in this thesis. Selkirk (2000) further argues that a language opts for either a cohesive or a demarcative strategy. Hansson (2003) has made an optimality theory analysis of spontaneous southern Swedish, and her results indicated a cohesive strategy.

Putting together the different studies on boundaries and prominence, the clearest indications of discourse segment boundaries seem to be cue phrases and pauses. Between these boundaries there are islands of new information conveyed through accented noun phrases, where the most prominent acoustic cue seems to be pitch accent. The picture which evolves is that both the string of words and the prosody interact in the division of the discourse into segments. However, is this interaction constant across different speaking styles where the string of words and the prosody might differ?

To get a better picture of how the contributions from the string of words and the prosody can vary between different speaking styles we devote the next section to a survey of spoken and written language as well as to findings from scripted and non-scripted speech with a focus on Swedish.

2.6 Spoken and Written Language – Different Realizations of Discourse

Comparing written text to spoken language we immediately find differences. Written text can easily be segmented on the basis of syntax and punctuation. An attempt to perform the same type of segmentation on spontaneous speech immediately reveals large differences between spontaneous speech and written text, even though the speech may be transcripted. In spoken language there are features such as restarts, hesitations and constructions which the grammar book would judge as ungrammatical. In addition, there is no punctuation to indicate the structure. Instead, the spoken language has a range of acoustic means to signal features such as information structure and attitudes, which the written language has just crude means to communicate through e.g. underlining etc. (Enkvist, 1974:190).

Gumperz *et al.* (1984) stress the cohesive role of prosody in spoken language arguing that prosody in spoken discourse establishes cohesion in much the same way as complex syntax and connectives do in written discourse. Based on this, we might hypothesise a difference in the distribution of labour between prosody and syntax in different speaking styles.

This section contains a brief survey of textual differences between written and spoken language, and also introduces a way used to quantify one aspect of this difference – the noun-verb quotient. Moreover we briefly survey some acoustic differences between read aloud and spontaneous speech with a special focus on Swedish.

2.6.1 Differences Related to the Verbal Content

Many researchers have studied differences between spoken and written language from a range of different aspects, see e.g. Biber (1988), Chafe (1982 1993) and Tannen (1982) for English and e.g. Einarsson (1978) for Swedish. Some of the features which are said to characterize a written style are according to Biber (1988:47) that i) written language is structurally more complex in terms of longer sentences and greater use of subordination, ii) written language is more explicit than spoken language, meaning that idea units and logical relationships are encoded in the text, iii) written language is less dependent on shared situation knowledge and iv) written language is better organised and planned. Even though these very broad generalizations are not without their problems, see Biber (1988:48) for an elaborate discussion, they are still capturing a tendency of frequently observed differences between written text and more spontaneous speech.

Related to the above features, i.e. the longer and more complex clauses in written text and the shorter and less complex clauses in spoken language, is also the idea of *nominality* in text. Nominality is used by Wells (1960) to describe the difference between a style with a high proportion of nouns and long noun phrases as opposed to a style with a

high proportion of verbs and shorter noun phrases. Thus, some of the complexity and longer clauses might be due to longer noun phrases. The dimension nominal – verbal is described by Hellspong and Ledin (1997) as mirroring the density of information in a discourse. A nominal style is said to be declarative and have a higher density of information, while a verbal discourse is said to be reasoning and have a lower density of information (Hellspong and Ledin, 1997). The nominal style is further said to demand a greater amount of both planning (from the sender) and processing (by the recipient). We thus find nominal style in written text rather than in spontaneous speech.

However, the "nominal" or "verbal" style does not mean that the features "many nouns" or "many verbs" come on their own. They are connected to other lexicogrammatical characteristics in the discourse. Hellspong and Ledin (1997) have listed some characteristics for nominal and verbal style, which are shown in table 2.5:

Nominal style	**Verbal style**
Many nouns at the expense of verbs and pronouns	Many verbs at the expense of nouns
Many long noun phrases	Few long noun phrases
Many attributive preposition phrases	Many adverbial preposition phrases
Few subordinate clauses	Many subordinated clauses
Long sentences are mostly due to many coordinations and enumerations	Long sentences are mostly due to long prepositional adverbials and subordinated clauses

Table 2.5: Some characteristics of nominal and verbal styles. Adapted from Hellspong and Ledin (1997:78).

Wells (1960) suggested that the relationship between nouns and verbs should be used as a measurement of nominality in a text. He called the measurement *Noun–Verb Quotient (NVQ)* and it is computed by the number of nouns divided by the number of verbs ($\frac{Nouns}{Verbs}$). The higher the NVQ, the more nouns in the text, and the more nominal the style. This makes it possible for us to compare the proportion of nouns and the proportion of verbs in one measurement.

Wells' idea behind the NVQ measure was that all sentences can be expressed in either a verbal or a nominal form, e.g. "when we arrive" (verbal) and "at the time of our arrival" (nominal). As Wells himself points out, this view is extremely simplified, but still useful. It allows us to compare the degree of nominal and verbal styles as two abstract extremes in texts. In the typical example of nominal style we find simple sentences elaborated with long noun phrases ("at the time of our arrival), while verbal style has many compound sentences and shorter noun phrases ("when we arrive"). Wells meant that these syntactic properties also implied other features connected to nominality: prepositions and adjectives often combine with nouns while conjunctions and adverbs combine with verbs.

In a more elaborated description of *Noun Quotient (NQ)* described by Melin and Lange (1986 2000) prepositions, participles, pronouns and adverbs are also used. The formula is given in equation 2.2. We see that nouns, prepositions and participles are regarded as "nominal" while verbs, pronouns and adverbs are regarded as "verbal".

$$\text{NQ} = \frac{Nouns+Prepositions+Participles}{Verbs+Pronouns+Adverbs} \quad (2.2)$$

Biber (1988:57) stresses the importance of good sampling when you attempt to compare spoken and written discourse. Discourses differ not only in the dimensions spoken and written, and Wells (1960) also suggests the dimensions i) Interactive – Edited text, ii) Situated versus Abstract content and iii) Reported versus Immediate. The first of these three dimensions, Edited – Interactive, is linguistically characterized by features such as longer words and more varied vocabulary (Edited) versus questions and first and second person pronouns (Interactive). The second dimension, Abstract versus Situated, is characterized by features such as nominalization and passives (Abstract) versus place and time adverbials (Situated). The third dimension (Reported versus Immediate) is characterized by past tense (Reported) versus present tense (Immediate) features. Comparing the three dimensions to the features included in the NQ measures shows that broadly speaking NQ would capture features from dimensions i and ii, however, not from dimension iii. Nevertheless, the NQ measure can be assumed to broadly index a discourse according to features generally regarded as frequent in written or spoken language, and thus roughly plot the discourse on a continuum between written and spoken discourse.

When categorising speaking styles the concepts of style and genre become slightly problematic. Instead the importance of communicative situations has been suggested. This means that the basis of the classification is not the content of the discourse itself, but the situation out of which the discourse originates. The importance of the communicative norm, and how it differs between different speech situations – or different speech *activities* is stressed by Allwood (1995), who suggests an activity based approach to speech, highly coloured by the social activity in which it appears. This approach highlights aspects of formality and informality between the dialogue participants.

Regarding different styles, there are indications on the lexicogrammatical level. We could assume that scripted speech would be more nominal than non-scripted speech. In addition, if the pattern indicated by NVQ and NQ holds we could assume differences in the part of speech distribution in the different speaking styles.

Can we expect similar differences in prosody?

2.6.2 Prosodic Differences Between Speaking Styles

Just as written discourse consists not only of prototypical text but also of e.g transliterated speech, spoken discourse consists not only of spontaneous speech. Spoken discourse

covers a continuum from spontaneous speech, produced at the moment of speaking, to read aloud speech where the verbal content can be carefully prepared long before the moment of actual speaking. Thus we can assume structural differences between different types of written discourses just as we can expect differences between speaking styles. Of special interest in this brief survey of prosodic differences in different speaking styles are differences related to the segmenting of discourse, in other words, differences related to the marking of boundaries and the bringing out of concepts.

Pausing is one of the stronger indicators of prosodic phrasing, and it has been observed to vary across speaking styles. Many researchers have reported a close overlap between pauses and syntactic boundaries in read aloud speech, see among others Goldman-Eisler (1972), Lehiste (1979), for English and Gårding (1967), Strangert (1990) and Bruce (1998) for Swedish. In spontaneous speech pauses are found also at other positions, see e.g. Gårding (1967), Strangert (1993) and Bruce *et al.* (1996) for Swedish.

Strangert (1993) has investigated the relationship between professional reading, non-professional reading and spontaneous speech in Swedish and found that professional reading has a higher articulation rate and shorter pauses which correlate with syntactic boundaries. Many syntactic boundaries did not have a correlate in silent pausing. In addition a specific kind of semantic pausing, i.e. pauses in front of semantically important words, were found. Non-professional reading was characterized by a slower articulation rate, longer boundaries, a more precise overlap of pauses and boundaries and a pause length which reflected the boundary strength. Spontaneous speech was characterized by a slow and hesitant style, a slow articulation rate, a great number of longer pauses, pauses both at syntactic boundaries and at other positions. Thus, a picture of spontaneous speech as slower and more hesitant and read aloud speech as faster and more direct is clearly evolving.

Studies of spontaneous dialogues and the read versions of the corresponding transcripts have been compared for English (Hirschberg, 1995) and for Swedish (Bruce, 1995), both studying dialogue. Hirschberg (1995) reported a difference in speech rate, e.g. the read speech was faster than the spontaneous one. In addition there were differences in the use of pitch contour in relationship to yes-no questions. These are often claimed to be uttered with a rising intonation which indeed also was the case in 55% to 80% of the cases in the read speech. However, in the spontaneous speech, in 43% of the cases the question were uttered with a falling intonation. Also Bruce (1995) reported differences in the pitch contours across the spontaneous and the read speech, most prominently a wider pitch range in the spontaneous speech.

Gustafson-Čapková and Megyesi (2001) report big differences in the pausing pattern between professional news broadcasting, non-professional reading and spontaneous elicited dialogue in Swedish. The study, carried out in the form of a perception experiment, shows that in professional announcing all silent pauses were perceived as boundaries, while many of the perceived boundaries were without a correlate in a silent pause. In

non-professional reading the majority of pauses were perceived as boundaries, and at the same time many of the perceived boundaries also had a correlate in a silent pause. In the dialogue many of the pauses were not marked as boundaries, but many of the boundaries had nevertheless a correlate in a silent pause. Similar results were also reported in Gustafson-Čapková and Megyesi (2002) and Megyesi and Gustafson-Čapková (2002).

In a careful study of prosodic features in spoken and written Swedish, Fant et al. (1989, 2000) report that pause durations differ between different kinds of boundaries. Pause duration ranges from between 50-100 milliseconds for short prompters to 1-2 seconds between sentences. Normal pause duration within sentences normally ranges between 300 to 600 milliseconds. Moreover, Fant et al. report that pauses at sentence boundaries have a relatively longer duration while final lengthening is more frequent at phrase boundaries.

There are fewer studies on differences with regard to prominence in read and spontaneous speech, and there are not such clear differences reported with regard to boundary marking in written and spontaneous speech, however, there are still indications of differences. For instance, Cutler (1997) reports from a study on English that a bigger difference between stressed and unstressed syllables, which is found in spontaneous speech, facilitates phoneme detection in the stressed syllables. However, such an effect was not found in laboratory read speech. This indicates differences in the strategies across different speaking styles.

Studies focusing on the connection between the information structure and the prosody have in many cases stressed the relationship between acoustic prominence and new information, semantic focus or topic, see e.g. Terken and Hirschberg (1994), Ayers *et al.* (1995), Horne and Philipson (1995), van Donzel (1999). However, the property of e.g. givenness does not change between spoken and written discourse because it is an information status, and not a syntactic feature. Thus, givenness relates to the same status in both written discourse and in spontaneous speech, which means that it is difficult to trace differences between speaking styles. Since the phrasing is closely related to syntactic boundaries differences between speaking styles are more visible.

Summarising the differences across speaking styles, the differences concerning boundaries are clearer than the differences concerning prominences.

2.7 Research Questions

It has been suggested that a discourse can be viewed as consisting of salient islands of content and between them boundaries which separate them from each other. Discourse segments can thus be viewed as consisting partly of the salient content and partly of the boundaries. The main research questions in this thesis concerns if, and in that case how, the discourse segmenting with regard to boundaries and prominence is affected by i) variation in the structure of the string of words (e.g. written – spoken) and ii)

variation in the prosody (e.g. access to prosody – no access to prosody). Does the string of words or the prosody contribute more to either aspect in one speaking style or another? Thus, we can formulate two questions: 1) How is the labour in the segmenting distributed between the string of words and the prosody? For instance, do subjects rely more on lexical cues than on prosodical cues in boundary annotation? and 2) Does this distribution of labour differ across speaking styles? For instance, if subjects rely more on lexical cues for boundary annotation in scripted speech, do they rely more on prosody in non-scripted speech?

A potential interplay between the string of words and the prosody is hypothesized, and the assumption is that if one dimension is varied, for instance if we vary the speaking style, a corresponding variation appears in another dimension, e.g. in the prosody. In effect, this means that the study compares selected acoustic variables across speaking styles. Investigations into both prosodic phrasing and prominence have reported that both aspects are characterized by a bundle of features rather than by one specific feature which is both necessary and sufficient. Perhaps the relationship is the same between speaking styles.

The third question concerns the theoretical relationship between prosody and discourse structure. How can our results be captured in a theory of discourse? In the design of the experiment as well as in the analysis, the framework of Grosz and Sidner (1986) is chosen. The aspect of boundaries is related to the linguistic structure and the aspect of prominence to the attentional state. If for instance prominence seems to play a larger role under some conditions, we might take this as evidence that also the attentional state plays a larger role in this condition than in another. Lastly a limited investigation of question segments related to the intentional state is carried out.

To sum up, we have three core questions concerning discourse structure. The first concerns the relationship between discourse structure as signalled by the string of words and by the prosody. The second question concerns how the relationship between the string of words and the prosody differ across different speaking styles, and the third question concerns the modelling of different speaking styles in a discourse theory. The core research questions are thus:

- Given a spoken language message, what does prosody contribute and what does the string of words contribute to the discourse structure in terms of "boundaries" and "prominent parts" in the the discourse segments?
- Does the interaction between the string of words and the prosody differ across speaking styles?
- How can we model differences between speaking styles including differences within the relationship between the strings of words and the prosody in a discourse theory?

To study the above questions, an experiment was carried out where subjects annotated boundaries or prominence in different speaking styles. Each of the two tasks in the experiment was carried out in two conditions: 1) with access to the speech signal and 2) without access to the speech signal. Each task as well as each experiment condition was carried out by separate groups of subjects. Investigations were made to see if the subjects' annotations converged, and if so, in what way they converged in the different speaking styles. In the next part of the thesis the experiment design and the results from the experiment are reported.

Method

Chapter 3

The Materials and the Design of the Experiment

DISCOURSE varies across speaking styles with regard to both lexicogrammatical structure and prosodic features. No single feature seems to be determinant but rather, there is a configuration of features. While listening to any spoken discourse the listeners have access to two main factors which contribute to their perception of structure in the discourse. Firstly the string of words, contributing with a syntactic structure as well as a focus of attention signalled through information structure features. Secondly, the prosody which contributes with prosodic phrasing and acoustic prominence. Do subjects rely more on one signal or another one in a certain speaking style?

In our study we approach discourse segmenting from two aspects: boundaries and prominence. The study is experimental and takes the listener's perspective; the procedure is to have inexperienced subjects annotate either boundaries or prominence in different speaking styles. Our aim is to study how the string of words on one hand and the prosody on the other contribute to the discourse structure. Therefore we need to distinguish between these two components and examine on one hand prosody and on the other structural differences in the string of words.

Our approach to fulfilling the above requirements was to set up an experiment where we had different groups of subjects annotate a number of different speaking styles, either with or without access to prosody. We had four separate subject groups: two working on a boundary annotation task and two working on a prominence annotation task, each of the tasks in two conditions: with or without access to the speech signal. Consequently, the experiment became rather elaborate, and we devote this chapter to a detailed description of the design of the experiment as well as to the discourse data and the subjects. We have divided the description of the design of the experiment and materials into four sections.

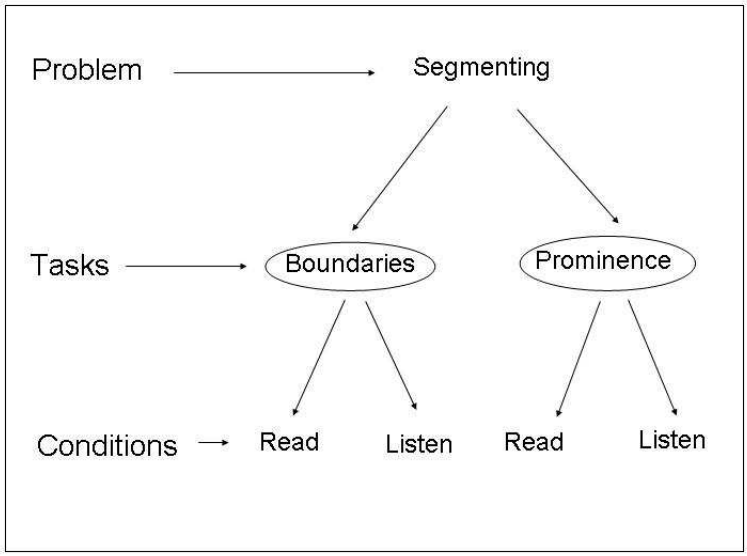

Figure 3.1: Architecture of the experiment.

The first section, section 3.1, contains a general overview of the architecture of the experiment. In the second section, 3.2, we describe the construction of the speech sample. In section 3.3 the subjects' annotation tasks are described, and the last section, section 3.4, contains brief information about the subjects.

3.1 Design of the Experiment

The aim is to study how on one hand prosody and on the other the string of words affect a listener's chunking of a discourse as well as their annotation of prominence. In figure 3.1 a schematic picture of the architecture of the experiment is presented. We can define the experiment as consisting of one problem, the segmenting problem, which is divided into two tasks, the boundary annotation task and the prominence annotation task. Specifically, each task is performed in two conditions: condition *Read*, where subjects annotate without access to the speech signal, and condition *Listen* where subjects make annotations with access to the speech signal.

The segmenting problem is studied in four different speaking styles. Thus, the materials are the same in all four conditions. This implies that all four separate subject groups annotate the same materials from the four speaking styles.

The four specific speaking styles range from scripted (read aloud) monologue to non-scripted (elicited spontaneous) dialogue. In this way we have included variations between

the different speaking styles in our material, and that enables us to study differences across speaking styles. Since our experiment also approaches the segmenting problem from two aspects represented by the two tasks, Boundaries and Prominence, it is in addition possible for us to study if subjects tend to agree more on e.g. the prominence annotation in one speaking style compared to another, as well as under one condition compared to another. For instance, do subjects agree more about where to annotate boundaries in one speaking style than in another?

The division of the materials into the conditions Read and Listen helps us control for the effect of prosody on the subjects' annotation of the speech materials. This allows us to study to what extent the string of words and the prosody respectively contribute to on one hand the annotation of boundaries and on the other the annotation of prominence. For instance, do subjects agree more about where to insert boundaries when they are allowed to listen to the speech signal than when they insert these boundaries solely on the basis of the transcripts? If there is such a difference, we can say that the prosody influenced the subjects towards a more uniform interpretation of the structure.

The variety of speaking styles allows us to study whether subjects approach the boundary and the prominence annotation tasks in different ways across speaking styles. For instance, do we find clearly different segmenting criteria in the read aloud speech as compared to the elicited dialogues?

The experimental procedure was the same in each condition (Read – Listen) of the two annotation tasks (Boundaries – Prominence): in each experimental condition inexperienced subjects made annotations with pencil in transcripts of the speech material (and in condition Listen they had access to the sound files). The subjects were not restricted regarding a time limit or a specific place. They could carry out their tasks in any place they liked, and they could take the time they needed. Together with the transcripts (and access to the speech signal in condition Listen) the subjects received the experimental instructions. Details about the instructions as well as of the speech sample and the transcription method are given in sections 3.3 and 3.2.

We considered to have the subjects make the annotations directly into a text-file on a computer, but we judged this procedure too complicated. Consequently we chose the procedure with pen and paper even though the computerised approach would have been more efficient in the organisation of the data.

It is important to stress that the use of subjects' annotations implies that the experiment captures the subjects' interpretation of the segmenting task. Thus we can relate the results to the understanding and interpretation of spoken Swedish. However, the results from the annotation cannot be related to the aspect of speech production, i.e. the experiment will capture preferences concerning how to interpret certain features, but not preferences concerning how to produce them.

We have related our study to the discourse theory of Grosz and Sidner (1986). This means that we make use of the three components of discourse that they suggest: the

linguistic structure, the attentional state and the intentional structure. We relate the investigation of boundaries to the linguistic structure since this is the surface structure of the discourse segments, i.e. the string that gets segmented. In addition, to this structure we relate the feature of pausing, since it is said by many researchers that this is one of the strongest prosodic boundary indications.

The study of prominence is related to the attentional state since this component keeps a record of concepts in focus. We have related the prosodic feature of accent to this component, since accent is the strongest prosodic feature for prominence.

A small study of question segments is related to the intentional state since a question can be taken as an example of a simple intention. Here we have not related to any specific prosodic feature, but have instead made an analysis of the general mark-up and annotation of the data.

Focality plays an important role in many approaches to discourse. Information about concepts in focus is important when the relationship between different discourse segments are to be determined. If discourse prominence, and thus e.g. focal accent, is a feature closely related to the discourse boundary, as well as to the realisation of the string of words in the discourse (i.e. connected to different realisations of discourse such as written text or spontaneous speech) it could be assumed that together with pauses this affects the discourse structure. Thus: how could the relationship between the realisation of the string of words and the prosody be captured in a discourse theory?

3.2 Collecting the Speech Sample

All annotations were done in transcripts of materials from four speaking styles. In this section we account for the sampling of the speech materials as well as for how it was transcribed. In addition we describe the final compilation of the samples of the speaking styles. We start with a survey of the speaking styles and the speech materials.

3.2.1 Selecting the Speaking Styles

All materials in our study are from spoken language. This enables us to have all materials in one condition "without sound" (transcripts only – condition Read) and "with sound" (transcripts and sound files – condition Listen). However, how shall we sample our spoken language data to include enough variety?

One of the most salient differences in spoken language is the difference between monologue and dialogue. Materials from both monologue and dialogue are included in order to capture this variation between non-interactive and interactive speech.

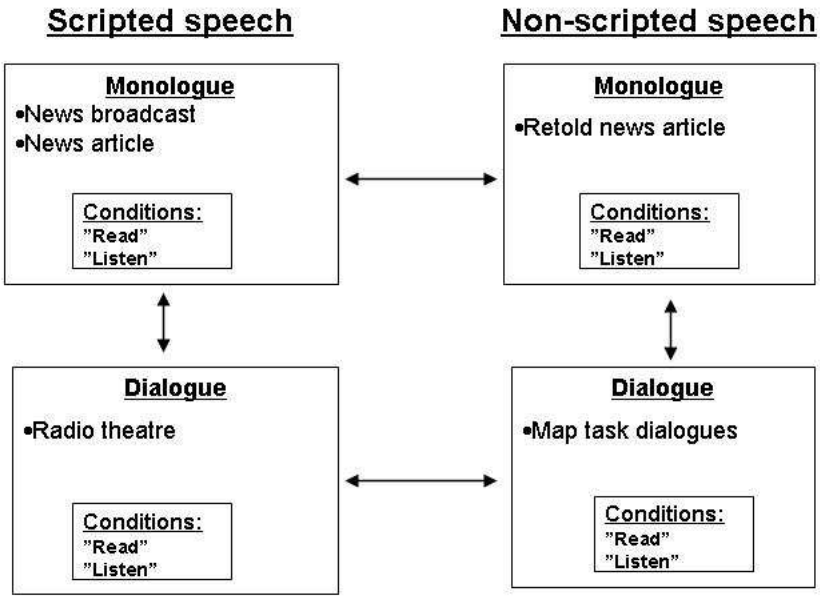

Figure 3.2: The speaking styles in the experiment.

Another prominent difference is the one between text read aloud and spontaneous speech. Since it is difficult to give a precise definition of when a speaker could be judged as "spontaneous" or "non-spontaneous", we chose, as previously mentioned, instead to use the distinction between scripted and non-scripted speech.

The selection of materials from interactive and non-interactive as well as scripted and non-scripted gave us materials from four categories:

- Scripted monologue
- Non-scripted monologue
- Scripted dialogue
- Non-scripted dialogue

A schematic picture of the materials is shown in figure 3.2. In each of the tasks (Boundary annotation and Prominence annotation), there are materials from all four speaking styles. Consequently, the whole experiment consists of two sets of materials shown in

figure 3.2, one for the boundary annotation task and one for the prominence annotation task. Each speaking style is described in greater detail in section 3.2.2.

Figure 3.2 shows schematically how we can compare the subjects' annotations of the materials one to another. We can compare them along the dimension of scriptedness and non-scriptedness (horisontally, within e.g. monologues), or along the dimension of interactiveness and non-interactiveness (vertically, within e.g. scripted speech). In addition, since all speaking styles are annotated in both condition Read and condition Listen we can trace the effect of prosody in each speaking style, e.g. within scripted monologue. Thus we can trace differences in subjects' segmenting of the materials as due to the prosody, as well as due to the string of words.

3.2.2 The Speech Recordings

Since our aim is to make comparisons between different speaking styles, the materials have to be comparable, with regard to both content and speakers. This means that the materials should ideally be comparable in content and as far as possible have the same speakers across speaking styles. Moreover, the recordings should be of a quality good enough for an acoustic analysis. In this section we describe in detail the materials with regard to speech type (monologue or dialogue), speech style (news broadcast, retold text etc.) speakers (male, female), recording and, where possible, the degree of formality and the speaker's attitude to the performance.

Scripted Monologue

The scripted monologue consists of two transcripts of news broadcasts by the Swedish radio (Dagens Eko), and one news article from the Swedish newspaper Dagens Nyheter. Each of the three materials were recorded with three speakers; two female (speakers a and b) and one male (speaker c). All three speakers were staff at the Department of Linguistics at Stockholm University.

For the news broadcast material, the speakers were instructed to read formally, as if their news reading was being broadcast, i.e. the intended listeners were a large group of people whom the speaker did not know. For the newspaper article, the speakers were instructed to read formally, but for a narrower group of people, such as their colleagues in the Department of Linguistics. For both types of material the speakers were instructed to talk as if they were reading to persons they could not see, and who could not see them. This made the speaker aware that they would not be able to communicate with the listeners with gestures or eye contact. The recordings were made in a soundproof room in the phonetics laboratory at the Department of Linguistics, Stockholm University.

Non-scripted Monologue

The non-scripted monologue consists of a retold news article from the Swedish newspaper Dagens Nyheter, the same news article that the speakers earlier read aloud. The speakers are the same three as were used for the scripted speech (speakers a, b and c), and they were instructed to talk as if they were re-telling the content of the news article to a colleague.

Also in this recording they were instructed to speak to a listener that would not see them, and that they would not be able to see. The recordings were made in a soundproof room in the phonetics lab at Department of Linguistics, Stockholm University.

Scripted Dialogue

The prototype of a scripted dialogue is a theatre play. It was not possible to ask the speakers from the recordings of the monologues to perform this task, so another solution was found; a recording of a play from the Swedish Radio, performed by the Radioteatern, was used. The play was selected to be as similar as possible to the non-scripted dialogues; thus, there should be two dialogue participants and the dialogue should be as task-oriented as possible. The play "Själen och Sankte Per" ("The Soul and Saint Peter") (Grimsrud, 2002) was chosen. This play is a dialogue between a newly deceased soul and Saint Peter, and together they are trying to make the soul remember if it was a man or a woman when it was alive.

The speakers are two actors, one male and one female (speakers e and f). The recording was obtained on a CD from the Swedish Radio, and it was converted into .wav format in the phonetics lab at the Department of Linguistics.

In the scripted dialogue we did thus not have access to the raw recordings and we did not know to what extent the radio play was edited. This is, however, not a problem since we primarily take the listener's perspective.

Non-scripted Dialogue

The non-scripted dialogues consist of spontaneous elicited speech in the form of Map Task dialogues. There are four speakers, three female (speakers a, b and d) and one male (speaker c), and three of those speakers (i.e. speakers a, b and c) are thus the same as for the monologues. The four speakers are divided into pairs, and each pair has two dialogues each, i.e. a total of four dialogues.

A Map Task dialogue is a way to elicit spontaneous speech, and is performed as follows: each of the two speakers in the dialogue has a map. One of the speakers has a path drawn on his/her map, and the task for this person (the "instructor") is to instruct the second dialogue participant (the "follower"), who does not have a path drawn on his/her

map) to follow the same path. As help both participants have landmarks on their maps. However, in order to elicit more dialogue the landmarks differ slightly between the maps.

The recordings of the Map Task dialogues were done by Pétur Helgason (Helgason, forthc) at the Department of Linguistics in a sound treated room, with the speakers sitting facing away from each other at a distance of approximately two meters (Helgason, 2002).

The data was sampled in such a way that the speakers in each speaking style were aware that a potential listener should be able to perceive the message solely through the audio channel. This means that the speaker should not have access to visual communication with the listener. With a sampling like the one described here, we can exclude a multimodal analysis of the data, and focus on the textual-acoustic dimension.

The script-based speaking styles are both directed at a large audience (radio broadcast), however, the news article (DN) has a somewhat smaller intended audience. The non-scripted speaking styles are both directed at a colleague.

3.2.3 The Compilation of the Speech Sample

The original recordings consisted of more than three hours of speech. Since an experimental approach with subjects who should carry out annotations of the materials was chosen, a smaller speech sample had to be constructed. For the dialogues, excerpts from each dialogue were selected, so that the overall time of scripted and non-scripted dialogues was the same. This means that the non-scripted dialogues consist of four excerpts, one from each dialogue, whereas the scripted dialogue consists of just one longer excerpt. In the scripted monologue each of the materials, i.e. the two radio broadcast transcripts and the newspaper article, was constructed from excerpts from each speaker. Thus, in the first news broadcast the beginning was read by speaker a, the middle part was read by speaker b, and the end was read by speaker c. The procedure was the same for the second radio news broadcast, as well as for the newspaper article, but the order of the speakers was changed. In 3.1 an overview of the sample is shown.

In the first column of table 3.1 we find the communication type, monologue or dialogue. The second column shows the speaking style, scripted or non-scripted, and the third column shows the type of speech material. The fourth column shows the duration of the sound file, given in minutes and seconds, the fifth column shows the number of words in the file, and the sixth and last column shows the speaker's identity. The duration and number of words are also given for each type of material, each speaking style and each communication type. The whole sample has a duration of 1 hour, 3 minutes and 34 seconds, and the total number of words is 9522.

The range of the excerpts was based on time, and not on the number of words, which means that a person who is speaking faster produces a larger number of words. For instance, speaker c was in general speaking very fast compared to speaker b, and an

Communi-cation type	Speaking style	Type of material	Sound file duration (min.sec)	#words	Speaker
Monologue duration 33.27 #words 4957	Scripted duration 18.49 #words 2872	Read, News 1 duration 5.50 #words 900	1.48	257	a
			2.00	291	b
			2.02	352	c
		Read, News 2 duration 6.45 #words 982	2.33	346	a
			2.09	272	b
			2.03	364	c
		Read, News DN duration 6.14 #words 990	2.37	364	a
			1.50	291	b
			1.47	335	c
	Non-scripted duration 14.38 #words 2085	Retold, News DN	9.26	1339	a
			3.08	464	b
			2.04	282	c
Dialogue duration 30.07 #words 4565	Scripted duration 14.47 #words 2245	Radio theatre	14.47	2245	e, f
	Non-scripted duration 15.20 #words 2320	Dialogue 1	3.28	515	a, c
		Dialogue 2	3.56	601	b, d
		Dialogue 3	4.00	529	b, d
		Dialogue 4	3.56	122	a, c

Table 3.1: Overview of the speech sample.

excerpt of two minutes from each of those two speakers would result in a greater number of words from speaker c. The excerpts were cut at positions where a break (i.e. a change of speaker) would pass as smoothly as possible, e.g. at a new paragraph etc. This implies that the speech excerpts are not of exactly the same duration.

3.2.4 The Transcription of the Materials

In order to enable a comparison of the results from the subjects' annotations across conditions and speaking styles, the annotations must be made on a comparable basis. This means that the transcripts of one speaking style must be identical in condition Read and condition Listen, and the form of the transcripts of the speaking styles must be as similar as possible. Further requirements are that the transcriptions should be easy to

read for the subjects so that the reading process should interfere the least possible with the annotation task. Since the materials would be tagged and parsed, the transcripts should also be suited as input to the tagger and parser. What transcription method would be optimal for these requirements? We devote this section to a description of the work on finding the most useful transcription form.

We first considered using the original texts for the written language styles, and transcriptions for the non-scripted parts. The idea behind this was to change the original data as little as possible. The texts, which were masters for the scripted speaking styles, all came from different sources. The news broadcasts came in text files, which were orthographic transcripts of news broadcasts from the Swedish Radio. The news article came as a word document which was retrieved from the Dagens Nyheter's internet-based news article service, and the radio play came in the form of a manuscript with stage directions. In a similar way, the non-scripted speaking styles came in different formats. The non-scripted monologues were present only in the form of sound files, while the non-scripted dialogues had a transliteration made for research purposes. Such a variety of formats would, however, give us a range of very different representations of the speaking styles, thus infringing on the requirement of as similar representation as possible.

For a more uniform transcription of the materials, there is a range of choices. Consider the phrase "Och det här är dagens eko kvart i fem" ("And this is today's echo, at a quarter to five") from the news broadcast. To transcribe it we could for instance choose between a phonetic transcript (example 3.1), a phonological transcript (example 3.2), a transcript with modified standard orthography (example 3.3) or a transcript based on standard orthography (example 3.4).

(3.1) [ɔdɛhˈæːʐɛdˈɒːgɛnsˈeːkʰɔkʰvˈaʈifˈɛmː]

(3.2) /ɔ dehæːr ɛ dɒːgens eːku kvartifemː/

(3.3) o(ch) de(t)här ä(r) dagens eko kvartifem

(3.4) Och dethär är dagens eko, kvartifem.

It would be possible to use the alternatives phonetic (example 3.1) and phonological (example 3.2) transcriptions for all four speaking styles. However, to an untrained eye they are not easy to read, and they are not suitable for the tagger and parser we planned to use. Thus, these alternatives infringe on the requirement of as small a reading effort as possible for the subjects, and as suitable an input as possible for the tagger and parser. In addition, the transcription itself is not a trivial task. With regard to these considerations we chose not to make use of phonetic or phonological transcriptions.

The modified standard orthography (example 3.3) could also be used for all four speaking styles, and it has the advantage of being fairly easy to read compared to the

phonetic and phonological transcripts. However, it still differs from traditional orthography e.g. in the use of parentheses. It has the disadvantage of being less suitable for the tagging and parsing, and it also requires more effort in the transcription phase. We considered that the alternative which would be most accessible for the subjects and at the same time most suitable as input for the tagger and the parser would be standard orthography (example 3.4). This alternative has the disadvantage of being less suitable than the other alternatives for transcribing spoken language, but we considered that the advantages outweighed the disadvantages. The transcripts could then be compared to spoken language as it is expressed in a written book. In our view this solution made the transcript reading the least intrusive for the subjects.

When it comes to the spelling we have in general conformed to standard Swedish spelling conventions, but in some cases we have used a more spoken language representation. In general we used standard orthography, as long as the speaker did not speak in some specific non-standard manner. This means that the word "är" ("be"), often pronounced [ɛ], is transcripted "är", unless it is not very clearly pronounced [e]. In that case it is transcribed "e". Sighs, hummings, etc are also orthographically transcribed.

During the reading all speakers made some small reading mistakes or hesitated. Thus, the original master texts had to be edited in order to include these phenomena. Otherwise, the transcripts would not correlate to the sound files. This would in turn give us unreliable results, since on one hand the subjects in condition Listen might have wanted to annotate e.g. a boundary in a stretch of discourse that was not present in the transcript. On the other hand the subjects in condition Read would not have access to the same string of words as the subjects in condition Listen. Example 3.5 shows a post-edited part in the read aloud speech before and after editing, the edited passages underlined.

(3.5) **ORIGINAL**

Det var den jagande apans behov av samordning som gynnade språkets framväxt. När våra förfäder skulle lägga ner stora villebråd krävdes planering, kanske över flera dagar. Den som då kunde beskriva och förstå ställen utöver "här och nu" fick ett naturligt försprång.

English translation:
It was the hunting apes' need for coordination which promoted the development of language. When our ancestors would bring down big game planning was needed, perhaps taking several days. Then, the person who could describe and understand places beyond "here and nownaturally gained an advantage.

EDITED

Det var den jagande apans behöv av samordning som gynnade språkets framväxt. När våra föräldrar skulle <fniss> när våra förfäder eee skulle lägga ner stora villebråd krävdes planering, kanske över flera dagar. Den som då kunde beskriva och förstå ställen utöver "här och nu" fick ett naturligt försprång.

English translation:
It was the hunting apes' need for coordination which promoted the development of language. When our parents would <giggle> when our ancestors eee would bring down big game planning was needed, perhaps taking several days. Then, the person who could describe and understand places beyond "here and nownaturally gained an advantage.

The transcription of the materials also included the punctuation issue, since the original punctuation differed between the transcripts. This was no problem in the boundary annotation task, since in this task all punctuation was removed. However, in the prominence annotation task and for the tagging and parsing, punctuation was desired. In the prominence annotation task it was necessary to increase the readability, and for the tagging and parsing punctuation was required in the input format. How could we best insert a similar punctuation in all transcripts?

In the scripted speech, the punctuation was present before the speech (i.e. the reading aloud), thus the punctuation affects the speaker's reading. If we inserted punctuation in the non-scripted speech, the punctuation would be inserted on the basis of the prosody, i.e. it would mirror the listeners' interpretation of the speech signal. Thus, the punctuation in the two representations would have different bases; in the scripted speaking styles the prosody would depend on the punctuation, while in the non-scripted speech the opposite relationship would hold.

If the punctuation were retained, it would keep the original information in the scripted speech, but we would have to insert new orthographic information based on prosody in the non-scripted speech. Presumably this would also bias the subjects' annotations. However, it was necessary to facilitate the reading, and in addition we needed punctuation in the input to the tagger and the parser. Therefore we still decided to make a transcription based on the original punctuation in the scripted materials, including inserting hesitations etc., and to insert sparse punctuation in the non-scripted speech. We come back to the issue whether this procedure might have affected the subjects in their annotations.

To sum up: regarding the representation we have chosen readability over exact sound representation. Standard orthography was judged to be the most helpful alternative for the subjects and was therefore chosen. Hesitations and false starts were transcribed in a standard orthographic way, as it could be assumed to be represented in e.g. a book when the author wants to express "natural speech". Punctuation was present in the prominence marking task to increase the readability, and these transcripts were also used for tagging and parsing.

By now the reader has a picture of the architecture of the experiment as well as of the speech materials. We now proceed to describe in more detail the two tasks, boundary annotation and prominence annotation, as well as their conditions (Read and Listen).

3.3 The Annotation Tasks

We have described our experiment as approaching the segmenting issue from two aspects: the segment boundaries and the prominent parts of the segment. These two aspects are mirrored by two tasks: the boundary annotation and the prominence annotation, each carried out by separate groups of subjects. In the boundary annotation task we had inexperienced subjects annotate boundaries in the four different speaking styles. For each task we had one group annotating condition Read and another group annotating condition Listen.

From previous studies, e.g. in (Gustafson-Čapková and Megyesi, 2001), (Megyesi and Gustafson-Čapková, 2001), we had learnt that it was important for the subjects to have some knowledge of the materials they annotated, otherwise they tended to ponder what the materials were about to such an extent that it disturbed the annotation. Thus, in both tasks we provided the subjects with short descriptions of the content of each of the materials, e.g. explaining what a map-task dialogue is, and that the monologue consisted of read aloud news and retold news. An overview of all the materials was also given in the instructions. Regarding the transcripts, we pointed out that the appearance differed slightly between the scripted and non-scripted materials, so that this would not cause the subjects any confusion.

For both tasks, boundary annotation and prominence annotation, the subjects were instructed to make the annotations of the transcripts in the same order as the transcripts were found in the overview, i.e. starting with the scripted monologue, after this the non-scripted monologue, and then the dialogues in the same order. The idea was to encourage the subjects to start with the materials that were considered to be the easiest ones, and then proceed to the more difficult dialogues. There was no control as to whether the subjects followed this order.

The annotation work was carried out in the subjects' free time either in the their homes or in the computer-lab at the Department of linguistics. The subjects were given a deadline of two weeks to carry out the task, and if they did not finish the task in time, the deadline was extended. In both tasks the subjects were encouraged to contact the experiment leader if they had any questions. Some subjects did this, but most questions concerned prolongation of the deadline.

3.3.1 The Boundary Annotation

In the boundary annotation task we study the aspect of the contours of discourse segments. Do the subjects' annotations of boundaries differ across the speaking styles? Do subjects' annotations differ within speaking styles, depending on whether they do not have access to the spoken message (condition Read) or if they can hear it (condition Listen)?

In order to instruct the subjects, we had to decide what kind of boundaries they should mark. What units are relevant in a discourse, and how shall they be described to the subjects? Since the notion of "discourse segment" is unfamiliar to most people and is in addition very vague, it is difficult to utilize. If the notion of discourse segment should be used, detailed guidelines had to be developed in order to give the subjects equal understanding of the concept. Such a learning phase together with the extensive annotation of the materials would make participation in the experiment very time-consuming, and it would be difficult to persuade the required number of subjects to carry out the task without offering better compensation than two cinema tickets. In addition, if we were to develop guidelines, we would get an annotation due to a specific discourse framework. What we wanted was an annotation procedure which allowed the subjects to make an intuitive segmentation and not to worry too much about whether they were "right" or "wrong".

In order to avoid these problems and also to facilitate the task for the subjects, we decided to make use of punctuation; the subjects were instructed to insert punctuation in paper transcripts of the speech. The assumption behind the use of punctuation is that punctuation reflects the structure of the discourse. Our assumption was that subjects would insert some kind of punctuation mark, not necessarily the same one, at positions where they felt some kind of boundary. Punctuation is present on many levels of an orthographic message representation, not only between for instance clauses and sentences in form of commas and full stops, but also between higher level discourse units in form of paragraph markings. In addition the subjects were familiar with punctuation, and so they would feel more confident using this method than learning new segmenting guidelines, and this in turn would make them bother less about whether they were "right" or "wrong", and focus better on the segmenting task. With this annotation of the data we would be able to compare strategies for subjects' structuring of different speaking styles by using punctuation.

In the boundary annotation task the transcripts were prepared in such a way that all punctuation was removed, and edited transcripts with the bare string of words was given to the subjects. Figure 3.3 shows an excerpt from the transcript of a scripted monologue prepared for the boundary annotation task. However, in the case of dialogues some information of the structure was retained also in the prepared transcripts; speaker information was retained, meaning that speaker change was indicated with a line break. It was judged that otherwise the subjects would have too much difficulty in carrying out the task.

Since the task of segmenting is divided into two conditions, condition Read (annotation without access to the speech signal), and condition Listen (annotation with access to the speech signal). The group which performed condition Read only got the transcripts, while the group that carried out condition Listen got the transcripts and a CD with the sound files of the materials. There was one group of subjects for each condition and no subject participated in more than one condition. With this setting we can examine the impact of the prosody on the segmenting task.

News broadcast, used for boundary annotation

och dethär är dagens eko kvart i fem nya storbanken finland tog första steget giftskandalen på hallandsåsen båstad vill ha oberoende undersökningskommission och historiskt möte i belfast idag i ekostudion helena sjöholm och marianne hasslow ja idag blev det alltså klart med ytterligare en storaffär i bankvärlden det är nordbanken och finländska merita som går ihop och bildar nordens största bank samgåendet blir därmed en komplicerad teknisk affär men det var den enda form vi kunde välja det säger dom båda bankledningarna meritas vesa vaino vainio och nordbankens hans dahlborg

...

English translation

and this is today's echo at a quarter to five new big bank finland took the first step poison scandal at hallandsåsen båstad wants independent commission of investigation and historical meeting in belfast today in the echo-studio helena sjöholm and marianne hasslow yes today yet another big business transaction was finished it is nordbanken and finnish merita who join forces and establish the north's largest bank the fusion will be a complicated technical transaction but this was the only form we could choose say both bank managements merita's vesa vainio and nordbanken's hans dahlborg

...

Figure 3.3: Excerpt of a news broadcast transcript used in the boundary task.

BEFORE annotation of punctuation
<Talare A> maria olsson och det är den tjugoförsta mars nittonhundra nittio sju
<Talare B> ja ska jag säga det samma eller eller det vill säga motsvarande lars andersson och det är väl fortfarande den tjugoförsta mars då i så fall
<Talare A> då ska vi se då har vi en en s karta här framför oss och jag har landstigit på en plats på den här ön och det börjar vid en en in i en bukt en ganska ovalt formad bukt inne på västra sydvästra sidan utav den här ön har du den också
<Talare B> jo då eee och tydligen så är ju formen på våra öar identiska så att så att själva den bukten ska det inte vara några problem att hitta

AFTER annotation of puncuation
<Talare A> Maria Olsson. Och det är den tjugoförsta mars, nittonhundra nittio sju.
<Talare B> Ja, ska jag säga det samma? Eller, eller det vill säga motsvarande. Lars Andersson, och det är väl fortfarande den tjugoförsta mars då i så fall.//
<Talare A> Då ska vi se, då har vi en en s karta här framför oss, och jag har landstigit på en plats på den här ön. Och det börjar vid en en in i en bukt, en ganska ovalt formad bukt, inne på västra sydvästra sidan utav den här ön. Har du den också?
<Talare B> Jo då! Eee, och tydligen så är ju formen på våra öar identiska, så att så att själva den bukten ska det inte vara några problem att hitta.

English transliteration
<Speaker A> *Maria Olsson. And it is the twenty-first of March, nineteen hundred ninety seven.*
<Speaker B> *Yes, shall I say the same? Or, or that is to say the corresponding. Lars Andersson, and well, it is still the twenty-first March then in that case.//*
<Speaker A> *Then let us see, then we have a a s map here in front of us, and I have disembarked on a place on this island. And it begins at a a in a bay, a rather oval shaped bay, in on the west south-west side of this island. Do you also have it?*
<Speaker B> *Oh yes! Eee, and evidently the shape on our islands is identical so that so that this particular bay should not be a problem to find.*

Figure 3.4: Example of boundary annotation in the dialogue.

komma (comma)	,
punkt (full stop)	.
frågetecken (question mark)	?
utropstecken (exclamation mark)	!
semikolon (semicolon)	;
kolon (colon)	:
parentes (parenthesis)	( )
tankestreck (dash)	–
anföringstecken (quotation marks)	" "
tre punkter (three dots)	...
nytt stycke (new paragraph)	//

Figure 3.5: The table of punctuation marks given to the subjects.

The subjects were instructed to make the annotations in the paper transcripts, and they were advised to use a pencil so that they could erase and make changes if they wanted to. They were also given an example of an annotation of an excerpt from one of the dialogues (the beginning of the dialogue). This example is shown in figure 3.4 (this specific excerpt was then discarded in the analysis). In addition the subjects were informed that the dialogue contained parts with overlapping speech, and that these parts would be very hard to annotate, but they were encouraged to do their best. If they found any part impossible to annotate, or if they felt very uncertain about their own annotation, they were instructed to put this part inside square brackets, thus signalling that they were unsure or discarded it. However, only one subject made occasional use of this "uncertainty signal".

The subjects were provided with a table of punctuation marks which they could use for the task (see figure 3.5). This was done with an aim to preventing the subjects from worrying about whether they were allowed to use one or another sign.

Boundary Annotation, Condition Read

In condition Read the subjects have access only to the string of words, and they should recognise the content and interpret a certain structure from the words forming clauses and sentences. The subjects were instructed to insert punctuation in the transcripts, and the result is a segmenting based on the string of words.

Boundary Annotation, Condition Listen

In condition Listen, in addition to the transcripts the subjects also have access to the speech signal. The segmenting is based on both the string of words and the speech signal. This means that prosodic cues to discourse structure may assist the subject in inserting punctuation.

The subjects were instructed to listen to the sound-files, and to insert punctuation in the transcripts on the basis of what they heard.

The original instructions for the boundary marking task, both condition Read and condition Listen, are found in appendix 1.

Summary of the Boundary Annotation Task

Based on the annotations from the boundary marking task we can study the subjects' annotation of structure in different speaking styles. We can also compare the results from condition Listen with the results from condition Read, and thus we can see if the speech signal added something in terms of structuring cues, or whether the results are similar to the results in condition Read. We can also examine whether we find differences between the speaking styles, i.e. whether the subjects seem to be more sensitive to prosodic cues in any specific speaking style.

With the annotations from the subjects we have data on how the subjects assign structure in different speaking styles, based both on the string of words alone (condition Read), and on the string of words together with the speech signal (condition Listen).

3.3.2 Questions, a Specific Kind of Boundaries

The approach with punctuation in the boundary marking task makes it possible to study a specific type of segments: segments ending with question marks – questions. This means that we have a way to study a segment intention, the intention to request. In this study we can investigate to what extent the subjects use the linguistic form and to what extent they use the prosody to interpret a specific segment as a question.

This study of questions is a subpart of the study of boundaries.

3.3.3 The Prominence Annotation

The task Prominence mirrors the aspect of what is most salient in a stretch of discourse. Our aim with this part of the experiment is to study whether prosody affects subjects'

annotations of prominence and if so, does it affect the annotations more in one speaking style than in another? Would subjects agree more on what is prominent in e.g. dialogue than in monologue, and would they agree more on what is prominent if they had access to the speech signal than if they did not. In other words, is prosody in general more important to the subjects' decisions in one speaking style or in another? In addition, how do the results relate to the results from the boundary annotation task?

To study the aspect of prominence we had subjects annotate what they perceived as prominent in the different speaking styles. This task is not as familiar to the subjects as the insertion of punctuation in the boundary annotation task. In this task the subjects were instructed to mark the word(s) or phrase(s) they perceived as prominent in paper copies of the transcripts of the different speaking styles.

The materials from the speaking styles were the same as those in the boundary marking task, but the transcripts differed slightly: In this task, existing punctuation was retained in the scripted conditions. In figure 3.6 an excerpt from a transcript of the scripted monologue (same excerpt as in figure 3.3) used in the prominence task is shown. In the non-scripted conditions, speaker information and a minimal amount of punctuation were inserted (see figure 3.6 and figure 3.7 for examples).

The choice to retain punctuation in the prominence marking task was made because we considered the task to annotate prominence without any help with the readability too difficult and too time consuming for the subjects.

The annotation method was to underline the word(s) or phrase(s) the subjects found prominent. Also for this task there were two conditions: condition Read where subjects annotated prominence on the basis of the transcript alone, and condition Listen where subjects annotated prominence on the basis of both transcripts and speech signal. An example of a possible annotation of an excerpt from a dialogue (same excerpt as in the boundary marking task) was included in the instructions. This example is shown in figure 3.7.

This example was given in order to instruct the subjects how to annotate, but also to give them a feeling for the granularity of the annotation. In this task, in the same way as in the boundary marking task, the subjects were informed about overlapping speech and given the option to put passages inside parentheses to show that they were uncertain about the annotation.

Prominence Annotation, Condition Read

In this condition the subjects received the instructions and the transcripts of the four speaking styles. They were instructed to annotate what they found prominent when reading the transcripts.

News broadcast, used for prominence annotation

Och dethär är Dagens Eko kvart i fem.
Nya storbanken, Finland tog första steget. Giftskandalen på Hallandsåsen. Båstad vill ha oberoende undersökningskommission. Och historiskt möte i Belfast idag. I Ekostudion Helena Sjöholm och Marianne Hasslow.

Ja, idag blev det alltså klart med ytterligare en storaffär i bankvärlden. Det är Nordbanken och finländska Merita som går ihop och bildar Nordens största bank. Samgåendet blir därmed en komplicerad teknisk affär, men det var den enda form vi kunde välja, det säger dom båda bankledningarna: Meritas Vesa Vainio och Nordbankens Hans Dalborg.
...

English translation

And this is Today's Echo at a quarter to five.
New big bank, Finland took the first step. Poison scandal at Hallandsåsen. Båstad wants independent commission of investigation. And historical meeting in Belfast today. In the Echo-studio Helena Sjöholm and Marianne Hasslow.

Yes, today yet another big business transaction was finished. It is Nordbanken and finnish Merita who join forces and establish the North's largest bank. The fusion will be a complicated technical transaction, but this was the only form we could choose, this say both bank managements: Merita's Vesa Vainio and Nordbanken's Hans Dahlborg.
...

Figure 3.6: Excerpt of a news broadcast transcript used in the prominence task.

An excerpt annotated for prominence
<Talare A> maria olsson och det är den tjugoförsta mars nittonhundra nittio sju
<Talare B> ja ska jag säga det samma eller eller det vill säga motsvarand lars hansson och det är väl fortfarande den tjugoförsta mars då i så fall
<Talare A> då ska vi se då har vi en en s karta här framför oss och jag har landstigit på en plats på den här ön och det börjar vid en en in i en bukt en ganska ovalt formad bukt inne på västra sydvästra sidan utav den här ön har du den också
<Talare B> jo då och tydligen så är ju formen på våra öar identiska

English transliteration
<Speaker A> *maria olsson and it is the twenty-first of march nineteen hundred ninety seven*
<Speaker B> *yes shall I say the same or or that is to say the corresponding lars andersson and well it is still the twenty-first march then in that case*
<Speaker A> *then let us see then we have a a s map here in front of us and I have disembarked on a place on this island and it begins at a a in a bay a rather oval shaped bay in on the west south-west side of this island do you also have it*
<Speaker B> *oh yes eee and evidently the shape on our islands is identical*

Figure 3.7: Example of prominence annotation.

Prominence Annotation, Condition Listen

In this condition the subjects received instructions and transcripts together with a CD with the sound recordings. The subjects were instructed to listen to the sound files and then mark in the transcripts what they found prominent.

The original instructions for the prominence marking task, both condition Read and condition Listen, are found in appendix 1.

Summary of the Prominence Annotation Task

The data from the prominence annotation task show us how subjects assign prominence markings in different speaking styles, making use of either the transcripts alone (condition Read) or both the transcripts and the speech signal (condition Listen).

3.4 The Subjects

Since all the results presented later in this thesis are based on subjects' annotations, a brief description of the subjects who carried out these annotations is in place. We describe what type of subjects participated in the task, and we also report some of the subjects' comments on the task.

All subjects were inexperienced with regard to discourse annotation, but they were not inexperienced with regard to language analysis since they were at least in the second semester of studies in Linguistics or Swedish language at Stockholm University, and all except one were native speakers of Swedish. The non-native speaker was fluent in Swedish. For the annotation work the subjects were rewarded with two cinema tickets.

Since the annotation task was extensive for the subjects, the collection of the data was time consuming. Many subjects did not finish their task, many found the task too extensive to complete. We asked approximately 140 persons, around 70 volunteered for the experiment, and of those 34 finished the task. The collection of these 34 annotations took approximately 7 months. This means that the annotation was not without effort for the subjects, and in chapter 8, Discussion, we will return to this fact and discuss whether this might have affected the results.

Our aim was to have 10 subjects for each condition, and this aim was reached in the prominence annotation task. However, in the boundary marking task we had a total of 14 subjects; 6 in condition Read and 8 in condition Listen. The motivation for not having a total of 10 subjects in each condition in the boundary annotation task was that the results were fairly stable between subjects. In the prominence annotation task the case was the opposite; the annotator agreement fluctuated and was influenced as we added a new subject. Therefore we considered it important to have as many subjects as possible in the prominence annotation task. On the other hand, in the boundary marking task the agreement was fairly stable starting from three subjects upwards, and we could not see the same fluctuations. When we had reached an outer time limit for the data collection it was therefore decided that 6 plus 8 subjects in the boundary annotation task would be sufficient.

In general the subjects found the annotation harder and more time-consuming than they had first expected. In particular, this was the case for the prominence annotation task. We return to this issue in chapter 8, Discussion, where we discuss if, and in that case how, the annotation task itself might have affected the results.

Chapter 4

Preprocessing and Preanalysis

SINCE a core feature in our experiment is the comparison between different speaking styles, it is important to know more precisely how the speaking styles in our data relate to each other. It is clear that they differ along the qualitative dimensions, scriptedness and interaction, but there are as yet no figures indicating clear quantitative stylistic differences.

To find a stylistic reference point for our data a comparison was carried out between our materials and data from previous research on differences in spoken and written Swedish. With these measures we do not intend to classify our data as belonging to some specific style, but only to relate it to previous research which thus can function as our "stylistic baseline". The method used in the comparison is the degree of nominality as computed with NVQ or NQ.

The materials were part-of-speech tagged as well as parsed with shallow parsing. In addition databases consistent across all speaking styles were constructed. The preprocessing work, including the part-of-speech tagging and the tag set, the parsing and the phrase categories, and finally the development of the database, is described in section 4.1.

Not only the transcripts but also the raw annotation data from the subjects had to be preprocessed to be comparable between subjects and conditions. In section 4.3 we present data on the annotation frequency in the boundary annotation task in order to give the reader an introductory overview of the annotated materials.

In order to measure the degree of inter-annotator agreement, the κ (kappa) statistics are often used. In this thesis we too will use this measure, and in the last section, section 4.4, we describe it briefly. κ is used in several different disciplines, such as medicine, linguistics and computer science, and it is somewhat confusing that different disciplines use different threshold values for the interpretation of κ. Therefore a brief orientation of the range of threshold values in use is given, and in addition it is stated how the κ value is used in our study.

4.1 Processing of the Materials

As described in chapter 3 the transcripted materials as given to the subjects came in slightly different formats depending on annotation task (Boundaries – Prominence). The formats and the decisions made are described in detail in chapter 3. However, we briefly highlight some of the points which are important for the preprocessing.

In the boundary annotation task all punctuation was removed, but in the dialogue transcripts structural information in the form of speaker information was retained.

In the prominence annotation task the punctuation and other structural information was retained. However, because of the different origins of the transcripts, the structural information differed between them. The scripted monologues were presented with the original punctuation retained, the non-scripted monologue was divided into paragraph-like sections to facilitate reading, the scripted dialogue was presented with the punctuation as well as the speaker information retained, and the non-scripted dialogues were presented with the speaker information.

The transcripts used in the prominence marking task, i.e. the transcripts with structural information, were used, with slight modifications, as input to the tagger and the parser. This included conversion to lower case and a division into one sentence unit per line ending with a full stop, and that a new paragraph was indicated by a blank line. In the case of the dialogues, the speaker contribution was used as the unit corresponding to a paragraph, and speaker change was indicated by a blank line. When the contribution consisted of two or more sentence-like utterances, these were put on separate lines, and a full stop was inserted after each of those utterances. In many cases, the utterance and and the contribution were overlapping units, i.e. the sentence unit and the paragraph unit were overlapping. After this editing the transcripts could serve as input to the tagger.

4.1.1 The Part-of-Speech Tagging

All transcriptions were automatically PoS-tagged with the TnT tagger (Brants, 2000) trained on SUC (Ejerhed *et al.*, 1992) with the PAROLE tag format (Leech and Wilson, 1996) by Beáta Megyesi.

The output from the tagger gave the content of the transcripts in one column, with the PAROLE tag for the part-of-speech in the second column. However, in our analysis we do not use the fine-grained PAROLE tagset but use a simplified tag-set based on the original tagging instead. In figure 4.1, the output format with the simplified tags inserted in the database is shown. The first column (Transcription) shows the transcription, the

second (PoS-tag) the part-of-speech tags. For the sake of clarity, an extra column with an English translation is inserted to the right in all examples. [1]

Transcription	PoS-tag	English
och	CC	*and*
dethär	PF	*this*
är	V@	*is*
dagens	NC	*today's*
eko	NC	*echo*
kvart	NC	*quarter*
i	SP	*to*
fem	MC	*five*
.	FE	*.*
nya	AQ	*new*
storbanken	NC	*big-bank*-DEF
,	FI	*,*
finland	NP	*finland*
tog	V@	*took*
förSta	MO	*first*
steget	NC	*step*-DEF
.	FE	*.*

Figure 4.1: The format of a part-of-speech tagged transcript.

The simplified tags exemplified in 4.1 give information about the part-of-speech, and the type of the part-of-speech. For instance, in figure 4.1, line one, we find the word "och". It is tagged CC, which means that the general part-of-speech is " conjunction", and furthermore this word is a conjunctive conjunction, and not a subjunction.

The simplified tag-set gives us a tag-set with 30 tags to be compared with the original PAROLE tag-set which contain 156 tags. However, for our materials, which have a total of 9522 words, the simplified tag-set still causes too many problems with sparse data. Therefore an even more compact tag-set was constructed, consisting only of the most general part-of-speech information in the PAROLE tags. This means that different types of a part-of-speech are all collapsed into one tag, e.g. conjunctions and subjunctions are merged into one single category of conjunctions. This tag-set, (tag-set "Compact") consist

[1] This translation is not an exact word for word translation, nor a fully fluent translation. Its purpose is to enable a person with no knowledge of Swedish to understand what the example is about. Where necessary, morpheme information (in capital letters) and/or a word-for-word translation are added.

of 13 tags. In table 4.1 the simplified tag-set plus the tag-set "Compact" are presented together with a description of each tag. Thus the simplified tag-set "Compact" offers a suitable tag-set for our relatively small corpus.

The tag-set was developed for Swedish written text, but in our materials we have also used transcripted speech. In order to cover some speech specific spoken language phenomena we manually added two tags to the materials.

- "TV" ("tvekan"), for hesitations and fragments due to restart or interruption.
- "ME" ("meta"), for sounds which could not be classified as hesitation (e.g. noise, giggling or highly reduced words impossible to interpret reliably).

Feedback expressions in the dialogues were tagged with "I" (interjection).

4.1.2 The Parsing of the Materials

In order to make an analysis on the phrase level, the transcripts were parsed with a shallow parser, described by Megyesi (2002), and also this work was carried out by Beáta Megyesi. The parsing was done in conjunction with the tagging, and therefore no extra preparation of the transcripts was done.

The phrase structure was represented by nine grammatical phrasal categories. Below we render the description of each phrasal category together with an example, as given by Megyesi (2002).

- ADVP – Adverb Phrase consists of adverbs which can modify adjectives or numeral expressions. *Example: very*
- AP – Minimal Adjective Phrase constitutes the adjectival head and its possible modifiers, e.g. ADVP and/or PP. *Example: very interesting*
- APMAX – Maximal Adjective Phrase includes more than one AP with a delimiter or a conjunction in between. *Example: very interesting and nice*
- NUMP – Numeral Expression consists of numerals with their possible modifiers, for example AP or ADVP.*Example: several thousands*
- NP – Noun Phrase may include the head noun and its modifiers to the left, e.g. determiners, nouns in genitive, possessive pronouns, numeral expressions, AP, APMAX and/or compound nouns. Thus possessive expressions do not split an NP into two phrases. *Example: Pilger's very interesting and nice book*

Compact tag-set (Part-of-speech)	**Simple tag-set (Part-of-speech and Type)**	**Description**
A	AF	Perfect participle
	AP	Present participle
	AQ	Adjective
C	CC	Conjunction
	CI	Infinitive mark
	CS	Subjunction
D	D0	Determiner (definite/indefinite)
	DF	Determiner (definite)
	DH	Determiner (interrogative/relative)
	DI	Determiner (indefinite)
F	FE	Major delimiter
	FI	Minor delimiter
	FP	Pairwise delimiters
I	I	Interjection
M	MC	Cardinal number
	MO	Ordinal number
N	NC	Common noun
	NP	Name
P	PF	Personal pronoun (definite)
	PE	Possessive pronoun (interrogative/relative)
	PH	Pronoun (interrogative/relative)
	PI	Personal pronoun (indefinite)
	PS	Possessive pronoun
Q	QC	Particle (compound)
	QS	Particle
R	RG	Adverb
	RH	Interrogative/relative adverb
S	SP	Preposition
V	V@	Verb
X	XF	Foreign word

Table 4.1: The simplified part-of-speech tag-sets.

- NPMAX – Maximal Projection of an NP includes one or more NP(s) with following PP(s) as possible modifier. *Example: Pilger's very interesting and nice book about politics*
- PP – Prepositional Phrase consists of one or several prepositions delimited by a conjunction and one or several prepositions delimited by a conjunction and one or several NPs/NPMAXs, or in elliptical expressions an AP only. *Example: about politics*
- VC – Verb Cluster consists of a continuous verb group belonging to the same verb phrase without any intervening constituents like NP or ADVP. *Example: would have been*
- INFP – Infinitive Phrase includes an infinite verb together with the infinitive particle and may contain ADVP and/or verbal particles. *Example: to go out*

In addition to the phrasal categories listed above, tags indicating a word's position in a phrase and a tag marking relative clauses were used:

- XB – the initial word inside the phrase X.
- XI – non-initial word inside the phrase X.
- O – word outside of any phrase (This includes e.g. conjunctions, interjections/feedback).
- RELCL – relative clause.

An excerpt of the parse of one scripted monologue is shown in 4.2.

4.1.3 The Acoustic Markup

In addition to the part-of-speech annotation and the shallow parsing the data was annotated with selected prosodic features related to boundaries and prominence. All measurements were done in Praat (Boersma and Weenink, 1996) by the author.

Selected Prosodic Correlates to Boundaries

We have chosen to study pauses as a prosodic feature related to boundary annotation, and in our data a pause means a silent interval longer than 50 ms. The threshold for pause duration was motivated by Fant and Kruckenberg (1989) who report short word prompts as having a duration of 50 – 100 ms. All pauses were measured and comments about breathing were recorded. No filled pauses were regarded as pauses in our data,

Transcription	Parse	*English*
och	O	*and*
dethär	NPB	*this*
är	VCB	*is*
dagens	NPB_NPMAXB	*today's*
eko	NPI_NPMAXI	*echo*
kvart	NPB_NPMAXI	*quarter*
i	PPB_NPI_NPMAXI	*to*
fem	NUMPB_NPB_PPI_NPI_NPMAXI	*five*
.	O	.
nya	APMINB_NPB	*new*
storbanken	NPI	*big-bank*-DEF
,	O	,
finland	NPB	*finland*
tog	VCB	*took*
första	NPB	*first*
steget	NPI	*step*-DEF
.	O	.

Figure 4.2: Excerpt from the parse of a scripted monologue.

since we wanted to keep track of whether a subject annotated a boundary before or after any noise. Thus, filled pauses were transcripted as a noise, e.g. "eee", and if this noise was situated at a silent interval longer than 50 ms., this interval was measured and classified as a pause.

Selected Prosodic Correlates to Prominence

The focal accent was selected as a prosodic feature related to the prominence annotation. All words annotated as prominent by the majority of subjects were annotated as focal or non-focal. This classification was based on the prosodic model for Swedish suggested by Bruce (1998). It was carried out by inspecting the F0, and typical focal accents were classified as focal while words with a F0 which did not show signs of focal accent peaks were classified as non-focal even though the word could be classified as perceptually prominent. Our motivation for this approach was that we already had a perceptual assessment from 10 subjects, and therefore the focality assessment should be based primarily on F0 and not be rooted in yet another perceptual assessment.

In addition to the classification of focality a number of acoustic features for each of the words annotated as prominent by the majority of subjects were measured:

- Average pitch
- Average standard deviation of pitch
- Duration in milliseconds
- Average intensity

The four acoustic measurements are a rather modest collection. However, they are features which many researchers mention as important for acoustic prominence.

The acoustic markup is not extensive, but we believe that it might be extensive enough to capture clear differences.

4.1.4 The Construction of the Database

The tagged and parsed data as well as the acoustic measurements were inserted in a database [2]. In this step the word units and the structural information such as original punctuation were separated into different columns, and the parts-of-speech of punctuation marks were discarded in the analyses. The words and the structural information were separated because the database should serve as matrix for the transcripts in both condition Listen (without punctuation) and condition Read (with punctuation). Thus, we needed to keep "noisy" information apart from "silent" information. In figure 4.3 the basic structure of the database for a scripted monologue is shown.

In the very left column, (Nr.), in figure 4.3 the line numbering is shown, in the second, (Transcript), the transcripted speech with one word per line. When computing part-of-speech frequencies, these are based on the parts-of-speech assigned to the word units in this column, i.e. the parts-of-speech for punctuation are discarded. This separation of the word tokens and structural information was done in order to render equal the databases for the boundary marking task (no punctuation) and the prominence annotation task (punctuation) as well as the scripted and non-scripted materials.

In the third column, (Punc. after), the original punctuation except paragraph indication is inserted, the punctuation marks thus having their original position after the word on the same line in the column "Transcript". For example, in the original transcript the full stop on line 10 in the column "Punc. after" occurs after the word "fem" on line 10 in column "Transcript".

In the fourth column, (Para. after), the paragraph marking is inserted. The strategy is the same as in the case of punctuation: the paragraph markings' original textual position

[2]To save space, the parse and the prominence annotations are not included in the following examples

1	2	3	4	5	6
Nr	**Transcript**	**Punc. after**	**Para. after**	**PoS-tag**	**English**
2					
3	**och**			CC	*and*
4	**dethär**			PF	*this*
5	**är**			V@	*is*
6	**dagens**			NC	*today's*
7	**eko**			NC	*echo*
8	**kvart**			NC	*quarter*
9	**i**			SP	*to*
10	**fem**	.	&&	MC	*five*
11	**nya**			AQ	*new*
12	**storbanken**	,		NC	*big-bank*-DEF
13	**finland**			NP	*finland*
14	**tog**			V@	*took*
15	**första**			MO	*first*
16	**steget**	.		NC	*step*-DEF
17	**giftskandalen**			NC	*poison-scandal*-DEF
18	**på**			SP	*on*
19	**hallandsåsen**	.		NP	*halland-ridge*-DEF

Figure 4.3: The basic information in the database for a scripted monologue.

is after the punctuation mark on the same line. Thus, the paragraph marking in line 10 follows the full stop on line 10 in column "Punc. after", which in turn follows after the word "fem" on line 10 in the column "Transcript".

The fifth column (PoS-tag) shows the part-of-speech for the word on the same line. Thus, on line 10 in the column "PoS-tag" we find the tag MC (cardinal number), which is the part-of-speech tag assigned to the word "fem" on line 10 in the column "Transcript".

In this example an extra column (6) is inserted with an English translation of the transcript from column 2.

To the information shown in figure 4.3, information about pause and prominence measurements as well as more context were added. An example of an extended database including pause info, context, transcript, original punctuation and part-of-speech tagging is shown in figure 4.4.

Column 1 shows the line numbers, column 2, (Pause, sec), the duration of the acoustic pauses in seconds. The actual pause is found in the context after the word on the same line in column 6, "Transcript". Thus, the pause, with a duration of 0.732365 sec., on line 10 is found in the context after the word "fem", also on line 10 in column 6, "Transcript".

1	2	3	4	5	6	7	8	9	10	11	12
Nr	Pause, sec	Comment	2 before	Before	Transcript	Punc. after	Para. after	After	2 after	PoS-tag	*English*
2											
3					**och**			dethär	är	CC	*and*
4				och	**dethär**			är	dagens	PF	*this*
5			och	dethär	**är**			dagens	eko	V@	*is*
6			dethär	är	**dagens**			eko	kvart	NC	*today's*
7	0,270157		är	dagens	**eko**			kvart	i	NC	*echo*
8	0,07345		dagens	eko	**kvart**			i	fem	NC	*quarter*
9			eko	kvart	**i**			fem	nya	SP	*to*
10	0,732365	breath	kvart	i	**fem**	.	&&	nya	storbanken	MC	*five*
11			i	fem	**nya**			storbanken	finland	AQ	*new*
12	0,45589		fem	nya	**storbanken**	,		finland	tog	NC	*big-bank*-DEF
13			nya	storbanken	**finland**			tog	första	NP	*finland*
14			storbanken	finland	**tog**			första	steget	V@	*took*
15			finland	tog	**första**			steget	giftskandalen	MO	*first*
16	0,50662		tog	första	**steget**	.		giftskandalen	på	NC	*step*-DEF
...											

Figure 4.4: Scripted dialogue with information about pauses added.

Column 3, (Comment), records special information about the speech signal. For example, on line 10 the notion "breath" is inserted which means that the speaker breathed in relation to the pause. The part-of-speech tags are found in column 11 (PoS-tag), and in this example an English translation is offered in column 12.

Columns 4, 5, 9 and 10 give a linear word context to the main word, in column 6, "Transcript". Columns 4 and 5 give the preceding context, while 9 and 10 give the subsequent context. The context columns allow us to extract separate lines from the database, and still have a context around the target word. The context shown in 4.4 is just an example, and we might also add new context columns, e.g. pausing before a given word and not just after, as was the case in our examples.

So far, the examples have been taken from scripted monologue. The mark-up of the dialogues differs in some respects from the mark-up of the monologues. The most prominent difference is, of course, the fact that we have two speakers in the dialogues. Thus, in the databases over the dialogues, information about speaker and speaker contribution is given. Figure 4.5 shows an example of the information in a non-scripted monologue.

Examining figure 4.5 we find that the line numbering is located in the very left column (Nr.), and in column two (Pause, sec) pause duration in seconds is given. Column three (Comment) shows the comments and column four (Transcript) the word. Column five (Speaker) gives information about the speaker. One dollar sign indicates that the uttered word on the same line is uttered by speaker one, two dollar signs indicates speaker two. Thus, the word "buktkanten" on line 88 in the column "Transcript" is spoken by speaker one and the word "okej" on line 89 is uttered by speaker two.

1	2	3	4	5	6	7	8
Nr	Pause, sec	Comment	Transcript	Speaker	Turn	PoS-tag	*English*
...	...	...	...	...	...	...	...
83			**och**	$		CC	*and*
84			**följer**	$		V@	*follow*
85			**den**	$		DF	*this*
86			**här**	$		RG	*here*
87			**vackra**	$		AQ	*beautiful*
88	1,359549	breath	**buktkanten**	$	&&	NC	*bay-edge*-DEF
89		overlap	**okej**	$$	&	I	*okay*
90			**i**	$		SP	*in*
91			**en**	$		DI	*a*
92			**mjuk**	$		AQ	*smooth*
93			**kurva**	$		NC	*curve*
94	0,243715		**rakt**	$		RG	*straight*
95			**eee**	$		TV	*eee*
96		overlap	**norrut**	$	&&	RG	*northward*
97			**ska**	$$		V@	*shall*
98			**vi**	$$		PF	*we*
99			**gå**	$$		V@	*go*
100			**bakom**	$$		SP	*behind*
101			**ryggen**	$$		NC	*back*-DEF
102			**på**	$$		SP	*on*
103			**sälen**	$$		NC	*seal*-DEF
104	0,152888		**då**	$$	&	RG	*then*
...	...	...	...	...	...	...	...

Figure 4.5: Example of the format of a tagged dialogue (non-scripted dialogue).

Column six, (Turn), indicates speaker contribution. Here we have a record as to whether the same speaker continues to speak or if there is a speaker change. A blank cell indicates "same speaker", one ampersand indicates change to speaker one and two ampersands indicate change to speaker two. Thus, line 88 shows that the current speaker is speaker one, in column "Turn" there are two ampersands which mean that next word will be uttered by speaker two. In line 89 the current speaker is now speaker two, but in the column "Turn" we find one ampersand. This means that after the word "okay", uttered by speaker two, speaker one will start to speak. This column allows us to keep track of whether an extracted word is turn final or turn medial. We could also easily add preceding turn context, and thus information about where we find turn initial words.

Lastly, column seven (PoS-tag) shows the part-of-speech tagging, and in this example column eight offers an English translation.

The mark-up in figure 4.4 is the basic level of mark-up, to which the subjects' annotations were added later.

4.2 Characteristics of the Speaking Styles

What makes discourses differ in style? Of course there are many features in language that influence what impression the listeners, or readers, will get from a specific discourse. For this comparison we will focus on one dimension: the lexicogrammatical dimension *nominal– verbal* which was described more closely in chapter 2.

We have compared our data regarding nominal style in the different speaking styles to data from Einarsson (1978) and Melin and Lange (1986 2000). The examples from previous research serve as a reference point for our data, but no claims are made about the "true style" of our speaking styles. The materials used by Einarsson (1978), come from the corpora Skrivsyntax (Teleman, 1974) and Talsyntax (Loman and Jörgensen, 1971). The same holds for Melin's and Lange's materials.

A note on the style categories in previous research is appropriate. Einarsson's *Informative prose, professional writers* is a range of text books, papers and informative booklets, all written by professional writers. *Student essays* are all essays written by students in secondary school (the Swedish gymnasium) as part of their education, and the subgroups *High grade*, *Low grade* and *All students*, are reported separately. *Debate* are more formal conversations between academics. All the debates had a specific theme and a moderator and were recorded for research purposes. The *Conversation* categories consist of two groups: academics or manual workers. The method of collection differs between the academics and the workers. In the case of academics, the informal conversations before and after the debates (described above) were recorded without the speakers' knowledge. Afterwards, the materials were used as *Informal conversation*. In the case of the workers, an interviewer asked questions and gave feedback, in order to elicit as much flowing speech as possible. The category *Speech* used by Melin and Lange contains all categories of spoken language mentioned above, but merged into one category.

We compared our materials with the data from Einarsson (1978) regarding NVQ, and with the materials from Melin and Lange (1986 2000) regarding NQ. Even though these two comparisons actually compare the same materials twice, i.e. our data with data from Talsyntax and Skrivsyntax, we decided to carry out both comparisons since the measures NVQ and NQ are computed in slightly different ways. In addition, Einarsson (1978) and Melin and Lange (1986 2000) have divided the Talsyntax and Skrivsyntax into different categories. Einarsson (1978) have made more fine-grained distinctions for the spoken language, while Melin and Lange (1986 2000) have more fine-grained distinctions for the written language. For practical reasons we account for NVQ and NQ in separate figures starting with NVQ.

Figure 4.6 shows the NVQ, i.e. the noun to verb measurement, for the transcripts from our materials and the results reported for different styles by Einarsson (1978).

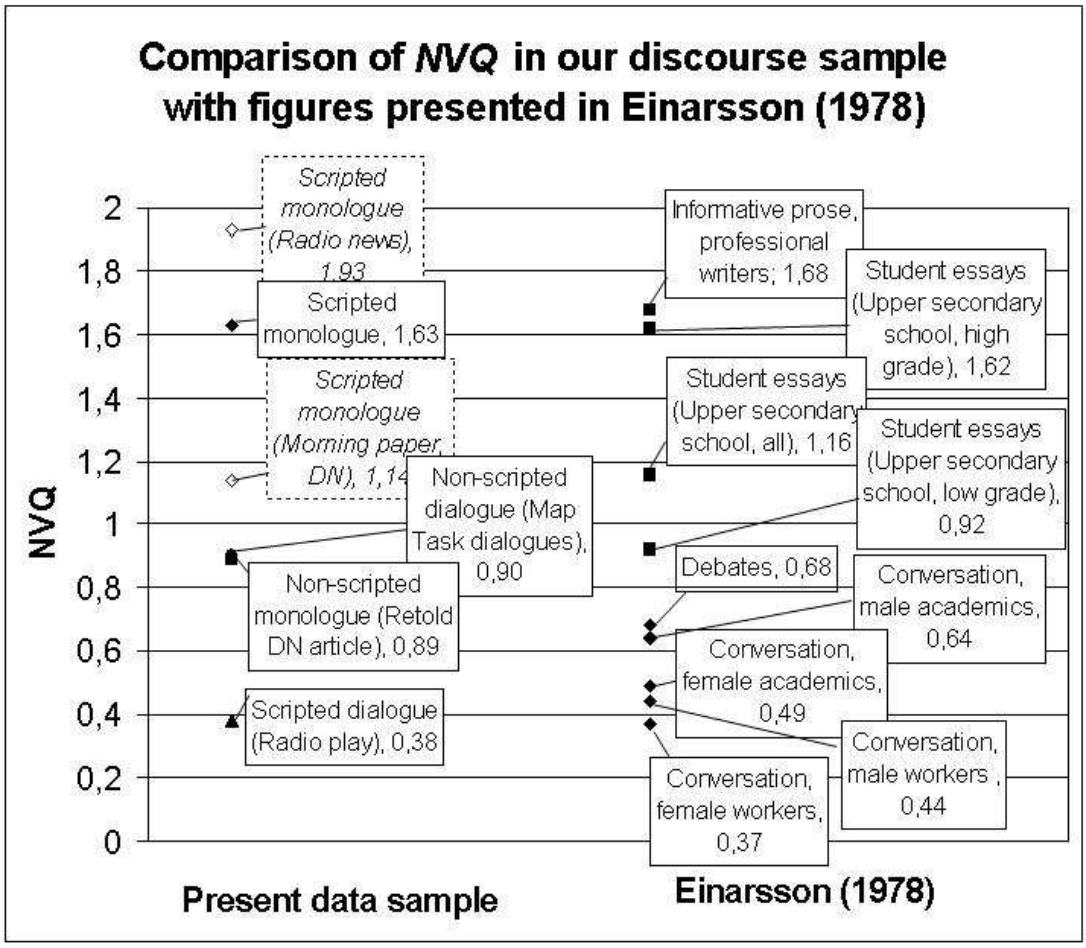

Figure 4.6: Comparison of NVQ measures for our data and data from Einarsson (1978).

On the left side of figure 4.6 the figures for our speaking style are presented, and on the right side the figures for Einarsson's (1978).

For our scripted monologue we show the figures for both the scripted monologue as a whole (Scripted monologue, 1.63) and the two subtypes of scripted monologue (Radio news and the morning paper DN, both italicised and in dotted boxes). This was done because of the great difference in NVQ between the two subtypes. Such a difference was not found between subparts of the other speaking styles. The news broadcast comes very high on the nominality scale, whereas the DN article is level with student essays.

Figure 4.6 shows that out of our four speaking styles, only the scripted dialogue (Radio play, 0.38) has a NVQ that correlates to spoken language in Einarsson's data (Female workers' conversation, 0.37). All three remaining speaking styles from our data have

figures close to different categories of written language. Scripted monologue as a whole (1.63) lies close to informative prose (1.68) and high-graded student essays (1.62), and non-scripted dialogue and monologue (0.90 and 0.89) lie close to low-graded student essays (0.92).

The figures in figure 4.6 show us that relative to previous research, the NVQ for our transcripts has a rather wide range, although the non-scripted monologue and the non-scripted dialogue are located close to each other (0.89 and 0.90), and are not level with spoken language in the data from Einarsson (1978).

The NVQ measure is a very rough way to measure style, therefore in addition we use the more elaborate NQ measure. In this version the prepositions, participles, pronouns and conjunctions, are included as described in 2.2. Figure 4.7 renders the NQ figures computed for our speaking styles and compares the NQ figures given by Melin and Lange (1986 2000:168). In these figures the category of *Informative prose* is divided into its subtypes, while the spoken language which (Einarsson, 1978) treated as different speaking styles are all merged into one category *Spoken language*.

Figure 4.7 shows the data from our speaking styles on the left side in the figure, while figures from (Melin and Lange, 1986 2000) are shown on the right side. Also here the two subtypes of the scripted monologue (Radio news and the morning paper) are rendered both as a whole and separately (in the dotted boxes).

In general the NQ measure ranks our transcripts in the same way as the NVQ measure, but the two non-scripted speaking styles move apart with the more complex NQ measure. This means that there are differences between the non-scripted monologue and the non-scripted dialogue which are not captured with the NVQ.

Figure 4.7 shows that the NQ measures are very high for the news announcing (1.50), higher than anything in the data from Melin and Lange (1986 2000). The scripted monologue as a whole in our data (1.16) is level with information booklets (1.19) and textbooks (1.18) as reported by Melin and Lange (1986 2000). The DN article is found to be slightly lower (0.72), at the same level as the student essays (0.72). The non-scripted monologue (0.56) and the non-scripted dialogue (0.45) are located in the gap between student essays and spoken language, while our scripted dialogue (0.22) is level with what Melin and Lange (1986 2000) mentions *spoken language* (0.25).

Since the degree of nominality is said to correlate with the degree of formality (Hellspong and Ledin, 1997), the most formal material in our data is the news broadcast. The DN article and the non-scripted materials are less formal, and least formal is the scripted dialogue. The distinction formal – informal is, however, not overlapping with the distinction scripted – non-scripted. Furthermore the figures show that our data is spread according to the degree of nominality. There is a large span between the lowest NQ value in our data (scripted dialogue, 0.23) and the highest one (scripted monologue, 1.16).

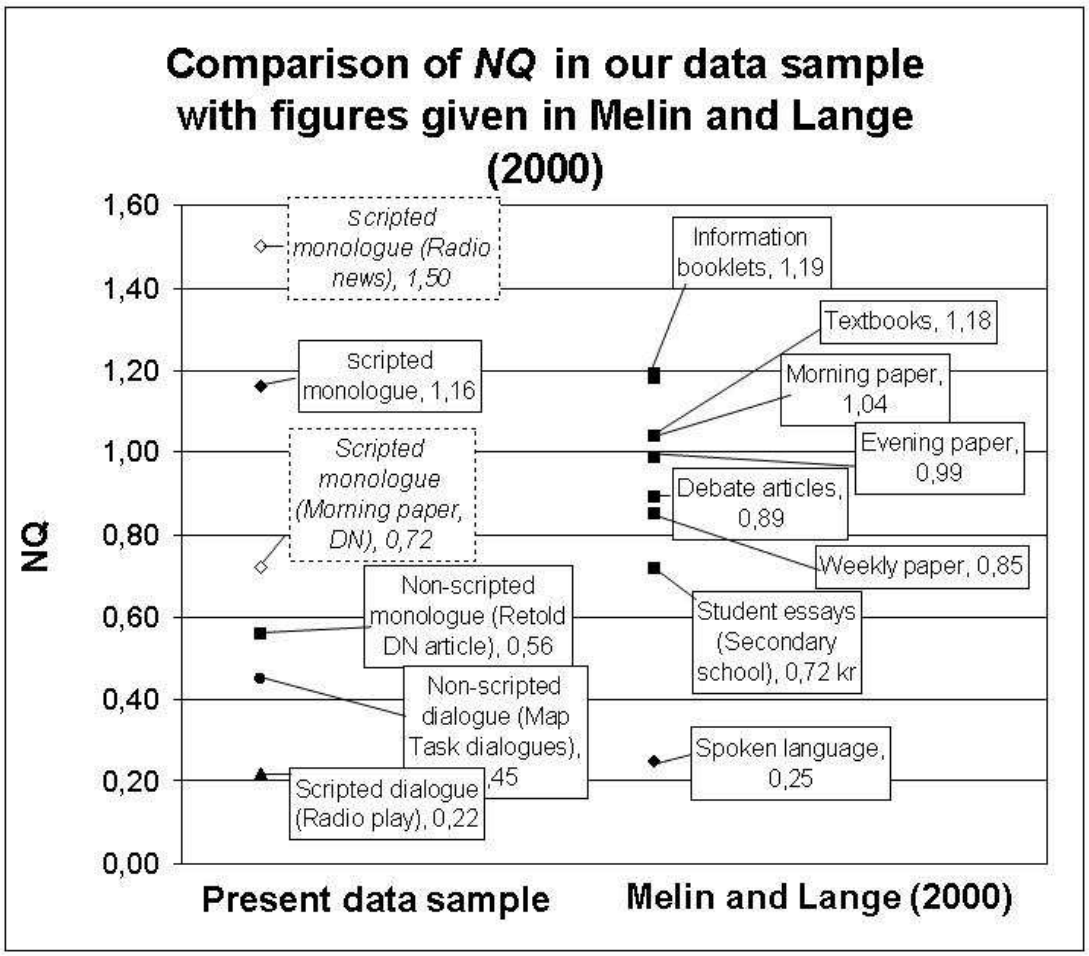

Figure 4.7: Comparison of NQ measures for our data and data from Melin and Lange (2000).

Both NVQ and NQ locate our speaking styles at different points on the scale of nominality. In the case of NVQ, the two non-scripted speaking styles are located very close to each other. However, when more features are taken into account as in the NQ measure, the two non-scripted speaking styles move apart. Thus, the NVQ and NQ validate our speaking styles as different from each other.

4.3 The Computation of Annotation Profiles

Our results are based on subjects' annotations, and consequently we need to examine how the annotators carried out the task. To show this more clearly we have made *annotation profiles* for every subject as well as for every transcript. These profiles are based

on the numbers of annotations made by each subject in each transcript. In this section we show how the annotation profiles were computed for the boundary annotations. The annotation profiles for the prominence marking task are computed in the same way but shown in chapter 6

A profile for a *subject* consists of the subject's global mean marking frequency computed on the mean marking frequency for each transcript annotated by the subject. A profile like this shows us whether a particular subject differs greatly from other subjects' annotation frequency.

A profile for a *transcript* consists of the mean of all subjects' mean marking frequencies computed for one transcript. This profile shows us whether the subjects' annotation frequency in general deviates in a specific transcript.

In the computation of the annotation frequency in the boundary annotation task *punctuation sites* were used. This is a way to handle the boundaries where some subjects put more than one punctuation mark, i.e. one subject might for instance have marked the sequence !". " (exclamation, end of quotation, full stop, beginning of quotation) while another subject in the same position just put a full stop. If nothing else is specifically stated, we have in our analysis chosen to collapse longer sequences such as !". " into one single marking of "punctuation site".

Before examining the annotation profiles we briefly remind the reader of some important features of the transcripts and the experiment conditions. In table 4.2 the number of words for each transcript is given. The number of words differs between the various subparts of the transcripts. Note, for instance, that in the non-scripted monologue, speaker A produced the greatest quantity of speech.

Speaking style	**Transcript**	**# Words**
Scripted monologue	Radio News 1	900
	Radio News 2	982
	Morning paper (DN)	990
Non-scripted monologue	Speaker A (female)	1339
	Speaker B (female)	464
	Speaker C (male)	282
Scripted dialogue	Radio play	2245
Non-scripted dialogue (Map task)	Dialogue 1	515
	Dialogue 2	601
	Dialogue 3	529
	Dialogue 4	522
Total	–	9522

Table 4.2: Number of words in each transcript.

The first column in table 4.2 (Speaking style) shows the speaking style, the second (Transcript) the subparts of the speaking styles and the third (# words) the number of words in each transcript.

Task	**Condition**	**# Subject**
Boundary	Read	6
	Listen	8
Prominence	Read	10
	Listen	10
Total	–	34

Table 4.3: Number of subjects in each experiment condition.

We would also like to remind the reader of the number of subjects in each experiment condition using an overview in table 4.3. The first column (Task) shows the task, the second column (Condition) the experiment condition and the third column (# subject) the number of subjects in each specific experiment condition.

We now proceed to examine the annotation profiles for the boundary annotation task. The same profiles for the prominence annotation are rendered in chapter 6.

4.3.1 Profiles for the Boundary Annotation Task, Condition Read

We start with an examination of the annotation profiles for the boundary annotation task in condition Read, the condition where the subjects inserted punctuation in the transcripts without having access to the speech signal.

Figure 4.8 shows subjects 1-6 horisontally, and the number of words vertically. The mean annotation frequency is indicated with a square for each subject. The mean annotation frequency shows how often on average a subject marked a punctuation site. The value was computed as follows: For a given subject, first the mean annotation frequency for each of the eleven transcripts was computed and then the subject's global mean, based on the eleven transcripts. The value we obtain shows how often on average a subject inserted punctuation, i.e. on average how many words there are per punctuation site. This means that a lower value indicates a higher annotation frequency (fewer words per punctuation mark, "shorter clauses") and a higher value indicate a lower annotation frequency (more words per punctuation mark, "longer clauses"). Thus, subject 1 has 6, which means that subject 1 inserted a punctuation mark on average after each sixth word. In figure 4.8 the range of the subjects' global means is also shown. The range for subject 1 lies between 4 and 8, which is a low value compared with some of the other subjects in the present condition. This shows that the mean in the different transcripts did not differ dramatically for subject 1.

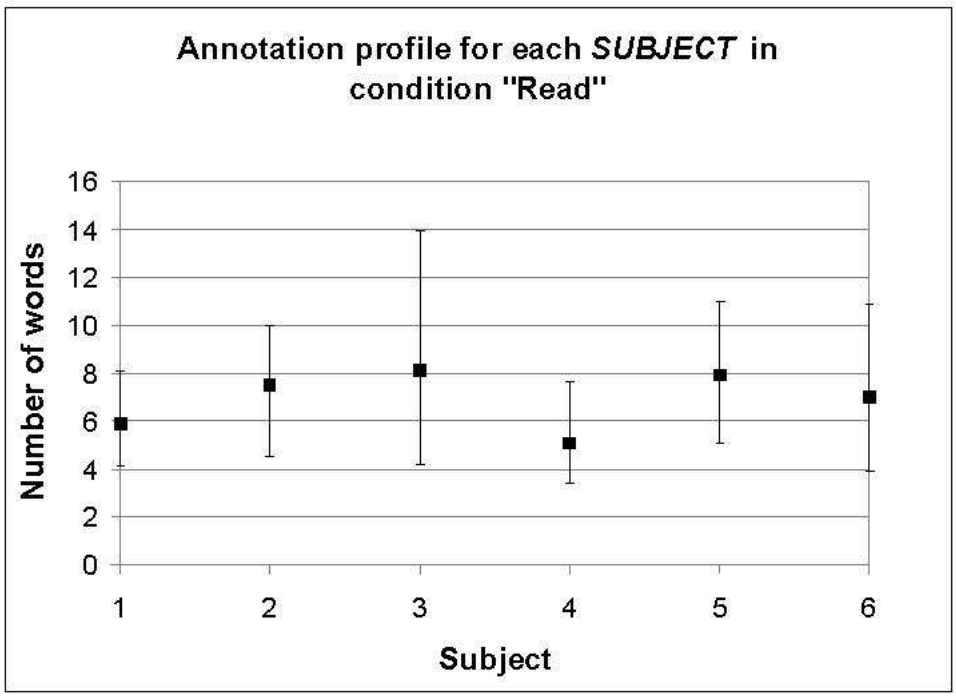

Figure 4.8: Annotation profiles for the subjects, condition Read.

The annotation frequency values for the subjects vary between 5 and 8. This means that the subjects inserted a punctuation mark on average every five to eight words, so these figures indicate that the average clause length is very roughly 5 to 8 words.

Examining the range in general we see that subject 3 has a great variation in the annotation frequency between the transcripts, between 4 and 14. This means that the mean annotation frequency differs a lot between transcripts for this particular subject. Also subjects 4 and 6 show a rather great variation, and examining the lowest and the highest values for all subjects, we find the lowest to be 3.5 (subject 4) and the highest 14 (subject 3).

The ranges indicate that the subjects' annotation frequency differs between transcriptions. Are the differences specific to "difficult" transcripts, or are they spread across all the transcripts? An answer to this question might be found in the annotation profiles for the transcripts in figure 4.9.

Figure 4.9 shows the transcripts (News 1, News2 etc) horizontally and the number of words vertically. Here, the mean annotation frequency indicates how often on average the subjects inserted a punctuation mark in a specific transcript. The values were computed as follows: for a given transcript, each subject's mean annotation frequency was computed, and then the global mean for the transcript based on all subjects' means

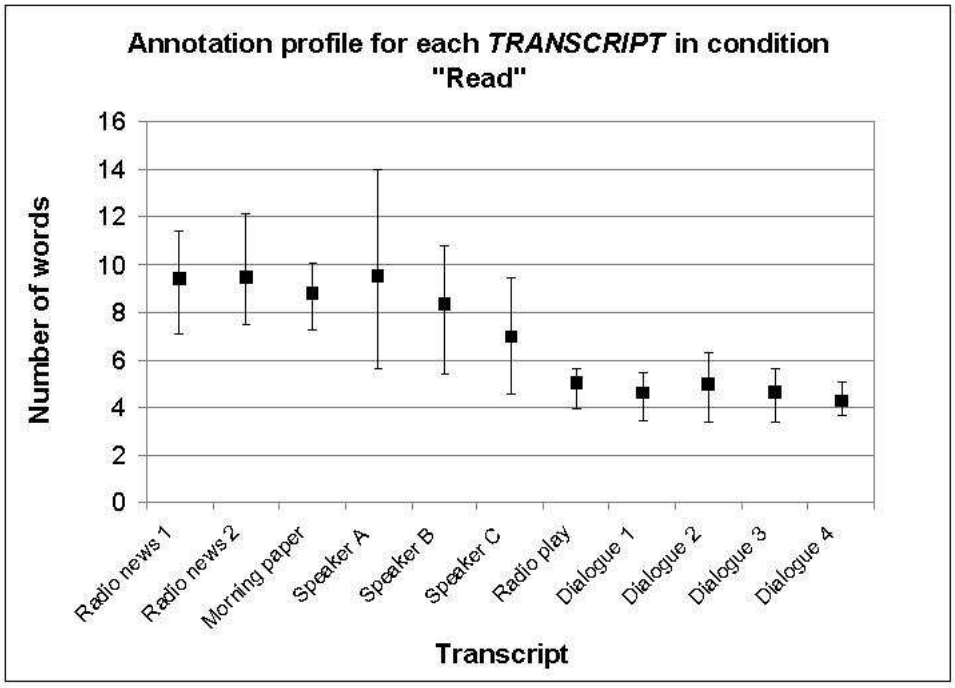

Figure 4.9: Annotation profiles for the transcripts, condition Read.

for this particular transcript. In this way it shows clearly whether a certain transcript deviates from the other transcripts regarding a subject's annotation frequency. A great range indicates a great variation in annotation frequency between subjetcs.

In figure 4.9 we find high values in the first transcripts (News1 to Speaker C, i.e. the monologues), the values then decline, and for the transcripts Radio Play – Dialogue 4 (the dialogues) the values are distinctly lower. This means that the monologues have a higher number of words per punctuation site while the dialogues have a lower number. Thus, on average the subjects insert punctuation more frequently in the dialogues than in the monologues.

In figure 4.8 the figures indicated that on average subjects inserted a boundary every four to six words. In figure 4.9 the picture is becoming more nuanced. The average number of words per inserted punctuation is higher in the monologues, about 7–10, and lower in the dialogues, about 4–5. This indicates longer clauses in the monologue transcripts and shorter clauses in dialogue transcripts.

There is a greater range of variation in the monologue transcripts than in the dialogue transcripts. The variation indicates that the subjects had different annotation strategies, especially in transcripts of Speaker A, Speaker B and Speaker C, the non-scripted monologues. Based on these figures, transcript of Speaker A, the first of the non-scripted

monologues, seems to give rise to the most varied annotation strategies among the annotators.

So far, the annotation profiles have concerned condition Read, where subjects made the annotations based on transcripts alone. We now proceed to condition Listen where subjects made the annotations based on both transcripts and sound files.

4.3.2 Profiles for the Boundary Annotation Task, Condition Listen

In condition Listen the subjects inserted punctuation in the transcripts, but in addition to the transcripts they also had access to the speech signal. The annotation profiles for the subjects and the transcripts in condition Listen are computed in the same way as the annotation profiles for subjects and transcripts in condition Read, described in the previous section. We start with the annotation profiles for the subjects.

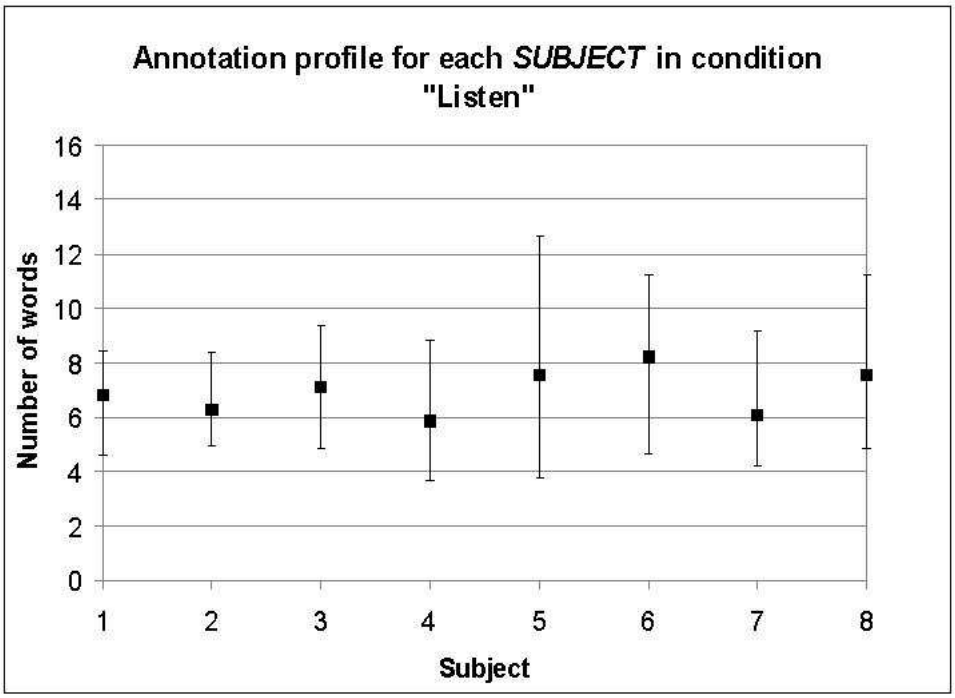

Figure 4.10: Annotation profiles for the subjects, condition Listen.

In figure 4.10 the annotation profiles for the subjects in condition Listen are presented. Horizontally we find the subjects (1–8) and vertically the number of words.

Since the subjects are not the same ones, and not even the same in number, as in the subjects' annotation profile in condition Read there is no point in comparing the individual subjects. The subjects' mean annotation frequency lies between 6 and 8,

which gives a slightly smaller variation than in the subjects' mean annotation profiles in condition Read, where the variation was between 5 and 8. These figures indicate that the extra information in form of access to the speech signal reduced the variation between the subjects' boundary marking, at least in terms of the length of the interval between the punctuation sites.

Compared to condition Read the subjects' punctuation marking is slightly more in agreement, both concerning the variation between all subjects' mean annotation frequencies and the range of the subjects' internal variation between transcripts.

In condition Read the monologues, in particular one of the non-scripted ones, contained a greater variation than the dialogues. There was also a difference in the annotation frequency between monologues and dialogues. Are these differences preserved in condition Listen?

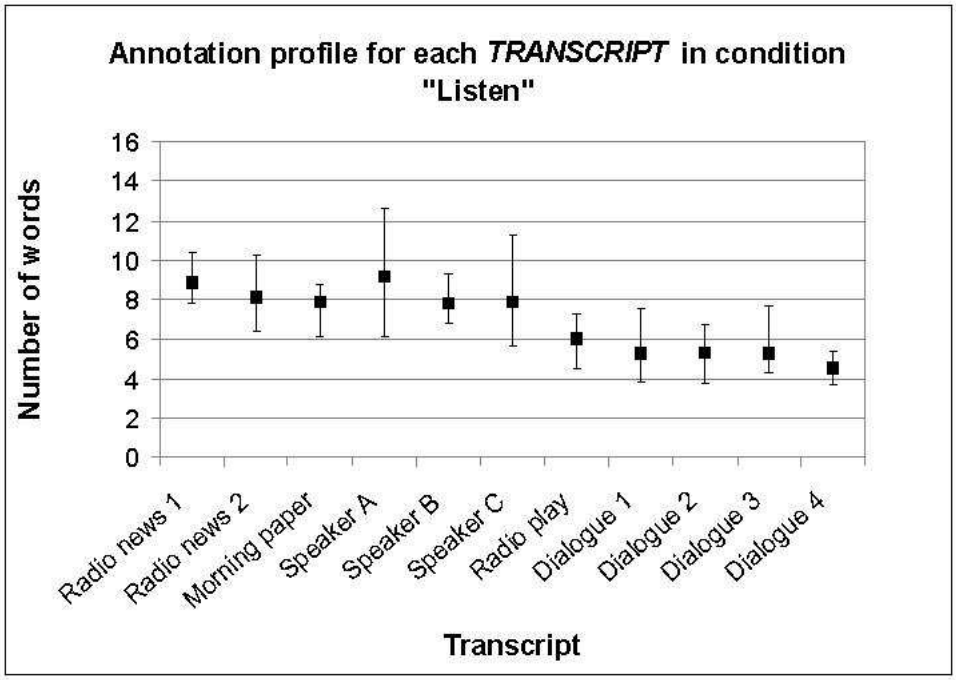

Figure 4.11: Annotation profiles for the transcripts, condition Listen.

Figure 4.11 shows the annotation frequency for the transcripts in condition Listen. Horizontally we find the transcripts (News 1, News 2, etc) and vertically the number of words.

Examining figure 4.11 we find that the same division between the monologues and the dialogues is present as in condition Read. The mean annotation frequencies for the monologues in general have a higher value (from 8 to about 9.5) than the dialogues (from 4.5 to 6). So, there is a higher number of words per punctuation site in the monologues than in the dialogues.

Inspecting the ranges of variations of the means in the transcripts we find that there are two transcripts that clearly differ from the others. Transcripts "Speaker A" and "Speaker C", i.e. two of the non-scripted monologues, show a considerably greater range of variation than the others. Between the ranges of variation in the other transcripts we do not find any dramatic differences, they are rather similar to each other.

4.3.3 Summary of the Annotation Profiles for the Boundary Annotation Task

Examining the annotation profiles, we find that the mean annotation frequencies by the subjects show similar properties in both conditions Read and Listen. Thus, comparing conditions Read and Listen we note that there is no dramatic difference between the subjects' mean annotation frequencies.

Comparing the annotation profiles for the transcripts in condition Read and condition Listen, we again see that the annotation profiles are rather similar. However, the non-scripted monologues seem to give rise to a greater variation in the subjects' mean annotation frequency than the other speaking styles. Perhaps this indicates that these transcripts were more difficult than other transcripts. Furthermore, the division between the monologues, having more words per punctuation site, and the dialogues, having fewer words per punctuation site, is becoming more distinct in condition Listen than in condition Read.

With the annotation profiles, we have a first picture of the subjects' annotations. Examining the annotation profiles for the transcripts it becomes clear that we can expect different results for the monologues compared with the dialogues. In addition, the non-scripted monologues are characterised by the greatest variation between subjects. However, in addition to such qualitative observations we also need to make precise quantitative comparisons of the annotations. For the measuring of inter-annotator agreement we have chosen to use the κ statistics, which we describe more closely in the next section.

4.4 The κ Statistics

As described in the previous section, the results will be based on the subjects' annotations, and therefore we have to measure the degree of agreement between them. For the measuring of inter-annotator agreement we use the κ (kappa) statistics. In this section we give a description of how to compute the κ coefficient, as well as a brief description of threshold values for κ.

In recent years the κ statistics have frequently been used for measuring inter-annotator agreement in NLP tasks such as testing tag-sets (Carletta *et al.*, 1997), (di Eugenio *et al.*, 2000), or checking to what extent annotators agree about the classification of

anaphora antecedents (Poesio and Vieira, 1998). The κ statistics are also used in other disciplines, for example in medicine and in computer science.

The κ statistics are suitable for computing inter-annotator agreement in classification tasks, i.e. where a number of subjects have to classify a number of points according to a number of specific choices. The formula for how to compute the κ coefficient is shown in equation 4.1.

$$\kappa = \frac{P_A - P_E}{1 - P_E} \quad (4.1)$$

In equation 4.1 P_A is the proportion of the annotators' agreement, and P_E is the random agreement. The result of a κ measurement is a value between 1 and -1, where 1 means complete agreement, 0 means random agreement, and -1 means complete disagreement. Thus, the higher the κ value, the better the inter-annotator agreement. The better the inter-annotator agreement, the more objective the annotations can be said to be.

No.	**Transcript**	**1**	**2**	**3**	**4**	**5**	**6**	**English**
1	och							and
2	dethär							this
3	är							is
4	dagens							today's
5	eko	,						echo
6	kvart							quarter
7	i							to
8	fem	.	:	.	.	.//	:	five
9	nya							new
10	storbanken	,		,	;		–	big-bank-DEF
11	finland							finland
12	tar							takes
13	första							first
14	steget	.	,	.	;	,	,	step-DEF
15	giftskandalen							poison-scandal-DEF
16	...							...

Table 4.4: A sample of an annotation matrix with subjects' punctuation.

As well as showing how to compute the κ coefficient, we also show in detail how the subjects' annotations were converted into the *punctuation sites*. In the illustration of how to compute κ, we follow the description given by Poesio and Vieira (1998) which means that we use the method for computing κ suggested by Siegel and Castellan (1988).

The start is the matrixes over the subjects' annotations. In table 4.4 a sample of such a matrix (an extract from the scripted monologue, condition Read) is shown. The subjects' annotations are inserted into a database similar to those we described in section

4.1.4. Starting from the left we find the line numbers (column "No."), the transcript (column "Transcript"), the columns containing the subject's annotations (columns "1"–"6" representing subject 1–6), and an English translation (column "English").

The punctuation marks are annotated after the word rendered in the column "Transcript". On line eight in column "Transcript", we find the word "fem", and on line eight in column "1" we find a full stop. This means that subject 1 inserted a full stop after the word "five" in the transcript. On line eight we also find punctuation marks in columns "2"–"6", which means that also subjects 2–6 inserted punctuation after the actual word. We also see that subject five (column "5") has inserted a sequence of punctuations on line eight: .// (full stop, slash, slash). This indicates that subject 5 annotated "full stop, paragraph boundary" after the word "fem".

Even though the annotators in this excerpt seem to have a similar idea of where to put the punctuation marks they do, however, differ in the actual choice of mark. Since we are interested primarily in *where* the punctuation marks are inserted, we chose to convert the actual signs into one common sign. The result of this conversion is shown in table 4.5.

No	**Transcript**	**1**	**2**	**3**	**4**	**5**	**6**	**English**
1	och	0	0	0	0	0	0	and
2	dethär	0	0	0	0	0	0	this
3	är	0	0	0	0	0	0	is
4	dagens	0	0	0	0	0	0	today's
5	eko	1	0	0	0	0	0	echo
6	kvart	0	0	0	0	0	0	quarter
7	i	0	0	0	0	0	0	to
8	fem	1	1	1	1	1	1	five
9	nya	0	0	0	0	0	0	new
10	storbanken	1	0	1	1	0	1	big-bank-DEF
11	finland	0	0	0	0	0	0	finland
12	tar	0	0	0	0	0	0	takes
13	första	0	0	0	0	0	0	first
14	steget	1	1	1	1	1	1	step-DEF
15	giftskandalen	0	0	0	0	0	0	poison-scandal-DEF
...	...							...

Table 4.5: A sample of an annotation matrix with punctuation sites.

In the table 4.5, all punctuation marks at a specific position are converted into "1", meaning "marking". The previously blank cells have got a "0", meaning "no marking". We now regard the matrix as representing a binary choice made by the annotators, i.e. they have chosen to mark, or not to mark. Thus, the annotations are divided into two categories: "positive" (marked) and "negative" (not marked). These two categories are

found in the two columns "Positive" and "Negative" in 4.6. These columns show the total of positive and negative assessments for each row. For instance, for line eight we saw in 4.5 that all six subjects had inserted punctuation marks. If we look at line eight in table 4.6, we find "6" in the column "Positive" and "0" in the column "Negative". We come back to the values in the last row, N, Pos, Neg, and Z, and what the column S_i stands for.

No	**Transcript**	**Positive**	**Negative**	**S_i**	**English**
1	och	0	6	1	and
2	dethär	0	6	1	this
3	är	0	6	1	is
4	dagens	0	6	1	today's
5	eko	1	5	0.666667	echo
6	kvart	0	6	1	quarter
7	i	0	6	1	to
8	fem	6	0	1	five
9	nya	0	6	1	new
10	storbanken	4	2	0.466667	big-bank-DEF
11	finland	0	6	1	finland
12	tar	0	6	1	takes
13	första	0	6	1	first
14	steget	6	0	1	step-DEF
15	giftskandalen	0	6	1	poison-scandal-DEF
N = 15	...	**Pos = 17**	**Neg = 73**	**Z = 14.13333**	...

Table 4.6: Matrix over the categories "Positive" and "Negative".

We shall now proceed to compute the κ coefficient based on the categories "positive" and "negative". An overview of the elements needed for computing κ is shown in table 4.7.

S_i	The percentage of agreement computed per assessment (row)
Z	The global percentage of agreement
T	The total number of assessments
N	The number of assessments
PA	The proportion of agreement based on the subjects annotations
PE	The proportion of random agreement

Table 4.7: The elements used for computing κ.

To compute the κ coefficient, we start by computing S_i for each row in the matrix. The formula for S_i is given in equation 4.2

(4.2) $S_i = \frac{1}{c(c-1)} * \sum_{j=1}^{m} n_{ij}(n_{ij} - 1)$

In example 4.3 and 4.4 we show the actual figures for rows 4 and 5 in columns "Positive" and "Negative" in table 4.6.

(4.3) $S_4 = \frac{1}{6(5)} * [0 + 6(5)] = (\frac{1}{30}) * 30 = 1$

(4.4) $S_5 = \frac{1}{6(5)} * [1(0) + 5(4)] = (\frac{1}{30}) * 20 = 0.666667$

Based on S_i the Z is computed as the sum of all S_i. For our sample $Z = 14.13333$ (see table 4.6), column S_i, last row.

In the next step we compute PA, the degree of the annotators' agreement. The equation as well as the full calculation for our excerpt is given in equation 4.5

(4.5) $PA = \frac{Z}{N} = \frac{14.13333}{15} = 0.94$

Thus, the Z value is divided by the number of assessments, N, which is found in the last row of column "No." in 4.6.

Next PE, the proportion of random agreement, is computed. To do this we need T the total number of classification assessments, i.e. $subjects * examples$. In our excerpt $T = 6 * 15 = 90$. We also use the sum of the annotations in each category, found in the last row of the columns "Positive" and "Negative". In equation 4.6 we show the equation, as well as the calculation for our excerpt.

(4.6) $PE = (\frac{Pos}{T})^2 + (\frac{Neg}{T})^2 = (\frac{17}{90})^2 + (\frac{73}{90})^2 = 0.0357 + 0.6579 = 0.6936$

In the last step we get the κ coefficient for our excerpt. The equation and calculation is given in 4.7.

(4.7) $\kappa = \frac{PA-PE}{1-PE} = \frac{0.94-0.6936}{1-0.6936} = \frac{0.2464}{0.3064} = \mathbf{0.80}$

Thus, the κ coefficient for the extract, presented in different forms in tables 4.4, 4.5 and 4.6, is 0.80.

What does, then, 0.80 mean? In general a κ value of 0.80 or over is taken to indicate a reliable agreement (the reader is encouraged to study the matrix in table 4.5, to get an impression of "a 0.80 κ agreement"!). However, the guidelines for interpreting κ differ between disciplines. In next section some of these differences are briefly described.

4.4.1 Guidelines for Interpreting κ

The guidelines for interpreting the κ value vary slightly between researchers in different disciplines. There is a wide range of suggested benchmarks, and therefore we offer a brief survey of benchmarks in three different disciplines: medical methodology, software engineering and natural language processing.

Landis and Koch (1977) have suggested a benchmark for use of κ in medical research. The task in this field could be for example to have a number of researchers choose one diagnosis from a set of diagnoses for a patient, having been given specific information about the patients health. The benchmark suggested by Landis and Koch (1977) is rendered in table 4.8.

κ value	Strength of agreement
≤ 0.00	Poor
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
≥ 0.80	Almost perfect

Table 4.8: Threshold values for κ according to Landis and Koch (1977).

As shown in 4.8 Landis and Koch (1977) have covered the whole range of κ between 0 and 1. Compare this with a benchmark for software engineering suggested by El Emam (1999). These threshold values are rendered in table 4.9.

κ value	Strength of agreement
≤ 0.45	Poor
0.46 – 0.62	Moderate
0.63 – 0.78	Substantial
≥ 0.78	Excellent

Table 4.9: Threshold values for κ according to El Emam (1999).

These threshold values differ from the values given by Landis and Koch (1977), most distinctly in the fact that El Emam classifies everything under 0.45 as a poor agreement. However, the remaining categories are relatively similar in interval to the ones suggested by Landis and Koch (1977).

The third example of threshold values comes from Carletta *et al.* (1997), who follow guidelines given by Krippendorff (1980). The work of Carletta *et al.* (1997) is in the field of natural language processing and concerns the test of a tag-set for dialogue coding. If we examine the threshold values used by Carletta *et al.* (1997), rendered in 4.10, we

find that they are even one step stricter than those established by El Emam (1999) in their use of κ.

κ value	Strength of agreement
0.67 – 0.79	Tentative conclusions can be made
$\geq$ 0.80	Reliable

Table 4.10: Threshold values for κ used by Carletta et al.(1997).

Thus, the suggested thresholds for interpretation of the κ value differ between researchers. There are stricter and less strict approaches, but it can be concluded that there seems to be a consensus among all researchers that values over 0.80 indicate a strong inter-annotator agreement and thus gives us a reliable classification.

Carletta *et al.* (1997) mention the interval 0.67 to 0.80 as allowing tentative conclusions. El Emam (1999) states that values in the interval 0.63 to 0.78 indicates a substantial agreement. Thus, both agree that the lower boundary for the second best category lies close to 0.65, whereas Landis and Koch (1977) suggest a slightly lower value: 0.60.

Carletta *et al.* (1997) do not accept a boundary lower than 0.67 as indicating any acceptable agreement. However, both El Emam (1999) and Landis and Koch (1977) use lower boundaries. El Emam labels the interval 0.45 to 0.62 as "moderate", and Landis and Koch use the same label for the interval 0.40 to 0.60.

In general, values over 0.80 indicate a high degree of inter-annotator agreement, and from 0.60 – 0.65 to around 0.80 a fairly good degree of agreement. Values in the range 0.40 –0.60 indicate a lower degree of agreement and do not give us a reliable annotation, while values lower than 0.40 indicate a poor degree of agreement.

We should keep in mind that the values for inter-annotator agreement have two dimensions of interpretation. The first concerns the question whether we can use the results from the annotations and build further on them. In this case we need high values, indicating a reliable degree of agreement and motivating validity in the classifications. The second dimension concerns the question of the annotators' degree of agreement under different circumstances, and we are primarily interested in differences between degrees of agreement in different conditions, i.e. how do different features in the classification task affect the agreement. In this case also lower values are of interest. Thus, if we want to motivate a certain categorization we need high values, but if we want to compare how different features influence the agreement, we can also use lower values.

The different threshold values in the above survey originate from how different disciplines use κ. This indicates that the usability of a certain threshold value varies between different tasks. For instance, Carletta *et al.* (1997) use the κ statistics to evaluate a given tagset. A high degree of agreement in the tagging shows that many of the subjects agreed about the analysis. This increases the probability that the tagset is feasible and

that that conclusions about the material drawn on the basis of the tagging are valid. If we want to use the κ value to motivate such a claim, a high value is needed.

Examining the task described by El Emam (1999), we find a difference compared with Carletta *et al.* (1997). El Emam (ibid) uses κ to evaluate a software process assessment, i.e. the evaluation of a software developing process. In a process assessment task, process requirements and process performance within an organisation are analysed and evaluated. In this particular case, the κ statistics are used to find points of strength and of weakness in the process. The κ values in this case function as a help concerning which points in the process should be improved. This means that conclusions are drawn about the agreement, instead of based on the agreement. To make statements about the degree of agreement does not require the same high κ value as if one wants to draw conclusions based on the agreement.

In the third case, the use of κ in medical research, the κ statistics are used to measure the agreement between researchers diagnosing a patient on the basis of a number of given criteria. This task is more similar to the task described by El Emam (1999) than to the task described by Carletta *et al.* (1997). This means that also in medical research the κ statistics are used to trace both weaknesses and strengths in the agreement, and in this way find which criteria give rise to agreement and which features give rise to disagreement. The κ value is thus not primarily used to motivate the "best" diagnosis but to study how researchers make use of the different diagnosis criteria. The differences between the use of the κ statistics in the tasks in these three fields of research both motivate and explain the differences in the threshold values given for κ.

In this thesis, we primarily use the kappa statistics for quantifying inter-annotator agreement and thus make it possible to compare degrees of agreement across conditions and speaking styles. This means that we do not use κ-values primarily in order to tell whether an annotation is "valid" or not, but rather as an index to make the results comparable to each other. With this use of κ, it is not a problem to use very low κ values, which would be the case if they were primarily used to prove a valid and objective annotation. In addition, this use of κ also removes some of the problems pointed out by di Eugenio and Glass (2004) with Siegel's and Castellan's (1988) way of computing κ compared with the method suggested by Cohen (1960).

4.5 Summary Chapter Four

In this chapter an overview of a number of rather disparate features from the preprocessing of the data is given. However, conversions and simplifications were made in order to make the data as comparable as possible across speaking styles. It is important to be aware of these in order to get a clear picture of the results of the studies. Here we have shown the principles according to which we have built the database and converted the subjects' original annotations into a more usable representation. In addition, we have

introduced an overview of the boundary annotation data as well as given a description of the κ statistics and how we use the κ values in this study. Thus equipped we can proceed to the studies.

Studies

Chapter 5

Discourse Boundaries

IN this chapter we study the boundary aspect of the discourse segments. Our aim is to account for selected boundary features in the different speaking styles, thus creating a picture of the differences and similarities across the four speaking styles in our data. In addition we study how access to the speech signal influences the subjects' annotations in condition Listen compared with condition Read.

5.1 Introduction to the Boundary Annotation Task

In this thesis, our central hypothesis is that the prosodic features of a discourse are related to the structural properties of the same discourse. Variation in structure would then be accompanied by variation in prosody, and this would in turn affect the strategy by discourse segmenting. However, transferring the view on prosodic phrasing conveyed by Bruce (1998), a discourse segment includes both the aspect of the boundary and the aspect of the content in between. In this chapter we account for the boundary aspect of the segmenting of our four speaking styles.

The study of boundary characteristics in our data is not primarily intended to reveal new findings about boundary marking. Our aim instead is to record a number of characteristics of the speaking styles in order to enable a comparison with the prominence marking study to be made, and thus to give a fuller picture of how a disocurse segment might change under different stylistic conditions. Thus, the focus in our study is on the comparative aspect. Our hypothesis concerns the relationship of the features from the boundary annotation to the linguistic structure in the disocurse theory of Grosz and Sidner (1986).

Many researchers have studied prosodic differences across speaking styles, e.g. Strangert (1992), Swerts (1997) and Hirschberg (2000). In all these studies significant differences in the pause pattern between speaking styles are reported.

Transcription	Parse	English
och	O	*and*
dethär	NPB	*this*
är	VCB	*is*
dagens	NPB_NPMAXB	*today's*
eko	NPI_NPMAXI	*echo*
kvart	NPB_NPMAXI	*quarter*
i	PPB_NPI_NPMAXI	*to*
fem	NUMPB_NPB_PPI_NPI_NPMAXI	*five*
.	O	.
nya	APMINB_NPB	*new*
storbanken	NPI	*big-bank*-DEF
,	O	,
finland	NPB	*finland*
tog	VCB	*took*
första	NPB	*first*
steget	NPI	*step*-DEF
.	O	.

Figure 5.1: Excerpt of the parse of a scripted monologue.

Both Fant and Kruckenberg (1989) and Strangert (1992) have reported a relationship between longer pauses and stronger boundaries in Swedish, and this was also observed by Swerts (1997) for Dutch. Thus, pausing is one of the strongest boundary cues in prosodic phrasing, and therefore we have selected pauses as the acoustic correlate to be examined. We are not examining features like change in F0 or final lengthening, which are also correlates to prosodic phrasing (see e.g. Swerts (1997), Bruce (1998) etc.) but are focusing on silent pauses.

As linguistic aspects of the string of words we have selected the phrasal and the part-of-speech context of the boundary annotations. This means that for each boundary we examine the phrasal category of the word after the boundary as well as the part-of-speech of the same word. Concerning the phrases we examine only the phrasal category immediately dominating the word, and not phrasal categories higher up in the discourse tree. Thus, upon reexamining the example from chapter 4 recapitulated in figure 5.1 this means that the phrase category examined is the very left phrasal label in the column "parse" (in this example O, NPB, VCB, NPB, NPI etc) while the other ones are discarded.

In order to get a measure of the degree of nesting in the different speaking styles, we examine the phrase depth, or rather the parse depth. This is done through a recording of

the nestedness of the parses in the parse. Thus, re-examining figure 5.1, we find that the parse depth is a record of the number of words with one phrase label attached (e.g. O, NPI, VCB), the number of words with two phrase labels attached (e.g. NPB_NPMAXB, NPI_NPMAXI) etc. It should be stressed that since we have a shallow parse, this does not give a total picture of the phrase depth. However, it offers a good base for comparison of parse depth across the speaking styles. Even though the term parse depth might be more suitable, we use the term phrase depth since the measurement is an indication of phrase depth.

The issues addressed in this chapter are: 1) Do the characteristics of boundary annotations differ across speaking styles within conditions Read and Listen? In order to study this we examine the prosodic and linguistic features described above in the context of the boundary annotations and 2) Do the characteristics of boundary annotations differ between the same speaking styles across conditions Read and Listen? This question is studied by comparing the subjects' boundary annotations in condition Read compared with the boundary annotation in condition Listen.

The chapter is structured as follows: in section 5.2 we present the level of inter-annotator agreement for the boundary annotation task. We examine both the general level of inter-annotator agreement for the subjects within conditions, i.e. to what extent did subjects agree on the marking in condition Read while annotating boundaries on the basis of transcripts alone, and to what extent did they agree in condition Listen, while annotating with access to the speech signal? In addition we compute the level of inter-annotator agreement across conditions using the boundaries annotated by the majority in each condition. In this case we study to what extent subjects in condition Read and in condition Listen agreed on the boundary marking.

In the first computation of the level of inter-annotator agreement we use all the boundary annotations made by all the subjects. However, when examining the level of inter-annotator agreement across conditions or the prosodic and linguistic features, we do not classify a boundary annotated by one subject as equal to a boundary annotated by eight subjects. Instead we have chosen to use the boundaries annotated by the majority of subjects, i.e. at least four subjects in condition Read and at least five subjects in condition Listen. In addition, all overlapping speech and file endings, where we could assume that subjects were largely influenced by the transcription or technical aspects of the recording, were disregarded in the analyses.

In section 5.3 we examine the phrase context of the boundary annotations in conditions Read and Listen. First we examine the phrase depth and then the phrase labels.

The part-of-speech contexts of the boundary annotations are examined in detail in section 5.4. In this section we study the part-of-speech label on the word immediately following a (majority) boundary annotation. In both the study of phrase labels and the study of part-of-speech labels, we search for different preferences for boundary indicators across the speaking styles. If there are such different preferences, it might be possible to

express this as different segment markers in the speaking styles, relating the findings to the marker hypothesis expressed by Green (1979).

In the section 5.5 we examine the boundary annotation context in terms of silent pauses of the majority boundary annotations in condition Listen. Thus, we examine only condition Listen, since it is obvious that the speech signal including the pauses did not influence the boundary annotations in condition Read.

5.2 Level of Inter-Annotator Agreement for the Boundary Annotation

Upon examining the data, it immediately became clear that the subjects did not agree in their use of specific punctuation marks. A certain level of agreement could be found in the use of question and exclamation marks, and the substudie of question segments is reported in 7). However, no clear agreement was present in the subjects' use of commas and full stops, which were the most frequently used punctuation marks. Moreover, subjects disagreed on how many punctuation marks to put in a certain position; sometimes one subject annotated a sequence as e.g. ?". (question mark, end of quotation, full stop), while another subject used only a . (full stop). Because of these results we chose to use only the punctuation sites, as described in the previous chapter.

A brief record of the number of boundaries as well as of the number of words for each speaking style is rendered in table 5.1.

	Scripted monologue	Non-scripted monologue	Scripted dialogue	Non-scripted dialogue
# of words	2858	2084	2212	2248
# of punctuation sites, Read	467	573	691	724
# of maj. boundary, Read	278	199	502	426
# of words/maj. boundary, Read	8.12	10.47	4.41	5.28
# of punctuation sites, Listen	542	653	733	789
# of maj. boundary, Listen	309	182	506	379
# of words/maj. boundary, Listen	9.25	11.45	4.37	5.46

Table 5.1: An overview of the boundary annotations.

The rows "# of punctuation sites" in table 5.1 show the total of positions in both condition Read and Listen where at least one subject annotated a boundary, the rows "#of maj. boundary" show the number of positions where the majority have annotated a boundary and the rows "#of words/maj. boundary" show the number of words per majority boundary. Please note that the general level of inter-annotator agreement, i.e.

the κ for Read and Listen, is computed on the basis of the punctuation sites, while the κ across conditions (Compare) as well as all examinations of linguistic and prosodic features use the boundary annotations made by the majority of subjects (# of maj. boundary).

As a crude measure of the articulation rate, the average time (in milliseconds) it takes to pronounce a word is given for each speaking style in table 5.2. In addition the average number of black characters per word is shown for each speaking style.

	Scripted monologue	Non-scripted monologue	Scripted dialogue	Non-scripted dialogue
ms./word	280	317	341	306
char./word	5.23	4.43	4.12	4.20

Table 5.2: Articulation rate in the four speaking styles.

The durations in table 5.2 are computed on the total speaking time in the sound files (minus total duration of silent pauses) divided by the total number of words. The table shows that in the monologues the scripted style is faster than the non-scripted one,while the case is the opposite in the dialogues. Comparing the non-scripted styles the dialogue is faster than the monologue, while the case is the opposite in the scripted styles.

The average number of characters per word indicates that on average the scripted monologue contains the longest words and the two dialogues the shortest, while the non-scripted monologue is placed in between.

Our first question concerns the subjects' agreement on where to annotate a boundary in the different speaking styles in the two conditions Read and Listen. Swerts (1997) reported a higher level of agreement between subjects annotating boundary with access to the speech signal compared with subjects annotating in transcripts alone in a study on Dutch spontaneous monologue. If this also holds for our data, we should see a higher level of agreement in condition Listen than in condition Read.

Below we present the level of inter-annotator agreement within each condition (section 5.2.1), and across conditions (section 5.2.2). In the inter-annotator agreement within conditions κ is computed on the basis of all the annotations, i.e. not only majority annotations, but across conditions we use only the majority boundaries.

5.2.1 Level of Inter-Annotator Agreement Within Conditions in the Boundary Annotation Task

In this section, accounting for the level of inter-annotator agreement within conditions we examine the extent to which subjects agreed on the boundary markings in each

condition and in each speaking style. Thus, we examine if there is a difference in the agreement depending on speaking style – do subjects agree more on where to mark boundaries in e.g. scripted monologue than in e.g. non-scripted dialogue? In addition we examine the extent to which subjects are influenced by the prosody, i.e. do subjects agree more about where to mark a boundary in scripted monologue when they have access to the speech signal than when they annotate boundaries on the basis of the transcripts alone?

It should be stressed that the κ value is used primarily as a reference for comparing the annotations in the different speaking styles, and even if we use expressions such as "high" or "low" value this should not be interpreted as an absolute claim about "over" or "under" specific thresholds.

Starting with condition Read, in table 5.3 we show the level of inter-annotator agreement expressed with κ values for the four speaking styles. We again remind the reader that κ is computed on all annotations by all subjects.

READ	Monologue	Dialogue
Scripted	0.73	0.78
Non-scripted	0.59	0.63

Table 5.3: Level of inter-annotator agreement of boundary annotation in condition Read.

Examining the figures we find that there are fairly high figures for the two scripted styles; scripted monologue has κ=0.73, and scripted dialogue κ=0.78. For the non-scripted speaking styles the figures are lower; κ=0.59 for the non-scripted monologue and κ=0.63 for the non-scripted dialogue. However, in both cases we find a higher level of agreement in the annotation of the dialogues than in the annotation of the monologues. When examining the linguistic and prosodic characteristics of the boundary annotations we have to be particularly careful when drawing conclusions about the non-scripted monologue. This was also the speech sample which contained the highest range of variations in the annotation profiles in the previous chapter. The figures for the non-scripted speech are in line with the figures found in Passonneau and Litman (1997:105) who reported that reliability labelling from spontaneous speech gave κ=0.63. However, in the case of non-scripted speech our figures differ from Passonneau and Litman's 1997, where we have κ=0.73 and κ=0.78 compared with Passonneau and Litman's figures for text read aloud, which were κ=0.56.

Proceeding to condition Listen we ask the question as to whether subjects' access to the speech signal will influence the level of agreement.

Table 5.4 presents the level of inter-annotator agreement expressed with κ for condition Listen. κ is also here annotated on all boundary markings.

Upon examining the level of inter-annotator agreement in condition Listen, it is higher in the scripted speaking styles, in the monologue κ=0.79 and in dialogue κ=0.70. In the

LISTEN	Monologue	Dialogue
Scripted	0.79	0.70
Non-scripted	0.58	0.56

Table 5.4: Level of inter-annotator agreement for boundary annotation in condition Listen.

non-scripted material, the level of inter-annotator agreement measured with κ is lower, κ=0.58 for monologue and κ=0.56 for dialogue. Comparing the figures from condition Listen with the figures in condition Read there is no clear indication that access to the speech should have increased the level of agreement. An increase in the level of agreement is indeed present in the scripted monologue, but in both dialogues the agreement has decreased, and in the non-scripted monologue it is about the same. Thus, our reported results are contrary to the increase Swerts (1997) reported when subjects annotated with access to the speech signal compared with when they annotated on the basis of text alone.

The level of inter-annotator agreement differs between the speaking styles in both condition Read and condition Listen. In condition Read, there is a tendency towards a higher level of inter-annotator agreement in the dialogues, while condition Listen shows a very weak tendency towards a lower one. A clear increase in the level of inter-annotator agreement in condition Listen compared with condition Read was present only in the scripted monologue.

5.2.2 Level of Inter-Annotator Agreement Across Conditions in the Boundary Annotation Task

There were no extremely great differences in the level of inter-annotator agreement in the speaking styles when they were compared across conditions Read and Listen, i.e. in the scripted monologue the level of agreement increased in condition Listen, while it decreased in the other speaking styles. However, do subjects agree across conditions? In principle, it would be possible with similar figures for level of inter-annotator agreement in each of the two conditions, but a low level of inter-annotator agreement across conditions, i.e. the subjects would agree on different things within each of the two conditions. However, if there is a high level of agreement across conditions, this means that subjects within both conditions have marked same positions as boundaries. A high level of inter-annotator agreement across conditions would indicate that subjects to a great extent mark boundaries on the basis of the string of words. If the level of agreement is low, this means that the marking on transcripts alone and on transcripts together with the speech signal were different, and this might indicate that the subjects were influenced by the speech signal when marking.

In order to investigate this question we compute level of inter-annotator agreement between the boundary annotations in the two conditions. Our method of measuring the level of inter-annotator agreement across conditions was based on the boundaries annotated by the majority of subjects.

However, selecting just a subset of the annotations, what proportions of data do we use? The "majority" could in some cases include a high number of annotations where all the subjects agree, and only a low number where a minimal majority agree. In other cases "the majority" can mean a smaller proportion of consensus and a greater frequency of minimal majority. In order to keep this variation under control we survey the distribution of annotations and point out what data we have used.

Figure 5.2 shows the distribution of boundaries annotated in condition Read. On the x-axis the number of subjects is presented, and on the y-axis the percentual proportion of boundaries annotated. Thus, it is possible to see what percentage of the boundaries in the different speaking styles is annotated by 6 subjects, by 5 subjects, by 4 subjects etc. The positions in the middle are where we find the highest level of disagreement between the subjects, i.e. about half of the subjects annotate a boundary and half of them do not, while the positions at the ends are the positions where we find the highest level of agreement, i.e. many subjects annotate or do not annotate a boundary. The "majority annotations" we use from condition Read means the annotations made by 4, 5 and 6 subjects, i.e. the positive majority agreement.

Upon examining 5.2 we see that in the scripted speaking styles a high proportion of the boundaries is annotated by 6 subjects, the proportion declines in the middle of the curve (3 subjects), and rises again at the end (1 subject). Thus, there is a relatively great proportion of stronger agreement, more positive (6-5 subjects) and negative (1-2 subjects), than of a weaker one(4-3 subjects). In the non-scripted dialogue the pattern is similar, but the curve is flatter. In non-scripted monologue we find a new pattern; a high proportion of negative agreement. In addition the disagreement interval (4-3 subjcts) is proportionally higher than in the other speaking styles.

Thus, the set of boundaries annotated by the majority of subjects (6-4 subjects) is a fairly great proportion of all the boundary annotations in the scripted speaking styles and in the non-scripted dialogue. In the non-scripted monologue the boundaries annotated by the majority constitute a smaller set of all boundaries annotated. However, for all four speaking styles we still use the majority set for comparison (i.e. annotations made by 4 or more subjects in condition Read), but bear in mind that the majority of subjects differ across speaking styles.

Figure 5.3 shows the distribution of boundaries annotated in condition Listen. In a similar way to condition Read the number of subjects is shown on the x-axis and the proportion of boundary annotations on the y-axis. In condition Listen there are 8 subjects (8-1), and the majority set is thus based on boundary annotations where 8-5 subjects agree.

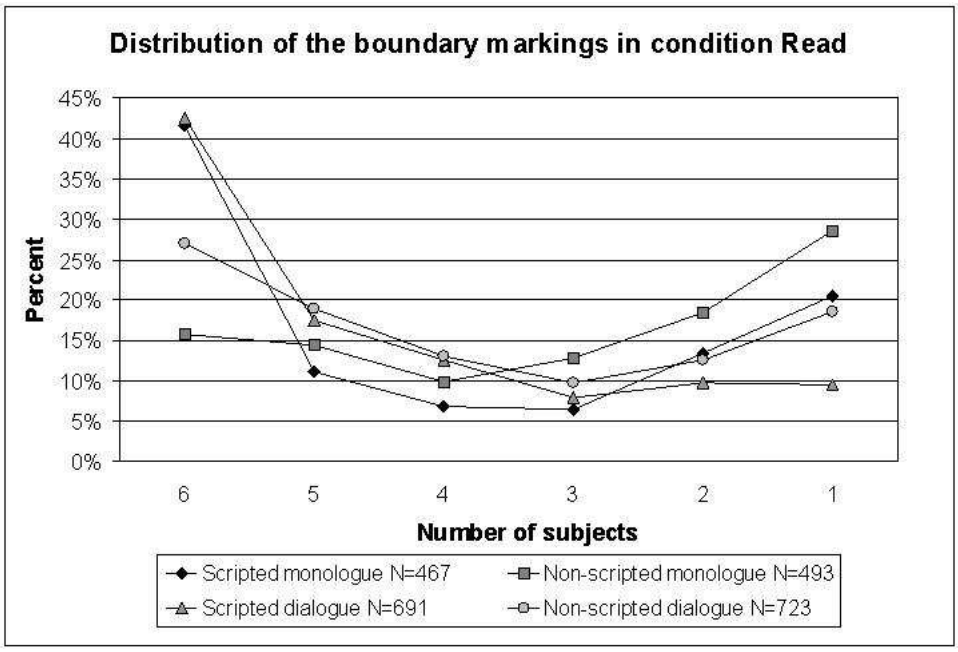

Figure 5.2: Proportion of boundary annotations per subject in condition Read.

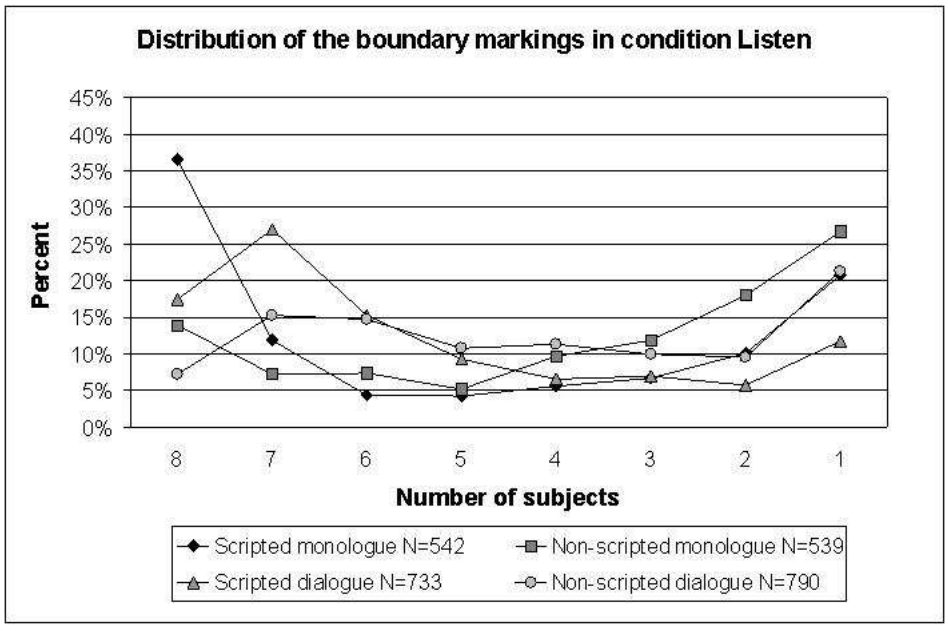

Figure 5.3: Proportion of boundary annotations per subject in condition Listen.

The pattern in condition Listen is different from that in condition Read. The same generally high proportion of positive agreement (8-7 subjects) is not present in condition Listen, the curves are in general flatter, and there is a higher proportion of negative agreement (2-1 subjects). Thus, the majority set includes a smaller proportion of boundaries in condition Listen than in condition Read. The proportion of boundaries with total agreement is clearly lower in condition Listen than in condition Read, especially in the dialogues. However, this might just indicate that one subject deviated in the annotation and that the occasions with total agreement between subjects were fewer.

The comparison of the annotations in condition Read and Listen is carried out by computing the level of inter-annotator agreement in terms of κ between the majority set in condition Read and the majority set in condition Listen. A high κ value indicates that the subjects agree on the same positions for the punctuation marks in condition Read as in condition Listen. A lower level of agreement indicates that the subjects agree about different positions in condition Read from those in condition Listen. The results from this comparison are shown in table 5.5

COMPARE	Monologue	Dialogue
Scripted	0.87	0.85
Non-scripted	0.71	0.77

Table 5.5: Level of inter-annotator agreement for boundary annotation (majority agreement) between conditions Read and Listen.

As shown, there is rather a high level of inter-annotator agreement between conditions Read and Listen, especially for scripted speaking styles ($\kappa = 0.87$ for scripted monologue and κ=0.85 for scripted dialogue). The non-scripted speech has lower figures, but they are still fairly high with $\kappa = 0.71$ for the non-scripted monologue and κ=0.77 for the non-scripted dialogue.

The κ values given in table 5.5 should, however, be very carefully interpreted. The figures are based on the boundary annotations where the majority of subjects agreed in the two conditions. Thus, the figures given in table 5.5 are not to be interpreted in the same way as the figures in tables 5.3 and 5.4, where the level of inter-annotator agreement was based on all the annotations from each subject in each condition. Since all minority annotations are removed, they were instead converted into an "artificial 0-agreement", biasing the result in a positive direction.

The decision to do this filtering was made since we wanted to base the figures on boundaries which could be assumed to be fairly solid across subjects, and thus investigate if we could see a solid tendency towards similarity or a difference in the annotations between the two conditions. If we did not carry out this filtering, boundaries in each condition made by only one subject would perhaps have influenced the result, and all these occasional annotations would result in figures which give a picture of low level of agreement. There is another consequence of this filtering of the data. Swerts (1997) states that

a higher level of agreement among annotators indicates stronger boundaries. In turn, stronger boundaries tend to correlate with boundaries higher up in the discourse tree, e.g. sentence and paragraph boundaries. Focusing on the majority boundaries we can presume that this set of boundaries consists of stronger ones, and thus makes better candidates for discourse segment boundaries than the weaker boundaries found within sentences.

The figures in table 5.5 indicate a generally high level of agreement between conditions Read and Listen. Examining the figures in 5.5 we find that there is a higher level of inter-annotator agreement in the two scripted speaking styles (monologue, κ=0.87 and dialogue κ=0.85) than in the non-scripted ones, i.e. in the scripted styles subjects mark the same positions as boundaries to a greater extent than in the non-scripted ones.

To sum up the three measurements of level of inter-annotator agreement: The *scripted monologue* has a higher level of inter-annotator agreement in condition Listen than in condition Read. It thus seems that access to prosody has influenced the subjects towards a greater agreement in the annotations, κ=0.73 in Read and κ=0.78 in Listen. This effect is not found in *non-scripted monologue* which has κ=0.59 in Read and κ=0.58 in Listen. In the *scripted dialogue* we note the opposite effect, i.e. the κ is higher in condition Read (κ=0.78) than in condition Listen (κ=0.70). The effect is the same in the *non-scripted dialogue* with κ=0.63 in condition Read and κ=0.56 in condition Listen.

The level of agreement across conditions is in general high, indicating that the majority annotations to a great extent are found at the same positions across conditions. Thus, it seems that having access to the speech signal has not influenced the subjects towards annotating very differently from those who annotated on the basis of text alone. In other words, much of the information used in the segmentation task is present in the string of words, at least for the majority of the annotations.

5.3 Phrase Level Properties in the Speaking Styles

In order to obtain a more detailed picture of some characteristics of the boundary annotations we examine a number of specific boundary contexts starting with the phrase context. The phrase mark-up is based on the (shallow) parsing, described in chapter 4. To study the phrase context we removed all the boundaries annotated by the majority of subjects, and examined the phrase context immediately following the boundary. Again we stress that in the boundary study, from now on we use the boundaries annotated by the majority of subjects, and for the sake of brevity we just call them "boundaries".We study four aspects of the phrase in each speaking style in conditions Read and Listen:

- The general phrase depth in the speaking style.
- The phrase depth at the boundaries in the speaking style.

- The general phrase frequencies in the speaking style.
- The phrase context of the boundaries in the speaking style.

The two aspects of the phrase depth are studied in section 5.3.1, the general phrase frequencies in the speaking styles in section 5.3.2, and the phrase context of the boundaries in section 5.3.3.

The phrasal and part-of-speech context is examined for the boundaries in the rows "# of maj. boundaries" in table 5.1 in both condition Read and Listen, whereas the pause context is considered only for condition Listen (the boundaries in the row "#of maj. boundary, Listen").

5.3.1 The Phrase Depth in the Speaking Styles and as Boundary Annotation Context

In the boundary annotation profiles in chapter 4 it was shown that the boundary annotation frequency differed between the monologues and the dialogues, i.e. there was a higher number of words per boundary in the monologues than in the dialogues. This difference in boundary annotation frequency might be related to the phrases in the monologues being longer than those in the dialogues. To investigate this question we study the distribution of the phrase depth based on the shallow parse in the different speaking styles. Starting with the general phrase depth, figure 5.4 shows the distribution of the phrase depth for each speaking style.

Figure 5.4 shows considerable similarities across the speaking styles. In all four styles the majority of words are involved in a phrase with depth 1, but there are slight differences between the speaking styles. In the scripted monologue 60% of the words are located to a phrase of depth 1, while the figure for the non-scripted monologue is 72%. The scripted dialogue has the highest figure with 87% of the words in a phrase of depth 1, with the non-scripted dialogue lies on 75%. At phrase depth 2 the scripted monologue has the highest frequency (20%), followed by the non-scripted monologue (11%), the non-scripted dialogue (10%) and the scripted dialogue (6%). A similar pattern across the speaking styles is found for the depths 3, 4 and 5. This indicates a higher proportion of longer parses containing nested phrases in the scripted monologue.

Comparing the figures of the phrase depths for the words in the context after a boundary in condition Read, we find even greater similarities. Examining figure 5.5 we note that the phrase depth context of boundary markings in the greatest proportion of phrases is of depth 1. This can be interpreted as an indication that the subjects prefer to insert a boundary annotation at the top level of a phrase, i.e. in front of sentence, clause and longer phrase breaks, rather than between more nested phrases. Since the pattern was similar in condition Listen a separate rendering is not made.

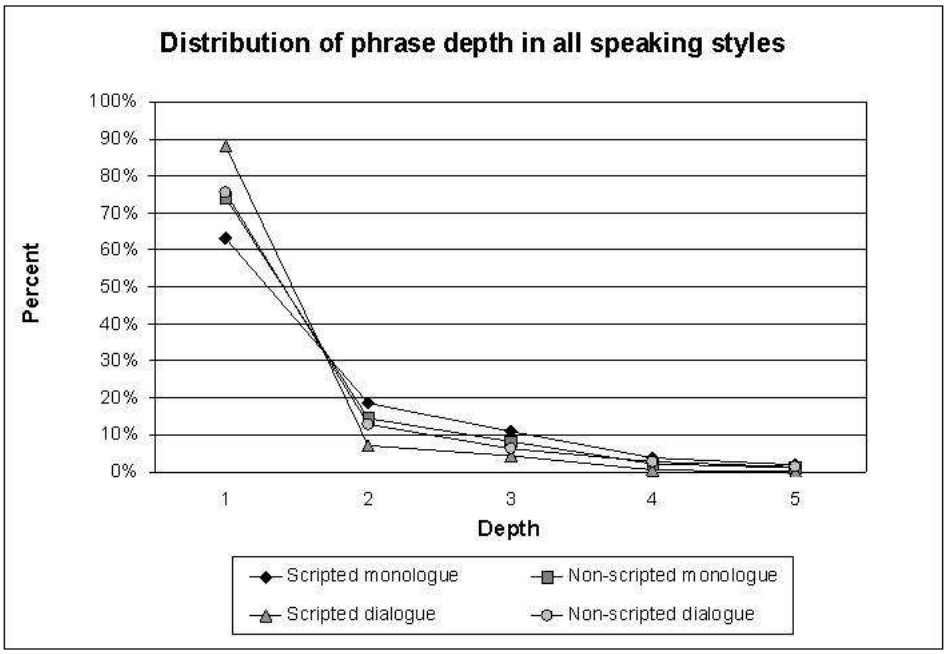

Figure 5.4: General phrase depth distribution in the four speaking styles.

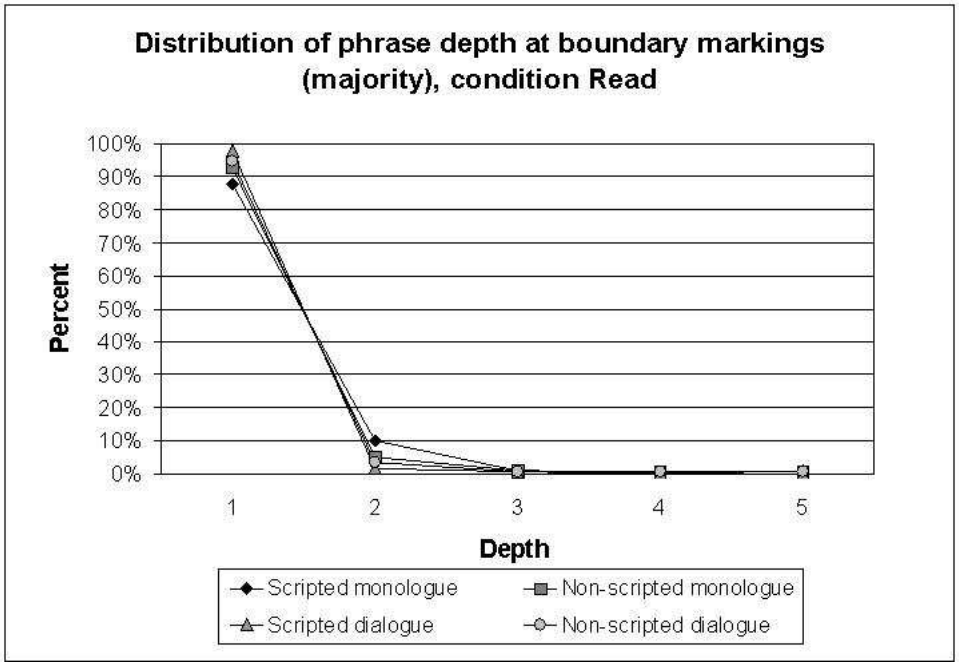

Figure 5.5: Phrase depth distribution at boundary annotations in the four speaking styles.

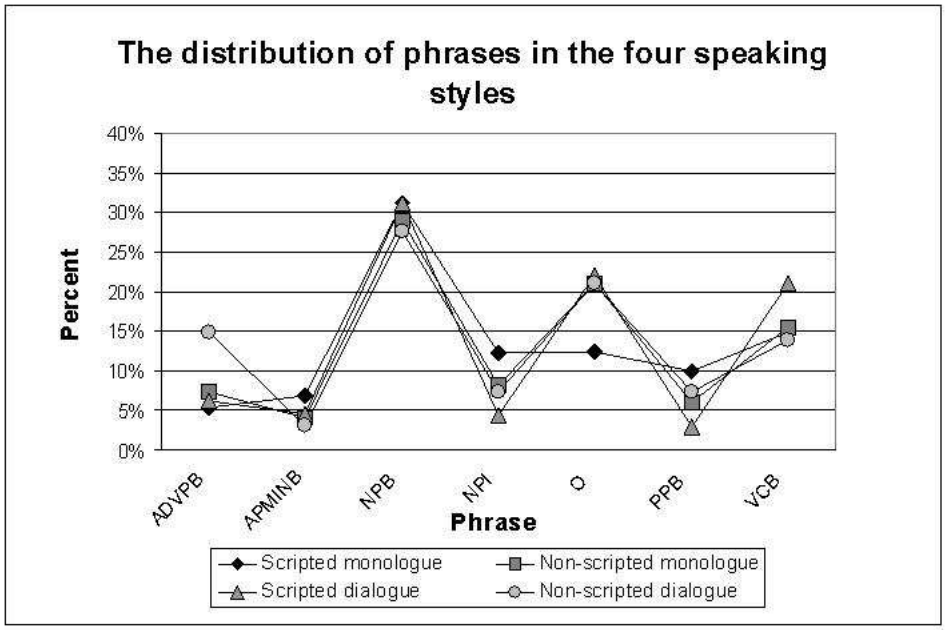

Figure 5.6: The distribution of phrases in the four speaking styles.

Our investigation of the phrase depth thus indicates that the scripted monologue has more nested phrases and the scripted dialogue fewer ones. In all speaking styles the boundary markings are to a great extent in front of a phrase of phrase depth 1. This is a very crude measure of phrase depth, however, it gives an indication of the differences across speaking styles and in addition indicates similarities across conditions Read and Listen.

5.3.2 The Phrasal Distribution in the Four Speaking Styles

The next aspect of phrases to be examined is the distribution of phrase labels in the speaking styles. As described in section 5.1 we study the distribution of the phrase labels immediately attached to a word. In other words, only the frequencies of the phrase immediately dominating a terminal node (a word) is given, and intermediate phrases are disregarded. In this section we study the general phrase frequencies, and in the next section the phrase frequencies of the boundary contexts.

We show phrase frequencies only in cases where the frequency reaches at least 5% in some speaking style. The general distribution of phrases in each speaking style is shown in figure 5.6.

Figure 5.6 shows that the phrase distribution is relatively similar across the four speaking styles, but with some minor exceptions. In the category ADVPB (beginning of adverb phrase) the scripted dialogue has a higher proportion of ADVPB than the other speaking styles. APMINB (beginning of minimal adjective phrase) and NPB (beginning of noun phrase) are similar across the styles while NPI (inside noun phrase) mirrors the phrase depth and is more frequent in the scripted monologue. The scripted monologue also has a lower proportion of O (outside the scope of any phrase, including conjunctions interjections/feedback). In the case of PPB (beginning of preposition phrase) the proportion varies across speaking styles, and in VCB (verb cluster beginning) scripted dialogue has a greater proportion.

The four speaking styles do not show dramatic differences in the phrase distribution. However, the scripted monologue is singled out most prominently as having a lower number of O, the scripted dialogue as having a greater number of VCB, the non-scripted dialogue as having a greater number of ADVP, while the non-scripted monologue in all cases agrees with the majority.

If the general phrase distribution predicts the positions of the boundary annotations, the phrase context of the boundary annotations in the four speaking styles ought to be similar to the general phrase distribution. If the boundary annotations depend instead on more specific boundary preferences, the pattern will differ.

5.3.3 The Phrase Context of the Boundary Annotations

In this section the phrasal context of the boundary annotations is inpected. We first examine condition Read, where figure 5.7 shows the phrase context after the boundary positions annotated by the majority of subjects. Also this figure shows contexts only where at least one speaking style has reached 5%. This delimits the set of phrases as compared with the general distribution, and the phrases that fell out are APMINB (beginning of minimal adjective phrase) and NPI (noun phrase inside).

The high degree of similarity between all four speaking styles which was present in figure 5.6 has now been reduced. There is still a high degree of similarity in the phrase context after boundary annotation in the two non-scripted speaking styles which, however, have a pattern different from the scripted one. Examining figure 5.7 we find the greatest differences in the categories NPB (beginning of noun phrase) and O (outside the scope of any phrase). The non-scripted speaking styles have a low degree of boundary annotations in front of NPB (monologue 21%, dialogue 17%) compared with the scripted speaking styles (monologue 50%, dialogue 40%). In addition the non-scripted speaking styles have a great proportion of boundary annotations in front of O (monologue 57%, dialogue 61%) compared with the scripted speaking styles (monologue 25%, dialogue 39%). The two contexts NPB and O are furthermore the most frequent boundary contexts. This divides the speaking styles into two groups: the scripted styles which have many boundaries in front of NPB (beginning of noun phrase) and the non-scripted styles which have many

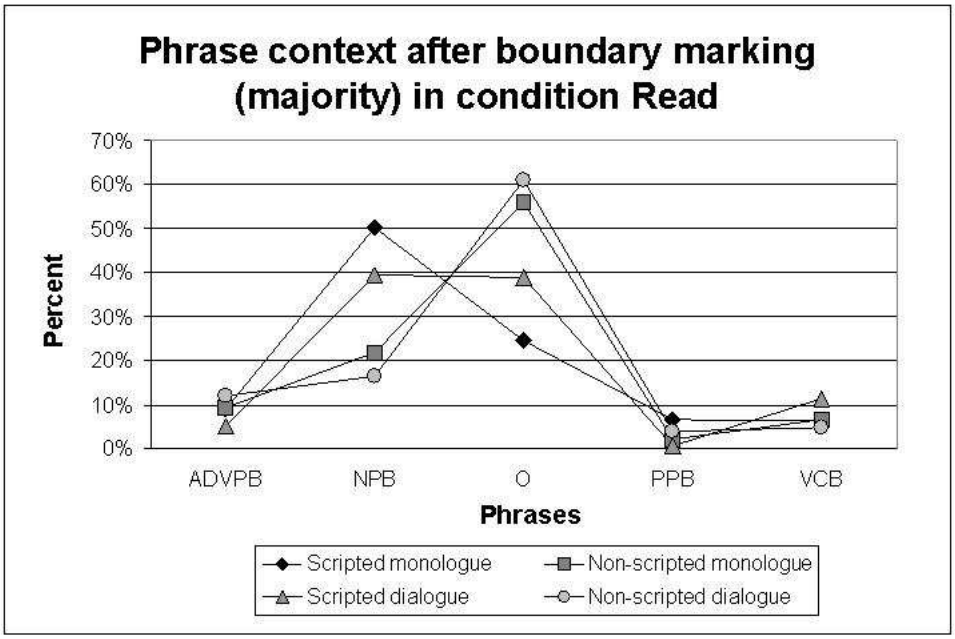

Figure 5.7: Phrase context of boundary annotation in condition Read.

boundaries in front of O (outside the scope of any phrase, here meaning conjunctions and interjections/feedback).

Turning to condition Listen, does the phrasal context of the boundaries differ from that in condition Read? In figure 5.8 we show the distribution of the phrase context of the boundary annotations agreed on by the majority of subjects in condition Listen.

The contexts of the boundary annotations in condition Listen do not differ dramatically from those in condition Read. The low figures for the ADVPB, PPB and VCB are preserved, but some minor changes are seen in the figures for NPB and O.

The figures for boundaries in front of NBP in scripted monologue and scripted dialogue are slightly lower in condition Listen than in condition Read, and also in non-scripted monologue the figures have declined, while the non-scripted dialogue is on the same level. In the case of the category O, the proportion of boundaries in the non-scripted monologue have increased from 55% to 60% while it in the non-scripted dialogue decreased from 60% to 55%. In the scripted monologue the proportion raised slightly from 25% to 29%, while the scripted dialogue still lies on about 40%.

A weak tendency is that in the scripted styles there are slightly fewer noun phrases in the context after boundary annotations in condition Listen than in condition Read, and a slightly higher proportion of O. The non-scripted speaking styles do not show any

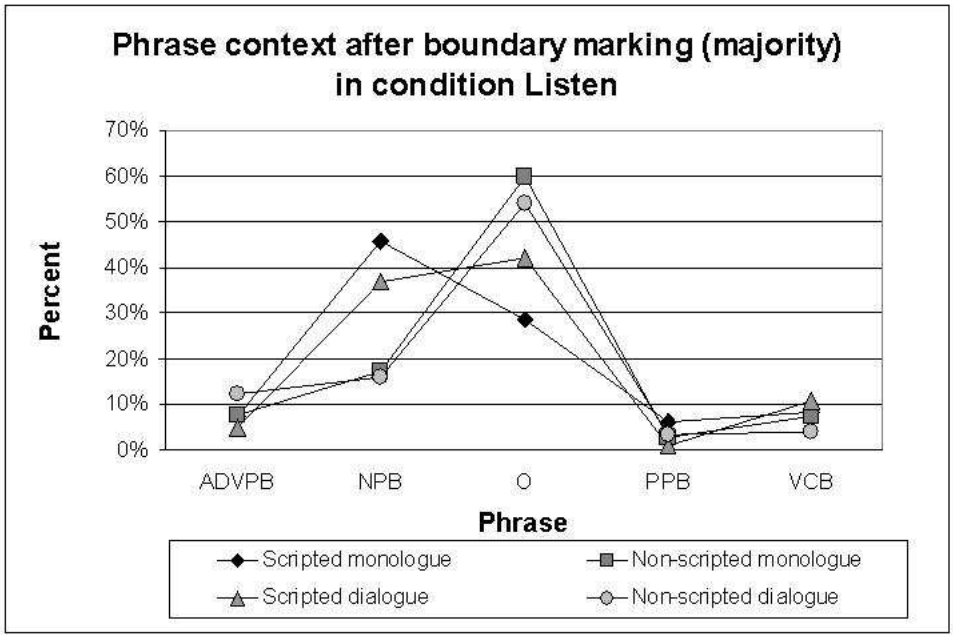

Figure 5.8: Phrase context of boundary annotation in condition Listen.

great differences between condition Read and condition Listen. Nevertheless, we can still single out the scripted and the non-scripted speaking styles as forming roughly two groups where the scripted styles have a higher proportion of NPB while the non-scripted ones have a higher proportion of O.

This indicates a similarity between scripted monologue and dialogue on one hand and non-scripted monologue and dialogue on the other. In a comparison of these results to the NVQ and NQ figures given in figure 4.6 and 4.7, the non-scripted monologue and dialogue also show similarities (NVQ 0.89 and 0.90 respectively and NQ 0.56 and 0.45 respectively). However, this is not the case for the scripted styles (NVQ 1.63 and 0.38 respectively and NQ 1.16 and 0.22 respectively). Why does the phrase context show such a high degree of similarity when the part-of-speech measurements NVQ and NQ show such a difference? To obtain a clearer picture of what lies behind the NBP and O frequencies, we proceed to the examination of the part-of-speech distribution. We examine the same four dimensions as for the phrases in this section, i.e. the general distribution in the styles and the distribution at boundary annotations in conditions Read and Listen.

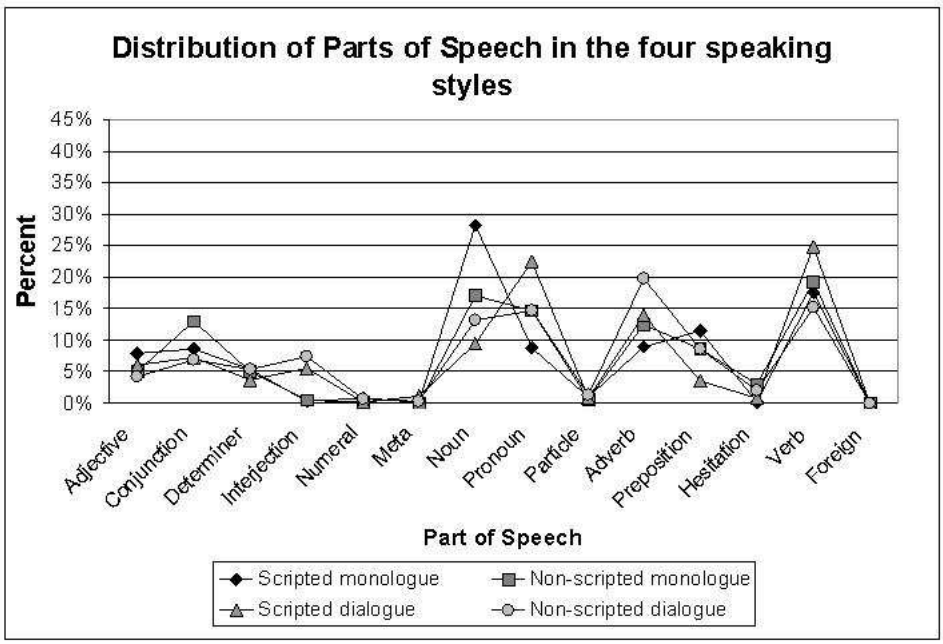

Figure 5.9: PoS distribution in the four speaking styles.

5.4 Word Level Properties in the Speaking Styles and as Boundary Context

Turning to the part-of-speech distributions we start with the general distribution (section 5.4.1) and then proceed to the part-of-speech context of the boundaries in section 5.4.2.

5.4.1 The Part-of-Speech Distribution in the Four Speaking Styles

Figure 5.9 presents the part-of-speech distribution in each of the four speaking styles. The tag-set used is the tag-set "compact", described in chapter 4. Figure 5.9 shows that the differences between the four speaking styles become more obvious in the part-of-speech distribution than in the phrase distribution.

The frequency of numerals, particles, foreign words and the two new tags meta and hesitation is very low in all four speaking styles. The frequency of adjectives in the four speaking styles shows a similar pattern to the APMINB phrases. The same holds for the verbs, the prepositions and to a certain extent also the adverbs. In the case

of the verbs, the non-scripted monologue and dialogue and the scripted monologue lie between 15% and 20% while the scripted dialogue lies at 25%. The frequencies for prepositions are between 4% (scripted dialogue) and 12% (scripted monologue), with the non-scripted styles at around 8%. The frequencies of adverbs ranges from 9% in the scripted monologue to 20% in the non-scripted dialogue, while the non-scripted monologue and dialogue lies at 13% and 14% respectively.

The parts of speech contexts in the speaking styles introduce differences from those in the phrase frequencies. The parts of speech contexts where there are differences across speaking styles are at the conjunctions, the determiners, the interjections, the nouns and the pronouns. In the phrase frequencies, the Conjunctions and the Interjections were included in the category of O (words outside the scope of any phrase) and the determiners, the nouns and the pronouns were included in NPB (noun phrase beginning), two of the phrase categories with very high general frequencies in all four speaking styles, and the two with the highest frequencies as context after boundary annotations. Thus, the splitting up of these categories revealed differences between speaking styles.

In the case of noun phrases (NPB) all four speaking styles lie at around 30%. However, inspecting the frequencies for the nouns and for the pronouns reveals a difference. The scripted monologue has the highest proportion of nouns (29%) followed by the non-scripted monologue (17%), the non-scripted dialogue (14%) and the scripted dialogue (9%). The figures thus mirror the NVQ and NQ figures.

The proportion of conjunctions is greatest in the non-scripted monologue (14%), followed by the scripted monologue (9%) and scripted and non-scripted dialogue (both 7%). In the case of interjections the proportion is very small in the monologues (close to 0%), but greater in the dialogues (scripted dialogue 5%, non-scripted dialogue 6%). Thus, what seemed to be a similarity in category O turns out to be a difference between monologue and dialogue, with a greater proportion of conjunctions in the monologues and a greater proportion of interjections in the dialogues. This is related to the fact that both interjections/feedback and conjunctions are included in the category O. These aspects were not captured by the NQ and NVQ figures.

The frequency figures for the pronouns are reversed compared with those for the nouns. The scripted dialogue has the highest proportion of pronouns (22%), the non-scripted monologue and dialogue come next (15%), and the scripted monologue has the lowest proportion (9%). This shows that the similarities across speaking styles regarding noun phrases turns out to contain a pattern where the dialogues have a greater proportion of pronouns and a smaller proportion of nouns, while the opposite is the case for the monologues. These figures are in accordance with the NQ and NVQ figures.

The general part-of-speech distributions offered a more diverse picture of the linguistic properties in the four speaking styles, especially concerning the phrase categories NPB and O. Since these were the two categories which constituted the most frequent boundary annotation contexts in all four speaking styles, we can assume that there are also greater

differences in the boundary annotation contexts between speaking styles in the part-of-speech context than in the phrasal contexts.

5.4.2 The Part-of-Speech Context of the Boundary Annotations

Turning to the part-of-speech context of the boundary annotations, figures 5.10 and 5.11 show the part-of-speech context after majority boundary annotations for each speaking style, the monologues on the first page and the dialogues on the second one. Each diagram in the figures shows one speaking style, and both conditions Read and Listen together with the general part-of-speech frequencies. The diagram for the scripted monologue (top first page) contains three series. The first series, represented by black diamonds, shows the context of the boundary annotations in condition Read. The second series, represented by dark grey diamonds, shows the context of the boundary annotations in condition Listen. The third series, represented by light grey diamonds, shows as a reference point the general part-of-speech frequencies for the speaking style, i.e. a recapitulation of the information in figure 5.9. In this way, it becomes visible whether the part-of-speech context of the boundary annotation is proportional to the general part-of-speech distribution or whether it deviates from this.

We comment on the results from each speaking style, starting with the non-scripted monologue (top, first page).

In the scripted monologue, the part-of-speech context after the boundary annotations, is in most cases proportional to the general part-of-speech distribution. However, there are some clear variations. One of these is the proportionally high percentage of conjunctions as boundary annotation context. In condition Read 17% of the boundary annotations are followed by a conjunction, and in condition Listen 21%. The proportion of pronouns in the context after a boundary is also rather great, around 16% in both condition Read and condition Listen. The proportion of verbs is rather small, around 7% in both condition Read and Listen. Thus, in scripted monologue the most frequent contexts after a boundary annotation are conjunctions, nouns and pronouns. In addition, there is no greater difference between condition Read and Listen.

The pattern of the non-scripted monologue (bottom, first page) differs from the one of the scripted monologue. The most frequent part-of-speech context after a boundary annotation is conjunctions. In both conditions Read and Listen the frequency of conjunctions as context is about 40%. Next come pronouns at around 15% in both condition Read and Listen. The frequency of nouns as boundary context is very low compared with the general frequency of nouns in the speaking style, around 4% in both condition Read and condition Listen. The frequency of adverbs lies at around 11% in both condition Read and Listen. The greatest difference between condition Read and Listen is found in boundary annotations at hesitations: 8% in condition Read and 14% in condition Listen. In the non-scripted monologue, similarly to the scripted monologue, the proportion of

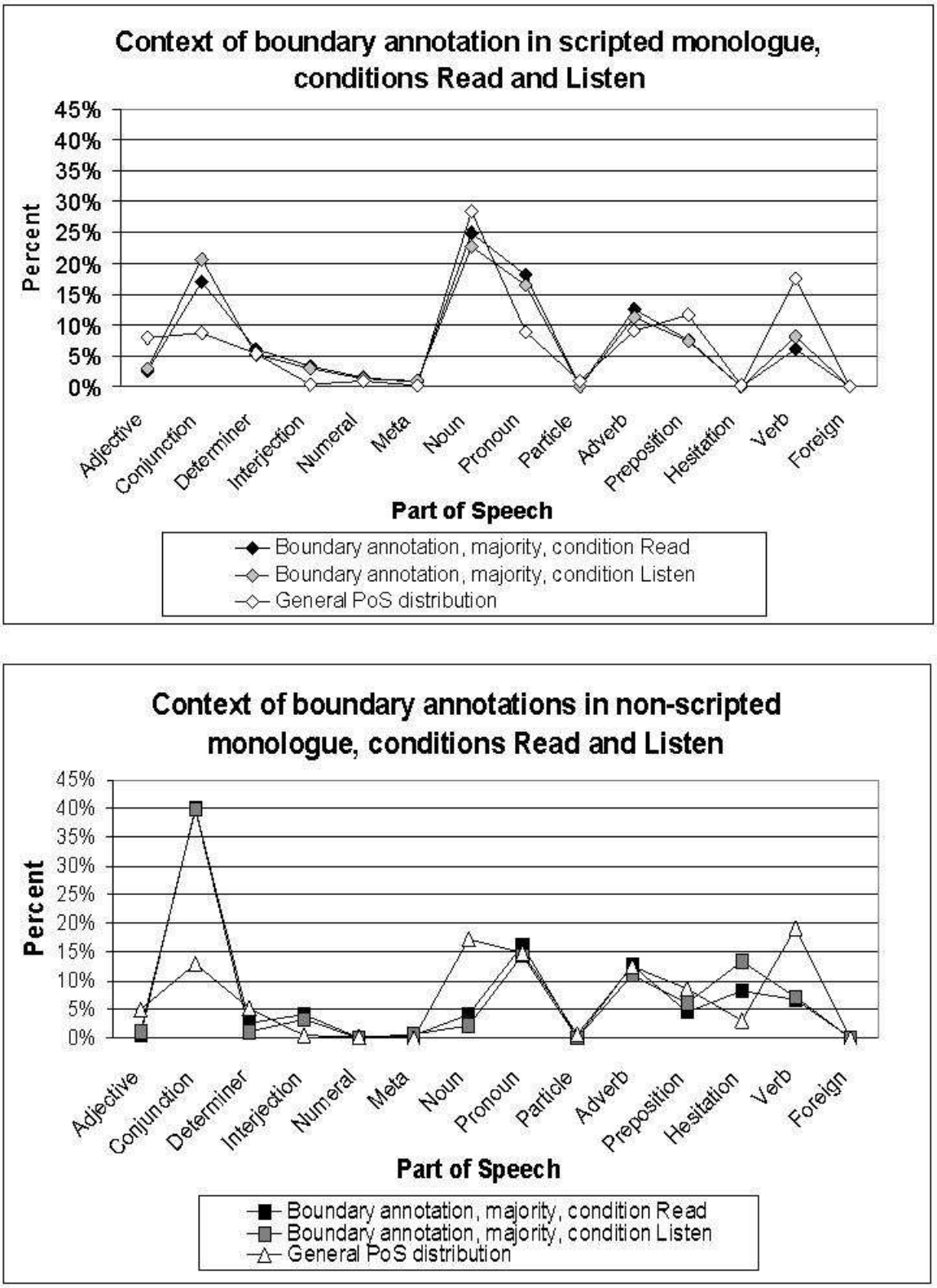

Figure 5.10: PoS context of boundary annotations in the monologues.

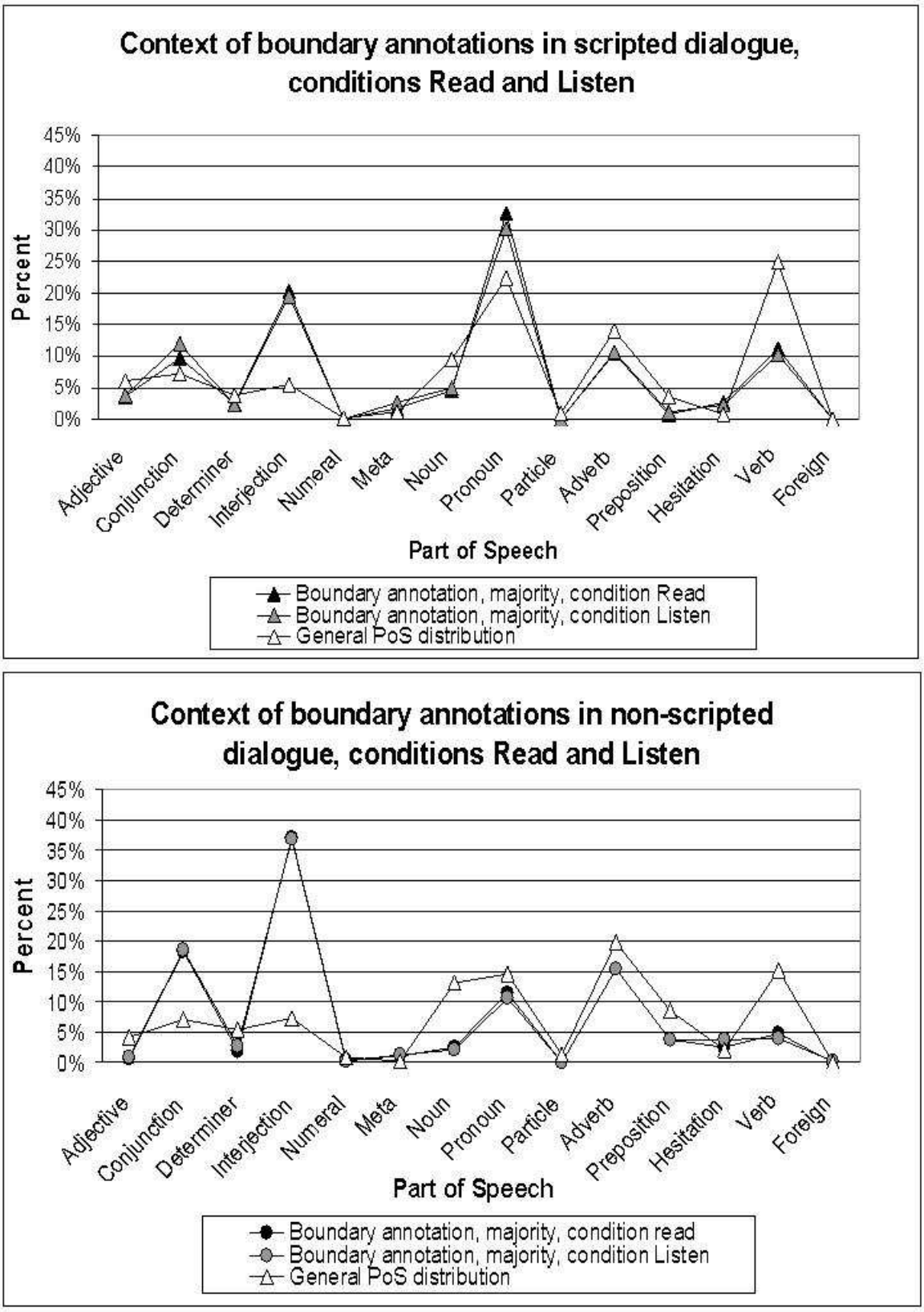

Figure 5.11: PoS context of boundary annotations in the dialogues.

verbs in the context after a boundary is smaller than the proportion of verbs in general in the speaking style, around 7% in both condition Read and condition Listen.

The scripted dialogue (top, second page) offers yet another pattern. The most frequent context after a boundary annotation is a pronoun in both condition Read and condition Listen (around 30%). In addition, the proportion of pronouns in the context after a boundary is greater than the proportion of pronouns in general in the speaking style. Also the interjections are more frequent as boundary context than in the speaking style in general (around 20% for both condition Read and condition Listen). Verbs, adverbs, and conjunctions all have a frequency of around 10% in both condition Read and condition Listen.

In the fourth speaking style, non-scripted dialogue (bottom, second page) the patterns differ again. Here the most frequent context after a boundary annotation is an interjection (feedback). The frequency of interjections as boundary context (around 36% in both condition Read and condition Listen) is considerably higher than the general frequency of interjections in the speaking style. Next come the conjunctions (18% in both condition Read and condition Listen), which also have a greater proportion in the context after boundaries than in the speaking style in general. Adverbs lie at 15%, nouns at 3% and verbs at 4%, all figures for both condition Read and Listen and all figures indicating a lower proportion as boundary context than in the speaking style in general.

An examination of the part-of-speech context of the majority boundary annotations yields a distinct pattern for each speaking style, but there are no great divergences between conditions Read and Listen. In the scripted monologue, the annotators have many boundary annotations at nouns and conjunctions, while the non-scripted monologues have many boundary annotations in front of conjunctions but also in front of pronouns. In the scripted dialogue the boundary annotations are located primarily in front of pronouns and interjections, and in the non-scripted dialogues they are to a great extent located in front of interjections and also to some extent in front of conjunctions and adverbs.

To sum up: boundary annotations in front of conjunctions are common in monologues, whereas in dialogues they are often located in front of feedback expressions (interjections). An examination of the part-of-speech contexts of the majority boundary annotations revealed a pattern which diverged across speaking styles. It is important to stress that these differences concern the majority boundary annotations, but not all kinds of boundary annotations in the data, i.e. the pattern concerns the boundaries on which the subjects agreed. We shall return to this issue in section 5.6.

5.5 Pause Context of the Boundary Annotations

The level of inter-annotator agreement was high between conditions Read and Listen concerning the majority of boundary annotations. It did not seem to be greatly influ-

	Scripted monologue	**Non-scripted monologue**	**Scripted dialogue**	**Non-scripted dialogue**
# of words	2858	2084	2212	2248
# of silent pauses	344	301	343	412
# of boundaries corresponding to a silent pause	277	128	277	225
# of majority boundaries	309	182	506	379
# of silent pauses without a corresponding boundary	67	174	66	187
# of boundaries without a corresponding silent pause	32	54	229	154
word/silent pause	8.31	6.92	6.45	5.46
word/boundary	9.25	11.45	4.37	5.94

Table 5.6: Overview of silent pauses and punctuation in each speaking style.

enced by the subjects having access to the prosody in condition Listen. This does not say anything about how the pause pattern differs across speaking styles. In this section we study how the majority boundary annotation in condition Listen corresponds to silent pauses longer than 50 ms. in the speech data. Since the speech signal obviously could not have influenced the annotations in condition Read, this condition is disregarded here. First an overview of the number of silent pauses and boundary annotations in absolute numbers is given, and thereafter a survey of three aspects of pause length: i) the mean pause length in the different speaking styles, ii) the mean pause length in the different speaking styles at boundary annotations made by the majority of subjects and iii) the mean pause length in the different speaking styles at silent pauses without boundary annotations.

Table 5.6 presents an overview of the number of silent pauses and boundary annotations. In addition, figures are shown for number of words, pause frequency and boundary annotation frequency (where the majority of subjects agreed on a boundary marking) in the different speaking styles.

Table 5.6 shows that there are differences between speaking styles concerning both pause frequency and how silent pauses and boundary annotations coincide. The pause frequency (row "word/silent pause") is highest (i.e. the pauses are occurring with the shortest interval) in the non-scripted dialogue, on average one silent pause per 5.46

words. The scripted dialogue and the non-scripted monologue have a lower pause frequency (6.45 and 6.92 respectively), and the scripted monologue has the lowest pause frequency, 8.31.

Comparing the frequency of the silent pauses (row "# of silent pauses") with the boundary annotation frequency (row "word/boundary"), there are fewer boundary annotations than silent pauses in all speaking styles except in the scripted dialogue. In the cases where there are more pauses than boundaries it is obvious that not all pauses can correspond to a boundary marking. However, how great is the overlap? We can express this relationship with the Precision and Recall. The Precision and Recall measurements give a picture of the relationship between the pauses and the boundaries. Precision will show the proportion of boundaries which corresponds to the silent pauses (see 5.1) and Recall will show the proportion of pauses corresponding to a boundary (see 5.2).

Thus, by Precision we mean the proportion of boundary annotations at silent pauses compared with the total of boundary annotations, as expressed in equation 5.2.

$$(5.1)\quad Precision = \frac{The\ number\ of\ boundary\ annotations\ at\ silent\ pauses}{The\ number\ of\ boundary\ annotations}$$

The second dimension, recall, is the proportion of boundary annotations at silent pauses compared with the total of silent pauses. The measurement is expressed in equation 5.2

$$(5.2)\quad Recall = \frac{The\ number\ of\ boundary\ annotations\ at\ silent\ pauses}{The\ number\ of\ silent\ pauses}$$

If the boundary annotations and the silent pauses are overlapping one-to-one, e.g if all boundaries are at silent pauses and all silent pauses are marked with a boundary, both precision and recall are high (100%), and in a diagram the plot is located to the top right corner. If many boundaries are not silent pauses but all silent pauses are boundaries, the precision (of boundaries at pauses) is low whereas the recall (of the pauses corresponding to boundaries) is high. This means that the plot would be located to the top left corner in a diagram.

Figure 5.12 shows the precision and recall for all four speaking styles, the precision on the Y-axis and the recall on the X-axis. The scripted monologue has a high precision (90%) together with a high recall (81%), while the non-scripted monologue has a lower precision (70%) together with a very low recall (43%). The scripted dialogue has a very low precision (55%) together with a fairly high recall (81%). The non-scripted dialogue also has a low precision (59%) and in addition a low recall (57%).

The Recall – Precision figures indicate that in the scripted monologue a high proportion of the boundary annotations are made at silent pauses (high precision) and many of the silent pauses correspond to a boundary annotation (high recall). In the case of non-scripted monologue, many of the boundary annotations are also silent pauses (relatively

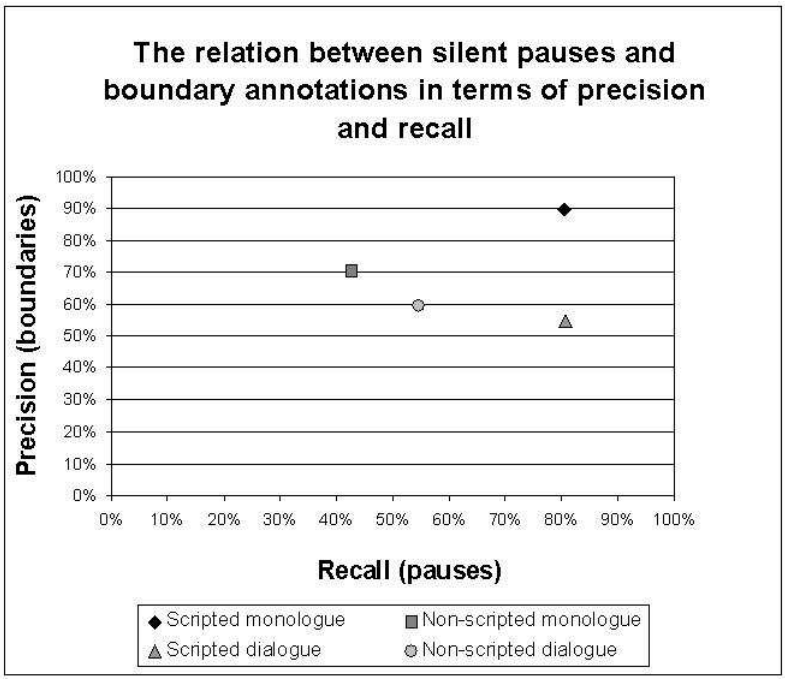

Figure 5.12: The distribution of pauses expressed with precision and recall.

high precision), but many of the pauses do not correlate with a boundary annotation (low recall). In the scripted dialogue few of the boundary annotations are made at a silent pause (low precision) but many of the silent pauses correspond to a boundary annotation (high recall). In non-scripted dialogue, few of the boundary annotations are made at silent pauses (low precision) and few of the silent pauses correlate with a boundary annotation (low recall).

In figure 5.12 we see that the scripted and the non-scripted speaking styles differ with regard to recall figures. Both scripted styles have a high recall, while the non-scripted styles have a low one. This means that there are fewer silent pauses which do not correspond to a boundary marking in the scripted styles than in the non-scripted ones, i.e. fewer "superfluous" pauses. Moreover, the monologues have a higher precision than the dialogues, the scripted styles differing more in the precision than the non-scripted ones. The higher precision indicates that a greater proportion of the boundaries are found at a silent pause in the monologues than in the dialogues. Thus, the higher recall (few "superfluous" pauses) seems to go with scriptedness and the higher precision (few "superfluous" boundaries) with monologues. In other words, there are many silent pauses that are also boundary annotations in the scripted styles, while we find many boundaries that are also silent pauses in the monologues.

	Scripted monologue	**Non-scripted monologue**	**Scripted dialogue**	**Non-scripted dialogue**
Pauses at boundaries				
# of pauses	277	128	277	225
Mean (sec.)	0.685	0.969	0.487	0.626
Maximum (sec.)	4.198	7.054	3.007	2.781
Minimum (sec.)	0.054	0.866	0.052	0.051
Pauses not at boundaries				
# of pauses	67	174	66	187
Mean (sec).	0.320	0.442	0.348	0.437
Maximum (sec).	2.817	1.677	1.240	1.505
Minimum (sec).	0.053	0.084	0.052	0.053

Table 5.7: Overview of general pause properties in the speaking styles.

The results might mirror the greater planning in the non-scripted speaking styles, resulting in more boundaries that not are annotated as discourse boundaries, as well as the interactive boundary cue of speaker change in the dialogues, resulting in boundary markings where no pauses are present. We come back to this issue in section 5.6.

5.5.1 Boundary Markings and Silent Pauses

Many researchers have pointed out the relationship between pause length and boundary strength, and in addition differences in the pause patterns across speaking styles have been reported. Since we use majority boundaries we could assume that the corresponding pauses should be rather long compared with the average, and we could also expect differences across speaking styles.

Before we examine the pauses at the boundaries in more detail, an account of the general pause properties in the speaking styles is given. The number of pauses in each speaking style, as well as the the mean pause length together with the range for pauses at boundary annotations and pauses not at boundary annotations, are shown in table 5.7

The overview in table 5.7 shows that the longest pauses are found in the monologues, whereas the dialogues have shorter pauses. In addition the pauses at boundaries are longer than the pauses not at boundaries.

In addition, an account of the the distribution of pause durations in each speaking style is given. In figure 5.13 an overview of pause lengths in different categories is presented.

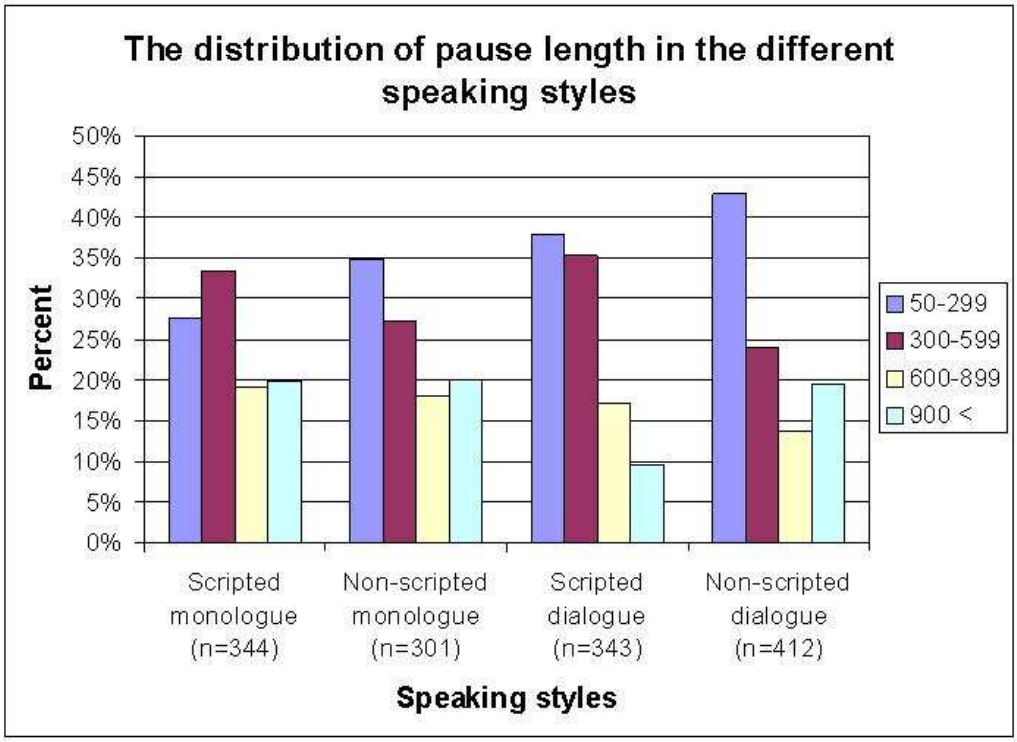

Figure 5.13: The mean pause length in the speaking styles.

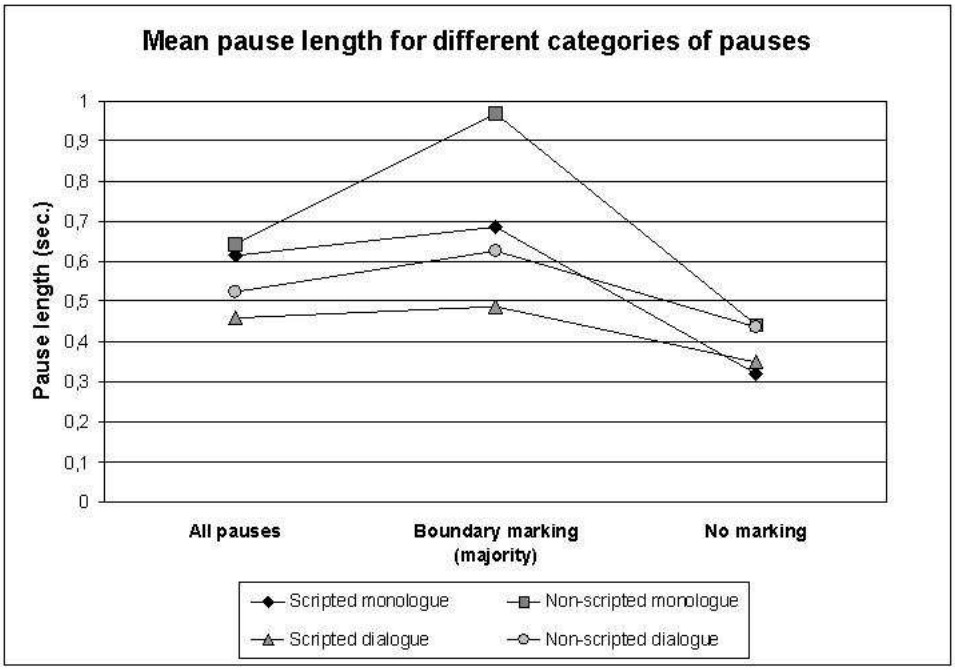

Figure 5.14: The mean pause length in the speaking styles.

The overview of the pausing features is in general in line with what previous research has reported. Figure 5.13 shows that the scripted monologue in general has fewer and shorter pauses. In the non-scripted monologue the pauses are both longer and more numerous. In the case of the dialogue it is important to stress that the scripted dialogue consists of an excerpt from a radio play. This implies that the speech is probably edited and that the pauses are influenced. This might be one reason why the pause durations are very short in the scripted dialogue. In the non-scripted dialogue the pauses are more frequent, but a large proportion consists of rather short pauses.

We now turn to a closer examination of the relationship between pause duration and boundary markings in the different speaking styles shown in figure 5.14. From left to right figure 5.14 shows the mean pause length for all silent pauses ("All pauses"), the mean pause lenght for all pauses at a boundary annotation made by the majority of subjects ("Boundary marking, majority"), and the mean pause length for silent pauses at positions not correlating to a boundary marking ("No marking").

Inspecting the figure we find that the mean pause length in most cases is shorter in the dialogues than in the monologues, and the scripted styles have shorter pauses than the non-scripted ones. In the case of the general pause length ("All pauses") the differences between the speaking styles is rather slight. The shortest mean is found in the scripted dialogue (0.46 sec.). The mean of the silent pauses in the non-scripted dialogue is a little bit higher (0.52 sec.), while there is an even longer mean pause length for both scripted and non-scripted monologue with 0.61 and 0.62 sec. respectively.

Pauses at the boundary markings are in general longer than pauses not coinciding with a boundary marking. Moreover, the differences between the speaking styles are clearer but their internal order is retained (e.g. non-scripted monologue longest duration). The shortest mean duration (0.49 sec.) is found in the scripted dialogue, whereas the non-scripted dialogue has a longer one (0.61 sec.). The scripted monologue comes next with an average pause duration of 0.69 sec., and the non-scripted monologue with a mean pause duration of 0.97 sec comes last.

The pauses are on average shorter when the pause is not corresponding to a boundary marking, and in addition there are smaller differences between the speaking styles. We find the shortest average duration in the scripted monologue (0.31 sec.), closely followed by the scripted dialogue (0.32 sec.). The non-scripted monologue and the non-scripted dialogue both have an average pause duration of 0.44. This means that here we find a difference in average pause duration between the scripted and the non-scripted speaking styles but no difference, or only a very slight one, between the monologue and the dialogue.

The difference between the pause duration at boundaries and the pause duration in pauses not corresponding to a boundary is significant in each speaking style on at least the level of $p<0.05$ using a t-test. Grouping the spaking styles, the monologues have significantly longer pause durations than the dialogues both at boundaries and at pauses not correlating to a boundary on at least the level of $p<0.01$).

Comparing the curves for the four speaking styles in figure 5.14, we find specific patterns for the monologues and the dialogues. In the monologues there is a greater difference between the pauses correlating to a boundary marking and the pauses not correlating to a boundary marking than is the case in the dialogue. In other words, the difference in pause length between "boundary pauses" and "non-boundary pauses" is greater in the monologues than in the dialogues.

Thus, in our data the pauses which are correlating to boundary markings are on average longer than pauses not correlating to a boundary marking, and in addition there are specific monologue and dialogue patterns in the difference between boundary pauses and non-boundary pauses.

5.5.2 Part-of-Speech Context of the Boundaries at Silent Pauses

Many researchers have pointed out that discourse boundaries are signalled by a combination of features, for instance a combination of a pause and a lexical discourse marker. The results from the study of the part-of-speech context of the majority boundaries indicated different patterns across speaking styles. Thus, certain parts of speech were more common as boundary context than other ones. Are then certain parts of speech more frequent as boundary markers together with a pause than "on their own"? In addition, if there are differences, do they vary across speaking styles? In order to study this question we have extracted the part-of-speech context of all majority boundary markings and categorized them according whether the boundary marking is correlating to a pause or not. The results are presented in figures 5.15 and 5.16. We start with an examination of the context of the boundary markings at silent pauses, figure 5.15.

Figure 5.15 presents the part-of-speech context after boundary markings at silent pauses in all four speaking styles. The patterns for the speaking styles are similar to the patterns of the part-of-speech context for the boundary markings given in 5.10 and 5.11. The scripted monologue has peaks for nouns (23%), conjunctions (20%) and pronouns (17%), while the non-scripted monologue has peaks for conjunctions (39%), hesitations (19%) and pronouns (14%). In the scripted dialogue the peaks are located to pronouns (26%) and interjections (24%), whereas the peaks in the non-scripted dialogue are found at interjections (46%) and conjunctions and adverbs (both 15%). Thus, the patterns from the part-of-speech contexts of the boundaries are in general preserved.

Figure 5.16 shows the part-of-speech context of boundary annotations not coinciding with a silent puse. Comparing the curves in 5.15 with the curves in 5.16 we find that there are some minor deviations. In the scripted monologue there is a lower proportion of pronouns (6% as compared with 17%) and a higher proportion of verbs (25% as compared with 6%). In the non-scripted monologue there is a greater proportion of adverbs (20% as compared with 7%) and no boundaries at hesitations (as compared with 19%). Please note, however, the low number of boundaries not corresponding to a silent pause in the monologues as compared with the dialogues.

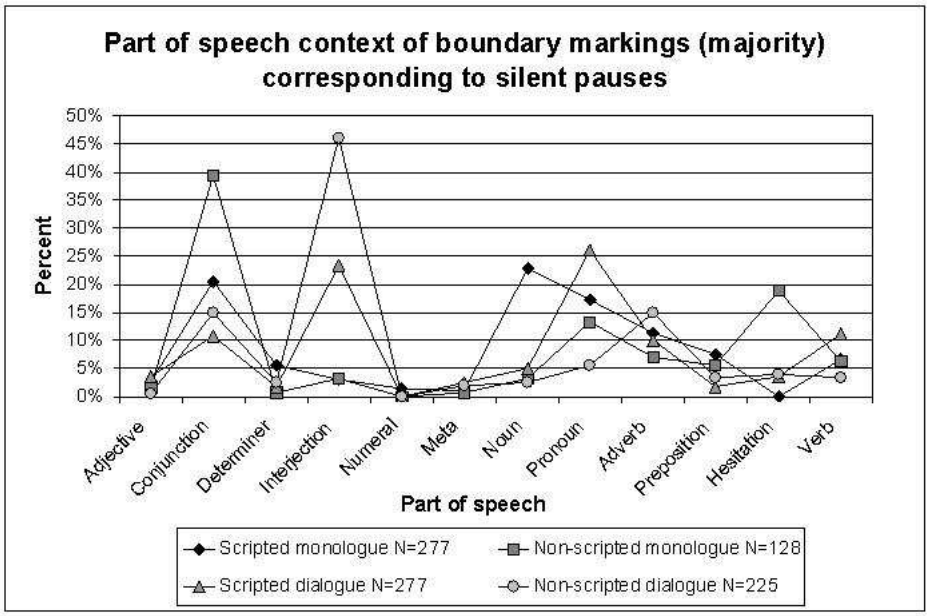

Figure 5.15: The part-of-speech context of boundary markings at pauses.

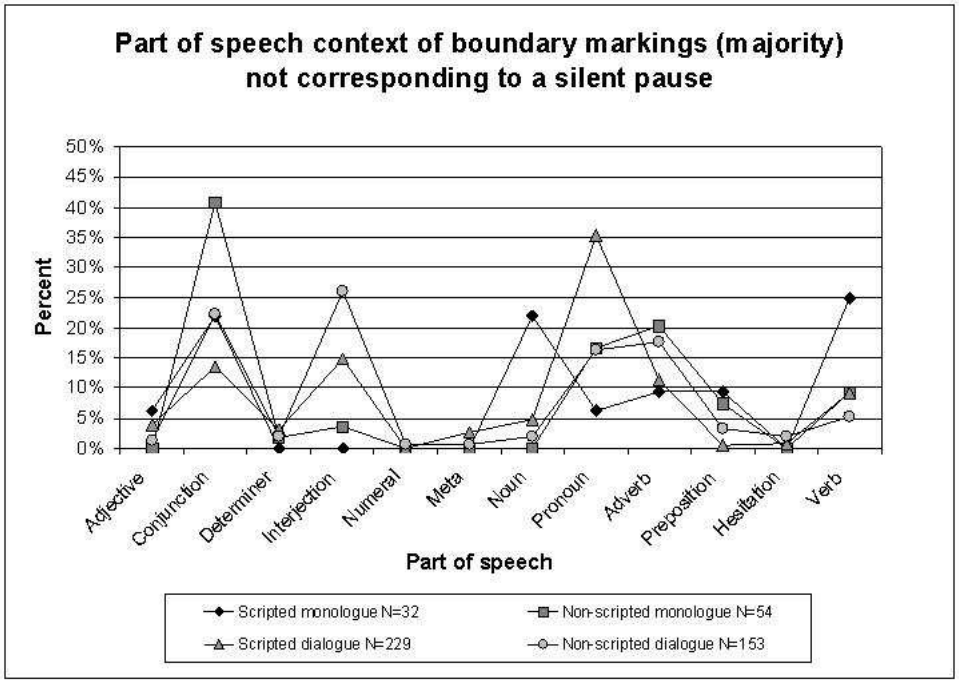

Figure 5.16: The part-of-speech context of boundary markings not at pauses.

In the scripted dialogue the same pattern for boundaries not corresponding to silent pauses as the one for boundaries corresponding to silent pauses is present. However, there is a higher number of pronouns (35% as compared with 25%). Moreover, the proportion of conjunctions and interjections is for both around 15%, meaning that the proportion of interjections has decreased while the proportion of conjunction has increased. In the non-scripted dialogue the dominance of the interjections has descreased (from 46% to 26%), while the proportion of pronouns has risen (from 5% to 16%).

The boundaries corresponding to pauses and the boundaries not corresponding to pauses in general follow the patterns for the four speaking styles. This indicates that the silent pauses support the part-of-speech preferences in the specific speaking styles, but they do not seem to make up such a strong boundary cue that the general context pattern in the speaking styles is ignored. This view is supported both by other studies and by our own inter-annotator agreement figures which indicated a high level of inter-annotator agreement across conditions Read and Listen.

However, there are some minor variations in the data. Subjects seem to be more likely to mark a boundary at certain combinations of part-of-speech and pause than at other combinations. For example, in the non-scripted dialogue there is a smaller proportion of boundaries in front of interjections/feedback in the context without a silent pause (26%) than in the context together with a silent pause (46%). A similar pattern is found in the scripted dialogue. In addition, there is a greater proportion of conjunctions in all four speaking styles in the context without a pause than in conjunction to a pause; the pattern is also similar for pronouns. These results might be related to where in a clause a certain part of speech is more frequent.

5.6 Discussion of the Boundary annotation Task

In the study of boundary annotations we have addressed two main questions: 1) Can we find differences in the subjects' boundary annotations of the speaking styles within conditions Read and Listen? If they do differ, what are the differences with regard to the linguistic and prosodic features we have studied? and 2) Do subjects' boundary annotations differ across conditions Read and Listen?

Firstly we saw that the level of inter-annotator agreement varies across the speaking styles within conditions, the scripted styles having a higher level of inter-annotator agreement than the non-scripted styles in both conditions Read and Listen. Secondly, studying the results from the inter-annotator agreement across conditions Read and Listen we noted that there were no great differences. The positions where the majority of subjects agree on a boundary are to a great extent the same in the both conditions. However, because of the filtering of the data a portion of the boundaries with a lower level of agreement were disregarded. Thus, we do not know how similar or different from one another these positions were across conditions.

In this section we discuss some of the features from the boundary study, specifically:

- The level of inter-annotator agreement
- The differences across speaking styles regarding the linguistic features
- The differences across speaking styles regarding pausing

The level of inter-annotator agreement differed across the speaking styles within conditions. In both conditions there was a higher level of agreement in the scripted styles than in the non-scripted styles. This indicates that the subjects had more problems in the annotation of the structure in the non-scripted styles. To some extent this might depend on the annotation task: The task was to insert punctuation, and it is plausible to assume that this is more difficult in transcripts of elicited spontaneous speech than in speech based on a written manuscript. However, this is not a problem in this study since our aim is to study discourses that are structurally different, and if this difference in the degree of applicability of punctuation indicates structural differences in the discourses, this is just supporting the view that our data are sufficiently different. The high level of agreement in the scripted styles would then be a consequence of the task, but the level of inter-annotator agreement would still indicate differences between speaking styles.

Another aspect of boundary annotation in relation to the subjects' agreement is the notion of *strong boundaries* used by Swerts (1997). A strong boundary is for instance a paragraph boundary, whereas a weaker boundary can be a phrase boundary. Swerts argues that a higher level of agreement between annotators indicates a strong boundary, whereas a lower level of agreement indicates a weak one. If we interpret "stronger" as also indicating "clearer", then we could conclude from our figures that the scripted speaking styles in general have clearer boundaries than the non-scripted ones. This implies that a paragraph conforming to written discourse could be assumed to have clearer boundaries than the corresponding length of spontaneous discourse.

Concerning the level of inter-annotator agreement within conditions we conclude that the figures support structural differences in our data, thus indicating clear boundaries in the scripted styles and less clear ones in the non-scripted styles. We leave the topic for now but shall return to it in chapter 8.

The inter-annotator agreement across conditions indicates a high level of agreement between the majority annotations in condition Read and the majority annotations in condition Listen. With regard to this it does not appear that the access to the speech signal influenced the annotations to any great extent. In the few cases where there is a genuine disagreement across conditions (e.g. where a majority agree on a boundary in condition Read, but where no subject in condition Listen has annotated a boundary) there are some clear indications:

- When subjects agree on a boundary in condition Read, but ignore this position in condition Listen, the position is often syntactically a sentence boundary, but

prosodically the F0 very clearly signals coherence and continuation. There are also cases in the dialogues where a speaker change is indicated as a boundary in condition Read but ignored in condition Listen.

- When the subjects agree on a boundary in condition Listen but not in condition Read, the position is often syntactically a clause boundary accompanied by a clear silent pause.

In most of these cases of clear disagreement there is a position where the subjects in condition Listen have annotated a boundary, whereas the subjects in condition Read did not. These clear cases (majority – 0) are very few, in total there are 22 cases where a boundary marking is present in condition Listen but not in condition Read. Out of 20, 18 have a pause which on average is 0.675 seconds. The two remaining cases have clear F0 resets.

There where also cases where the majority of subjects in condition Read but not in condition Listen annotated a boundary. One example from the scripted monologue (news broadcast) is shown in 5.3 and 5.17. The form in 5.3 is the same as in the subjects' transcriptions. The example contains an introductory description followed by a register of names, i.e. a kind of list structure.

(5.3) i ekostudion marianne hasslow och li hellström

English translation:
in the echo studio marianne hasslow and li hellström

Figure 5.17 shows the spectrogram with the F0 for the utterance. Under the spectrogram an aligned transcription of the utterance is given, where (R) indicates a boundary annotated in condition Read, while (R,L) indicates a boundary in both condition read and condition Listen. Thus, the subjects have inserted a boundary between "studio" and "marianne" in condition Read and after "hellström" in both conditions Read and Listen. The time axis is shown in seconds, and at the bottom of the figure the utterance is transcripted with Swedish orthography.

In the example in figure 5.17 the acoustics, both the absense of pause and the absense of F0 fluctuation, argue against a boundary. This indicates that in this case subjects in condition Read annotated on the basis of the text structure where a boundary is plausible, while the subjects in condition Listen seem to have annotated on the basis of the speech signal which does not support a boundary marking. Thus, there are cases in our data where the subjects seeem to be influenced by the prosody, these cases are, however, rather few.

In addition to differences across speaking styles indicated by inter-annotator agreement, there are differences on the level of both phrases and parts-of-speech. Starting with

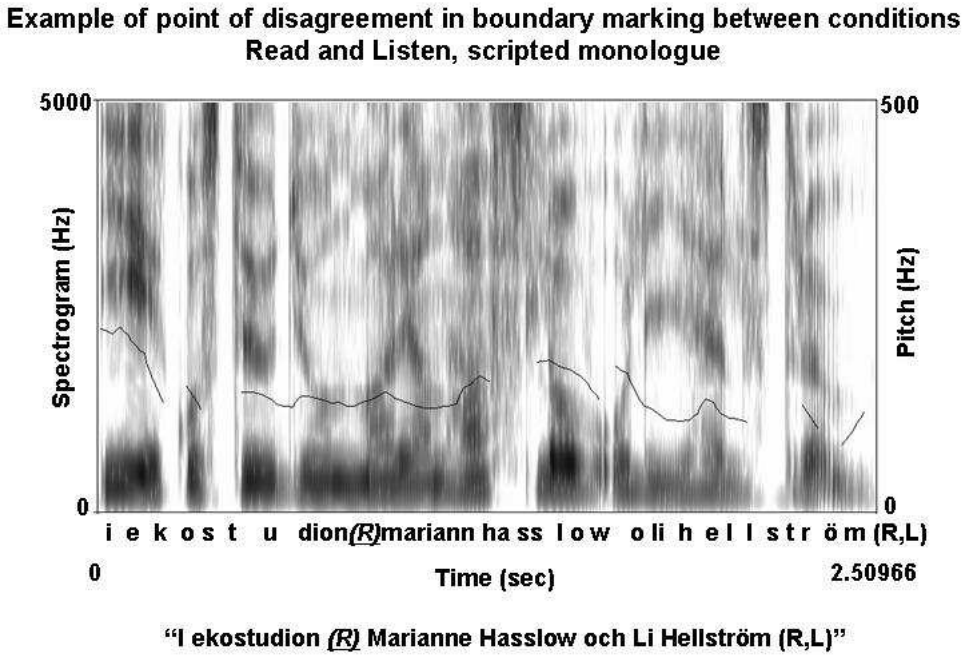

Figure 5.17: Example of disagreement in the boundary annotation.

the phrasal analysis there were indications of differences in phrase depth between the speaking styles. In addition a certain set of phrases was more frequent in the context after a boundary annotation. However, the differences between speaking styles were not very clear.

The differences across speaking styles become clearer when examining the part-of-speech context of the majority boundary annotations in the different speaking styles. Each speaking style shows a rather specific profile of the part-of-speech frequencies in the context after a boundary annotation. Relating this to the idea of markers Green (1979), we could describe the speaking styles as being characterized by different sets of markers corresponding to strong boundaries. Moreover, because of the high level of agreement across conditions Read and Listen, these markers seem to do relatively more work than the prosody in the signalling of strong discourse boundaries.

There is more to be said about the differences across the speaking styles concerning the distribution of the linguistic features. We leave this topic for now but shall return to it in chapter 8.

Studying the acoustic context of the boundary annotations we find that there are rather specific differences in the overlap between boundary annotations and silent pauses in the four speaking styles. Generally, the scripted speaking styles have a greater number of

the pause positions marked as boundaries (higher recall), while dialogues have a greater number of boundaries not corresponding to a pause (lower precision), (see figure 5.12). The lower recall in the non-scripted speaking styles, meaning that there are many pauses which do not correspond to a boundary annotation, can be assumed to be connected to the presence of planning pauses in non-scripted speech.

Why, then, do dialogues have generally lower precision, i.e. more boundary annotations not coinciding with a silent pause? This should not be connected to planning pauses, since the tendency is the same in both scripted and non-scripted dialogue. Instead, it might be connected to a preference by the subjects to mark a boundary on the basis of the graphic transcription. Each speaker contribution is transcribed on a new line, and the subjects in condition Read might have preferred to annotate boundaries at the new line/speaker change positions. If we have many such speaker change positions without a pause, the result would be many boundary markings without a corresponding pause – as we find in the dialogues.

The overlap between boundary marking and speaker contribution was studied and found to be higher in condition Read. Thus, this indicates that subjects in condition Read were influenced by the graphic layout in their marking, annotating a boundary at speaker change. However, with access to the speech signal many shorter contributions, e.g. feedback expressions, were not regarded as independent clauses, so they were just included in the other speaker's current utterance. Thus, the access to the speech signal seems to have influenced the annotators towards an annotation less limited to the layout. However, without this clear segmenting limitation, the speaker change as boundary signal became less clear and thus introduced more disagreement which resulted in a lower κ value. Such a tendency by subjects to segment at speaker change and feedback expressions is also reported by Arim *et al.* (2003).

There were also other indications that the subjects in condition Read were influenced by the layout in the annotation, but we shall come back to this issue in chapter 8.

The study of boundary annotations indicates different types of boundary contexts in the different speaking styles for the majority of boundary annotations. There are differences in the level of inter-annotator agreement between speaking styles, indicating that the boundaries were less clearly signalled in the non-scripted speaking styles. The high level of agreement across conditions could be interpreted as indicating that the lexical boundary cues are rather strong compared with the prosodic.

The part-of-speech analysis revealed that the boundary annotations were found in different contexts in the different speaking styles. The pause pattern was also different across speaking styles with longer pauses in the monologues than in dialogues. In addition a greater proportion of the pauses were boundary markings in the scripted speaking styles, whereas a greater proportion of boundaries corresponded to pauses in the monologues.

Since the point of this study is the relationship between boundaries and prominence, we shall discuss this in more detail in the general discussion in chapter 8. The issue of the

relationship to discourse theory is also left for the general discussion, since it depends on the results from the two other studies. Instead we proceed to the next chapter and the prominence annotation study.

Chapter 6

Discourse Prominence

IN the previous chapter, the results from the boundary annotation study were examined. We now turn to another aspect of the segmenting of the discourse: the study of the prominent parts. In this study we account for a number of features of words which the subjects annotated as prominent, and how these features might vary across different speaking styles. We also study how the subjects annotations are influenced by their having access to the speech signal compared with annotations based on the transcripts alone.

6.1 Introduction to the Prominence Annotation Task

In this study, as well as in the study of boundaries, our central hypothesis is that the prosodic features of a discourse are related to the structural properties of the same discourse. We see discourse prominence as one of the two aspects of a discourse segment which together with the boundaries form a whole unit. Furthermore, we believe that the acoustic and lexicogrammatical properties of a discourse influence the composition of the segment. In the previous study we investigated the aspect of the boundaries across speaking styles, and we now turn to the aspect of prominence across speaking styles. Our hypothesis is that the aspect of prominence could be related to the attentional state in the discourse theory of Grosz and Sidner (1986).

Some researchers have studied the aspect of prominence across different speaking styles. However, in general the focus has been not on discourse structure, but rather on acoustic correlates to focus, e.g. by Heldner (2001), or the relationship between focus, acoustic correlates to focus and the information structure, see e.g. Swerts (1997) or van Donzel (1999).

In this thesis we study the relationship between a number of lexicogrammatical and prosodic features and the subjects' annotations of prominence. As lexiogrammatical

features we study the same features as in the study of boundaries, i.e. the phrasal properties based on a parse, and the part of speech properties based on tagging. Thus, we study the phrasal and part of speech properties of the words annotated as prominent in the same way as described for boundaries in chapter 5.

The prosodic correlate to prominence is the focal accent as described by Bruce (1998:107). Important acoustic correlates to accent and focal accent in Swedish pointed out by many researchers are F0 pattern, longer segment duration and higher intensity, see e.g. Gårding (1967), Strangert and Heldner (1995) and Bruce (1998). The focal accent pattern together with measurements of F0, duration and intensity are the selected acoustic correlates which are examined in this study, and we examine each of these features for the words annotated as prominent. For a closer description of the measurements see chapter 4.

In the study of prominence, as in the study of boundaries, two central issues are addressed: 1) Do the characteristics of prominence annotations differ between the different speaking styles within conditions Read and Listen? and 2) Do the characteristics of prominence annotations differ between the same speaking styles across conditions Read and Listen? In other words, will the patterns of the prominence annotations differ between the four speaking styles, and will the subjects' annotations of prominence be influenced by their having access to the speech signal?

The data in the study of prominence consists of the same samples of scripted and non-scripted monologue and dialogue as were previously used in the boundary annotation study. The mark-up is also the same except that the starting point for this study is the subjects' annotations of prominence.

In chapter 3 the design of the experiment as well as the prominence annotation task are described in detail. A very brief description is repeated here. The subjects' task was to underline the words, phrases or other units they found prominent. One group of subjects did this in transcripts of the data (condition Read) and another group did this in transcripts of the data but they also had access to the speech signal (condition Listen). The experiment instructions gave the subjects a sample annotation from non-scripted dialogue, so as to provide some guide concerning the granularity of the prominence marking. Part of the instruction is found in 6.1, the words marked as prominent are underlined and an English word for word translation as well as a more idiomatic paraphrase is given below. The full instructions are found in Appendix 1.

(6.1) <Talare A> då ska vi se då har vi en en s karta här framför oss och jag har landstigit på en plats på den här ön och det börjar vid en en in i en bukt en ganska ovalt formad bukt inne på västra sydvästra sidan utav den här ön har du den också
<Talare B> jo då och tydligen så är ju formen på våra öar identiska

English transliteration:
<Speaker A> then shall we see then have we a a s map here in front of us and I have landed at a place on this here island and it begins by a a into in a bay a

rather ovally shaped bay in on the west-THE southwest side of this here island have you it too?
<Speaker B> yes then and evidently so is shape-THE on our islands identical

English translation
<Speaker A> then let us see then we have a a s map here in front of us and I have landed at a place on this island and it begins by a a into a bay a rather ovally shaped bay on the west southwest side of this island do you haveit too?
<Speaker B> oh yes so evidently the shape of our islands is identical

The structure of the chapter is as follows: in section 6.2, a survey of the annotation profiles in the prominence marking task is given, followed by some general remarks on the prominence annotation task, including inter-annotator agreement statistics (section 6.3). In relation to this statistical information we want to stress that in the same way as in the study of boundaries we use two different sets of data: 1) all positions where at least 1 subject annotated a word as prominent (prominence sites) and 2) the positions where the majority of subjects have annotated a word as prominent (prominence annotations). Set 1 is used in the annotation profiles as well as in the kappa measurements within conditions, but not in the kappa measurements across conditions and not in the following examination of linguistic and prosodic features.

In section 6.4.2 the phrase and part-of-speech context is examined. Many researchers have reported that prominent information comes at the end of a sentence. One of the questions in this section is: will the words annotated as prominent be located more to embedded contexts than boundaries? In addition we examine whether the phrasal and part of speech properties differ across speaking styles for prominent words.

The acoustic features of the words annotated as prominent are examined in section 6.5. In this section we focus on a description of the prosodic and acoustic features and measure the words annotated as prominent in the well-known dimensions previously mentioned. We do not aim to investigate new correlates, but to describe some already known correlates and compare them across speaking styles and then relate them to the features of the boundaries.

In section 6.6 we study the relationship between the prominence annotations and boundary annotations. Also in this section we explore the idea that new information, which is often accented, comes at the end of a sentence.

The chapter ends with a discussion in section 6.7.

6.2 Annotation Profiles for the Prominence Annotation Task

The aim is to show on one hand how individual subjects on average annotated prominence and on the other hand how specific transcripts were annotated. Annotation profiles per subject and per transcript are presented in this section. The corresponding annotation profiles for the boundary annotation task were presented in chapter 4, section 4.3.

A profile for a *subject* consists of the subject's global mean marking frequency, computed on the mean marking frequency for each transcript annotated by the subject. A profile like this shows us whether any subject on average differs from the other subjects' average annotation frequency. In the prominence annotation task there are 10 subjects in each of the conditions Read and Listen.

A profile for a *transcript* consists of the mean of all subjects' mean marking frequencies computed for one transcript. This profile shows us whether the subjects' annotation frequency on average deviates in a specific transcript.

Before studying the annotation profiles, let us have a brief review of some statistics covering the words annotated as prominent in the two conditions Read and Listen. The figures are shown in table 6.1.

	Scripted monologue	Non-scripted monologue	Scripted dialogue	Non-scripted dialogue
# of words	2858	2084	2212	2248
# of prominence sites, Read	1690	1386	1276	1104
# of prominent words, Read	384	191	384	299
# of words/prom. word, Read	7.44	10.91	5.76	7.52
# of prominence sites, Listen	1701	1489	1169	1094
# of prominent words, Listen	380	197	320	240
# of words/prom. word, Listen	7.52	10.58	6.91	9.37

Table 6.1: Number of words and prominence markings (majority) in the speaking styles.

Table 6.1 shows differences between the speaking styles regarding the number of words annotated by the subjects as prominent in the different speaking styles. In general fewer prominence annotations are found in the non-scripted speaking styles. It is also clear that the number of positions annotated by a minority of the subjects (# of prominence sites) is much higher than the number of positions annotated by the majority of the subjects (# of prominent words).

The rows "# of words/prom. word" shows how often, on average, the majority of subjects have annotated a word as prominent in each speaking style in condition Read and

condition Listen. For example, in scripted monologue on average about every 7th word is annotated as prominent. However, we saw in the annotation profiles that this figure can differ between specific transcripts.

In addition to the above statistics we recapitulate the average articulation rate for the speaking styles from chapter 5 in table 6.2. In addition the average number of characters per word for all the words as well as the number of characters per word annotated as prominent is shown.

	Scripted monologue	Non-scripted monologue	Scripted dialogue	Non-scripted dialogue
ms./word	280	317	341	306
char./word	5.23	4.43	4.12	4.20
char./ prom. word	6.80	6.77	5.47	6.06

Table 6.2: Articulation rate in the four speaking styles.

We briefly remind the reader that the speech durations in table 5.2 are computed on the total speaking time in the sound files (minus total duration of silent pauses) divided by the total number of words. The average number of characters per words indicates that on average the words are longer in the scripted monologues than in the three other speaking styles. A comparison of the average number of characters per prominent word with the average number of characters per word indicates that the words annotated as prominent are on average longer than the other words. In addition the words annotated as prominent are longer in the monologues than in the dialogues.

6.2.1 Profiles for the Prominence Annotation Task, Condition Read

In this section we give a general picture of the prominence annotations by means of the annotation profiles. The corresponding section for the boundary annotation task is found in chapter 4, section 4.3.1. As in the annotation profiles for the boundaries, we account for the distribution of annotations by subject and by transcript in both conditions Read and Listen.

Starting with condition Read, figure 6.1 shows the annotation profile for each *subject* (1-10) in condition Read. Horizontally we find the subjects (1-10), and vertically the number of words per prominence annotation. Please note the difference in scale on the y-axis between the boundary annotation profiles in chapter 4 (scale 0-16) and the prominence annotation profiles presented here (scale 0-35). The mean annotation frequency for each subject is indicated with a black square, e.g. for subject 10 we find the black square at 5, which means that for all speaking styles subject 10 annotated on average

every 5th word as prominent. The bars show the range of variation of the annotation frequency across the transcripts annotated by the subject, i.e. the minimum and maximum annotation frequency for the 11 transcripts which the subject annotated. Subject 10 had a range of variation between 3.5 and 8, meaning that in some transcripts on average every 3.5th word was annotated as prominent (highest frequency), while in another transcript on average every 8th word was annotated as prominent (lowest frequency), and the annotation frequency for the rest of the transcripts lies within these limits. Note that a low number of words (few words per annotation) results in a high annotation frequency (the prominence annotations were made more often), and vice versa. For a closer description of how the measurementss were computed, see chapter 4, section 4.3.

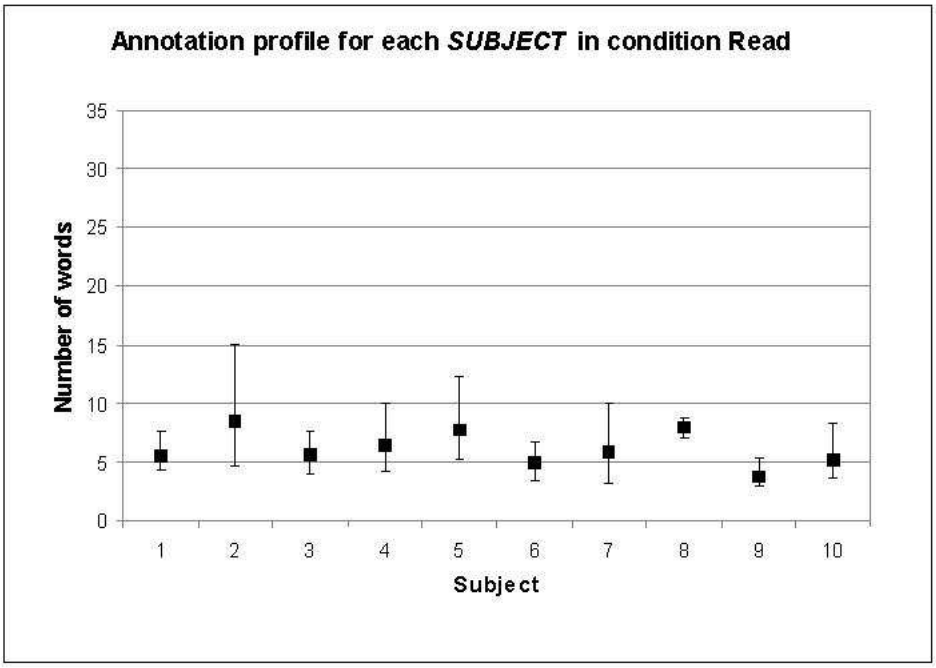

Figure 6.1: Annotation profiles for the subjects in the prominence marking task, condition Read.

Figure 6.1 shows that between the subjects the mean annotation frequencies vary from 4 to 8. In the profile of the boundary annotation task (see 4, section 4.3.1) we find a boundary annotation on average after every 5th to 8th word, so on average the boundary annotation frequency and the prominence annotation frequency per subject are made with about the same intervals.

In a similar way to the boundary annotation profile, there is a considerable range of variation between the subjects concerning the average prominence annotation frequency. For instance, subject 2 in 6.1 has a range between 4.5 and 15, meaning that in one transcript subject 2 has annotated every 4.5th word as prominent, while in another transcript every 15th word was annotated as prominent. The minimum value is 3.5 (subject 7) and the maximum 15 (subject 2). This could be compared with the minimum

and maximum values of the range in the boundary annotation profile which were 3.5 and 14 respectively.

There may be different reasons for the wide range of variation for the individual subjects. For instance, if the subjects on purpose annotate the speaking styles with different frequency, the individual variation would then show up as a clear difference in the average annotation frequency examined per transcript. This was also the case in the boundary annotation task, where a difference between the monologues and the dialogues was found; subjects marked a boundary in the dialogues more often than in the monologues. In addition two specific transcripts in non-scripted monologue, speaker A and speaker C, in general had a relatively wide range of variation, meaning that different subjects had very different annotation frequencies in the transcripts. Do we find a similar pattern in the prominence annotation?

Figure 6.2 presents the annotation profiles per *transcript* in condition Read. The transcripts are shown horizontally, and the number of words again vertically. The mean prominence annotation frequency in the transcripts lies between 5 and 7.5, and the range of variation is between 3.5 (transcript 7) and 15 (transcript 10).

In the boundary annotation task differences in the subjects' individual range of variation correlated with speaking styles, e.g. a lower frequency in the monologues compared with the dialogues. In the prominence annotation the variation is less clearly correlated with specific styles. The values are slightly higher for the non-scripted monologue (speaker A, B, and C), indicating that subjects in general might have annotated prominence with slightly longer intervals in the non-scripted monologue than in the other speaking styles. The range of variation is wide, especially for speaker A in the non-scripted monologue and in the non-scripted dialogues 2, 3 and 4, indicating a clearer disagreement between the subjects on how to annotate these transcripts compared with the other ones.

The annotation profiles per subject in the prominence marking task, condition Read, show a mean annotation frequency and a range of variation rather similar to the one we found in the boundary annotation task. In that task, however, the subjects' variation seemed to be distributed in a specific pattern over the transcripts, whereas this was not the case in the prominence annotation task.

How are the prominence annotations influenced by subjects having access to the speech signal? Let us proceed with the annotation profiles for the prominence marking task in condition Listen.

6.2.2 Profiles for the Prominence Annotation Task, Condition Listen

Corresponding annotation profiles for the boundary annotation task are found in chapter 4, section 4.3.2.

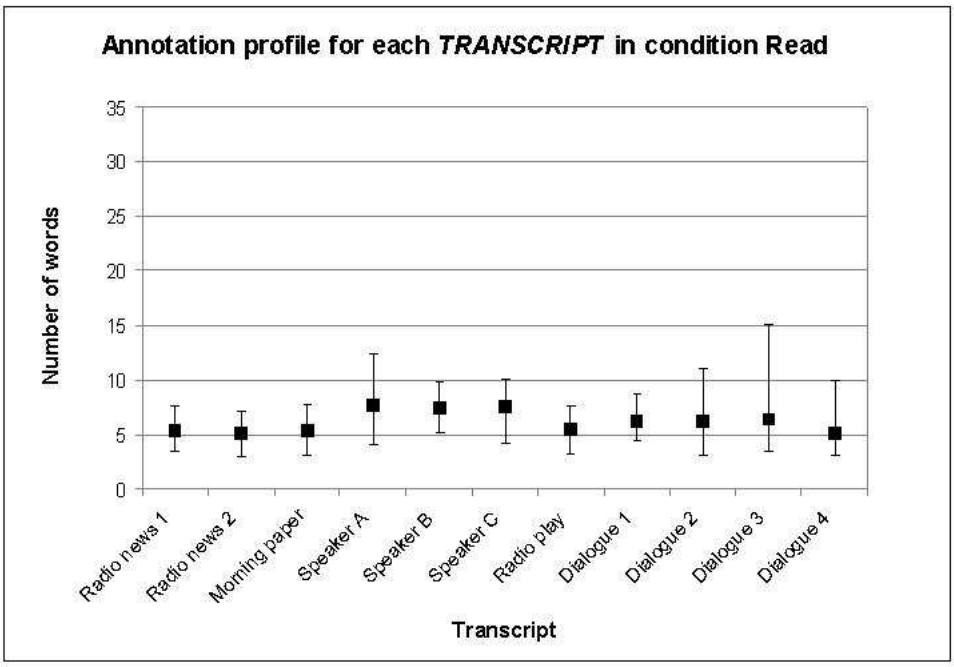

Figure 6.2: Annotation profiles for the transcripts in the prominence marking task, condition Read.

Figure 6.3 shows the annotation profile for each *subject* in condition Listen. Since the subjects are not the same ones across conditions, there is no point in comparing them individually with the subjects in condition Read, but we note the great difference in the subjects' annotation profiles between conditions Read and Listen. The mean annotation frequencies for the subjects vary between about 3.5 and 15 which could be compared to 4 and 8 in condition Read. The range of variation by the subjects in condition Listen differs between subjects and is in a number of cases extremely wide. For instance subject 3 shows a range of variation between 6 and 33.5, meaning that in some transcripts this subject had annotated on average every 6th word as prominent, and in another transcript the same subject had annotated on average about every 33rd word as prominent. In a similar way, a wide range of variation is also found for subject 6 (5 to 28.5). On the other hand some subjects, e.g. subjects 7 and 8, have a very narrow range of variation. All this indicates that the subjects in condition Listen have carried out the prominence marking task in very varied ways.

Based on the observation that a particular subject's annotation strategies are clearly different from those of other subjects, we can assume that the great differences influence the inter-annotator agreement. This applies specially to the the difference between on one hand subjects 3 and 6, annotating with long intervals and very differently across

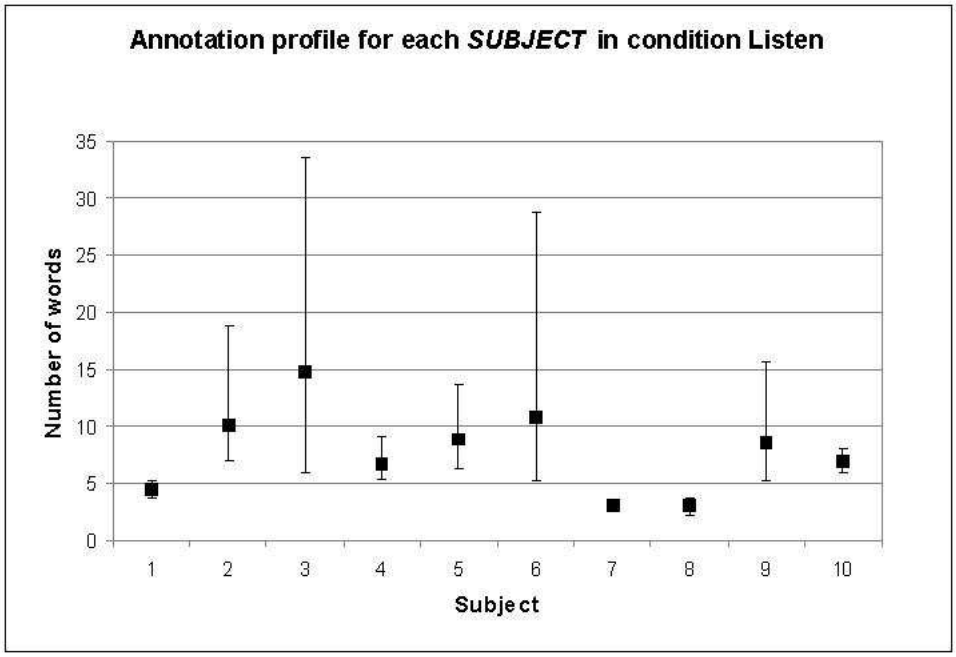

Figure 6.3: Annotation profiles for the subjects, condition Listen.

transcripts, and on the other hand subjects 7 and 8, annotating often and in a very similar way to one another across transcripts.

How is the variation indicated in the annotation profiles for the subjects distributed over the transcripts? Is the wide variation found in figure 6.3 evenly spread over all transcripts, as was the case for the prominence annotations in condition Read, or is there a specific pattern as was the case in the boundary marking task? In figure 6.4 the annotation profiles for the *transcripts* in condition Listen are presented.

Figure 6.4 shows that the average annotation frequency in the transcripts lies between 5 and 10. However, the range of variation is wide in all transcripts, indicating that the subjects annotated the transcripts very differently. Thus, there are no clear differences between e.g. monologue and dialogue as was the case in the boundary marking task. Instead the profiles are rather similar to the profiles in condition Read (see figure 6.2). The greatest variation is – again – found in transcripts "Speaker A" and "Dialogue 3", both of which are parts of the non-scripted speech. The smallest variation is found in "News 1" (scripted) and "Dialogue 4" (non-scripted). In addition, in condition Listen the scripted monologue (News broadcasts 1 and 2 and the morning paper) stand out from the other speaking styles in having words annotated as prominent more often.

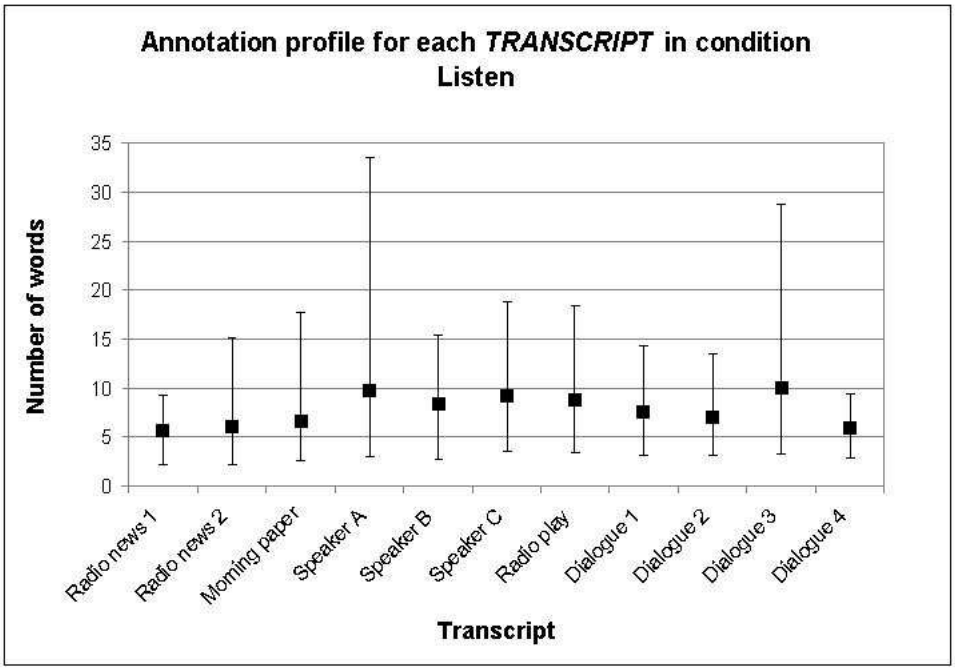

Figure 6.4: Annotation profiles for the transcripts, condition Listen.

The variation in mean annotation frequency for the transcripts in condition Listen is far greater than in condition Read, indicating a high level of disagreement between the subjects on how to annotate the transcripts. This means that it seems likely that having access to the speech signal made the subjects annotate more differently than if they did not have access.

6.2.3 Summary of the Annotation Profiles for the Prominence Annotation Task

Examining the annotation profiles for the prominence annotation task we found a wide variation of the average annotation frequency between subjects. In condition Read the non-scripted monologue has on average slightly longer intervals between the prominence annotations than the other speaking styles do, and in condition Listen the scripted monologue on average has slightly shorter intervals between the prominence annotations. Comparing both conditions we find in both cases a slight difference in average annotation frequency between scripted and non-scripted monologue, whereas the dialogues fluctuate more.

Since the wide range of variation in the subjects' annotation profiles is not unevenly distributed over the transcripts (as was the case in the boundary annotation task), it is plausible to assume that there was a disagreement between the subjects regarding how to annotate the transcripts.

Some transcripts seem to evoke more disagreement than others; if one subject shows some extreme annotation frequency, the extreme values seem to be found in specific transcripts: Speaker A, Speaker C or in Dialogue 3, a fact that indicates that these transcripts were harder for the subjects to annotate.

Having a picture of how the prominence marking task was carried out, we now proceed to examine the inter-annotator agreement in the prominence marking task.

6.3 Inter-Annotator Agreement for the Prominence Annotation

The annotation profiles indicated that the subjects had annotated prominence with rather different strategies. In this section we use the κ statistics to get a measure of these differences in terms of inter-annotator agreement.

For all speaking styles inter-annotator agreement is computed on the full sets of data for all 10 subjects. However, since the subjects differed in their average annotation frequency, and since some subjects had a very wide range of variation, inter-annotator agreement was also computed for a reduced set of subjects. The subjects in this reduced set were selected primarily on the basis of similarity in mean annotation frequency and secondly on the basis of similar or lower range of annotation frequency. Thus, from two subjects with similar mean annotation frequency, the subject with a narrower range would be selected. Where one subject has a narrow range but a more divergent mean annotation frequency, this subject would be excluded and a subject with less deviating mean annotation frequency and wide range would be selected. A higher level of agreement for such a reduced set would indicate that a more uniform annotation frequency also gives a more uniform annotation of prominent words, i.e. that subjects who mark at the same intervals also to a great extent mark the same words. This means that subjects who mark frequently might mark the same core set of words as subjects who mark sparsely, but in addition they also mark many words in between. These words in between can be supposed to cause noise and thus lower the level of agreement.

6.3.1 Inter-Annotator Agreement Within Conditions in the Prominence Annotation Task

Starting with the inter-annotator agreement in condition Read, table 6.3 shows the agreement figures expressed with κ for the prominence annotations (for a closer description of the κ statistics see chapter 4, section 4.4). The inter-annotator agreement in table 6.3 is computed on the full set of 10 subjects.

READ, 10	Monologue	Dialogue	# of subjects
Scripted	0.33	0.46	10
Non-scripted	0.37	0.43	10

Table 6.3: Inter-annotator agreement for prominence annotations, condition Read (10 subj.).

As might have been expected from the annotation profiles, the level of inter-annotator agreement is in general very low. The monologues have a lower level of agreement than the dialogues, κ=0.33 and κ=0.43 for scripted and non-scripted monologue and κ=0.46 and κ=0.37 for the scripted and non-scripted dialogues.

Reducing the data and computing inter-annotator agreement in the set of annotators who have the most similar annotation frequencies noticeably changes the κ values. In table 6.4 the inter-annotator agreement between subjects 1, 3, 4, 6 and 8 is shown. The reduced set of subjects was selected on the basis of similarity in mean annotation frequency and range as described above. In this case, the interval for "similar" mean annotation frequency was between 5 and 8. Thus, subjects 2, 5 and 7 were discarded because of their wide range, subject 9 because of lower mean annotation frequency and subject 10 because of a combination of low mean annotation frequency and a wider range than the selected subject 6. The motivation for selecting 5 subjects and not e.g. 6 or 4 is to make the reduced sample as similar as possible to the reduced sample for condition Listen where the optimum sample concerning similarity was based on 5 subjects.

READ, 5	Monologue	Dialogue	# of subjects
Scripted	0.47	0.59	5
Non-scripted	0.51	0.54	5

Table 6.4: Inter-annotator agreement for prominence annotations, condition Read (5 subj.).

Table 6.4 shows that the level of inter-annotator agreement computed on the reduced set of data is still low, but substantially higher than the figures shown in table 6.4. [1] Thus, the agreement increases when data from subjects with more similar annotation frequencies are used.

[1] In fact, the same experiment with a reduced set was made for the data in the boundary annotation task, but the variation in the inter-annotator agreement was just marginal, i.e. the figures in this case were more stable across subjects and not sensitive to similar changes.

An examination of table 6.4 shows that also here the lowest level of agreement in the monologues (scripted monologue, κ=0.47 and non-scripted monologue κ=0.51) and the level of agreement in the dialogues is higher (scripted dialogue, κ=0.59 and non-scripted, κ=0.54). Thus, the difference between monologue and dialogue as well as scripted monologue and dialogue as the two extremes is retained.

The results from the two sets of data support the interpretation that the low κ values to some extent are due to the subjects' different understanding of the task. However, since the results are proportionally similar, there seems to be a core set of words where the annotators to a great extent agree.

In the next section we show how subjects' access to the speech signal has influenceed the inter-annotator agreement.

In a similar way to condition Read we compute κ for two sets of data: The full set of data with all 10 subjects and a reduced set of data with the 5 subjects with the most similar annotation frequencies. The figures for the inter-annotator agreement based on all 10 subjects is shown in table 6.5.

LISTEN, 10	Monologue	Dialogue	# subjects
Scripted	0.35	0.42	10
Non-scripted	0.36	0.39	10

Table 6.5: Inter-annotator agreement for prominence annotations, condition Listen (10 subj.).

The level of agreement in condition Listen is about as low as that in condition Read. The pattern from condition Read with lower agreement for the monologues compared with the dialogues is also found in condition Listen, however the difference is smaller: scripted monologue, κ=0.33, non-scripted monologue κ=0.36 and scripted dialogue, κ=0.42, non-scripted dialogue κ=0.39.

Table 6.6 shows the figures for the reduced set of data computed on subjects number 1, 4, 5, 9 and 10. Also here the reduced set is selected on the basis of similarity in mean annotation frequency and range. Thus, subjects 2, 3 and 6 are discarded on the basis of diverging range, and subjects 7 and 8 on the basis of a very high (few words per prominent word) annotation frequency.

LISTEN, 5	Monologue	Dialogue	# subjects
Scripted	0.45	0.55	5
Non-scripted	0.46	0.52	5

Table 6.6: Inter-annotator agreement for prominence annotations, condition Listen (5 subj.).

In a similar way to condition Read, the inter-annotator agreement in condition Listen increases when computed for subjects with more similar annotation frequency, and

the proportional difference between the speaking styles are retained also in this case. The monologues have a lower level of agreement (scripted monologue, κ=0.45 and non-scripted monologue, κ=0.46) while the agreement in the dialogues is slightly higher (scripted dialogue κ=0.55 and non-scripted dialogue, κ=0.52.

Even though the level of inter-annotator agreement in general is very low in the prominence annotation task, there is a similar pattern in the prominence annotation of condition Read and the boundary annotation in condition Read: higher inter-annotator agreement for the dialogues. Moreover, in the prominence annotation task, as in the boundary annotation task, this advantage for the dialogue is reduced in condition Listen. In the boundary marking task some of this effect could be traced to the transcription of the dialogues and the subjects' preference to mark a boundary at speaker change. This explanation would, however, not be applicable in the prominence annotation.

The proportional stability across the full sets of data compared with the reduced sets of data supports the interpretation that to some extent the subjects agree on a core set of prominent words, but since the granularity of annotation is very different across subjects, the subjects with high annotation frequency introduce noise. This in turn indicates that the subjects were unsure about the task, and that more definite instructions would have influenced the agreement positively. Thus, even though the κ value in general is low, it seems that subjects agree on a core set of words, and depending on how they interpreted the granularity of the task they introduced more or less "extra annotations" which lowered the level of agreement.

6.3.2 Inter-Annotator Agreement Across Conditions in the Prominence Annotation Task

The level of inter-annotator agreement in the prominence marking task is in general very low. However, do the annotators agree across conditions on the positions when they after all agree within conditions?

If there is a high level of agreement across conditions, this can indicate that to a great extent the subjects were guided by the string of words (i.e. the word/constituent order), since the subjects in both conditions have access to the same string of words. However, the more the subjects (in condition Listen) have relied on information present only in the speech signal, the lower we can expect the level of agreement across conditions to be.

To make a comparison across conditions we use the same method of computing inter-annotator agreement between the conditions as in the boundary annotation task, i.e. we compute inter-annotator agreement for the positions where the majority of subjects have annotated a word as prominent in condition Read and in condition Listen. For a more detailed description of the method see chapter 5, section 5.2.2.

To give the reader a picture of the proportions of the data used in this comparison across conditions, we show the distribution of the subjects' annotations, starting with condition Read (see figure 6.5).

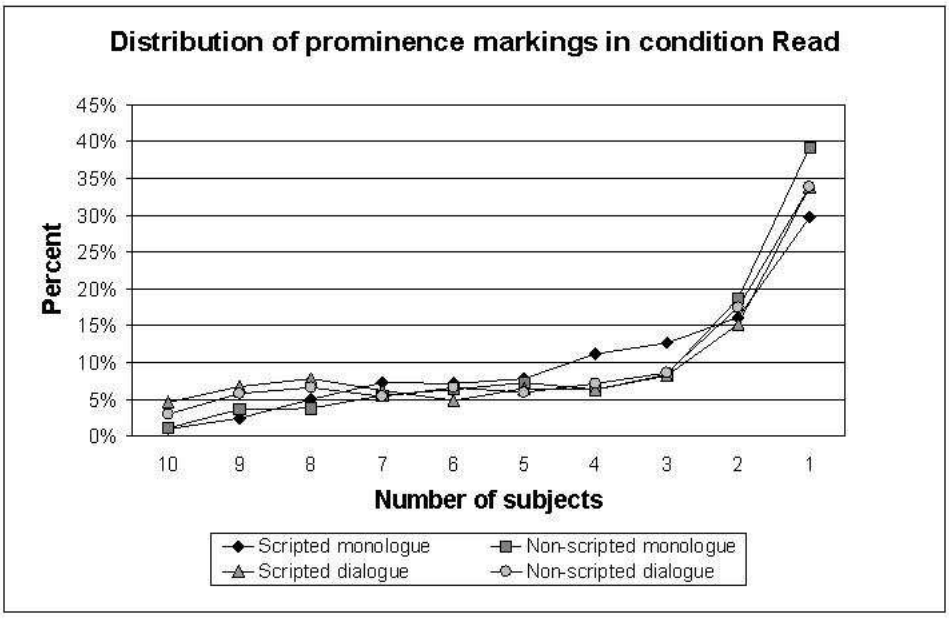

Figure 6.5: Proportion of subjects per prominence annotation in condition Read.

Horizontally we find the number of subjects who agree on a prominence marking, and vertically we find the proportion as a percentage of the data of such an agreement. Thus, the proportion of the prominence markings where all 10 subjects agree is very low, in the non-scripted speaking styles close to non-existent, and in the scripted speaking styles around 5%. The proportion of data is similarly small in all cases where we can talk about a majority agreement (at least 6 subjects), and the proportion rises dramatically first at agreement by at least 2 subjects. There is no greater difference across the speaking styles, and very crudely the majority annotations constitute about a quarter of all prominence annotation positions.

The proportions in condition Listen, given in figure 6.6 are distributed in a very similar way to those in condition Read, with the difference that the level of agreement between 10, 9 and 8 subject is even slightly lower. Also in this case the set of majority annotation positions constitutes roughly a quarter of all prominence annotation positions.

Now to the comparison of the majority agreement positions across conditions Read and Listen. If these positions are the same across conditions we will get a high κ value, but if there is a disagreement between condition Read and condition Listen about where the prominent positions are, we will get a low κ value. The result is found in table 6.7.

Also in the comparison between condition Read and condition Listen we find inter-annotator disagreement rather than inter-annotator agreement. The κ values are in

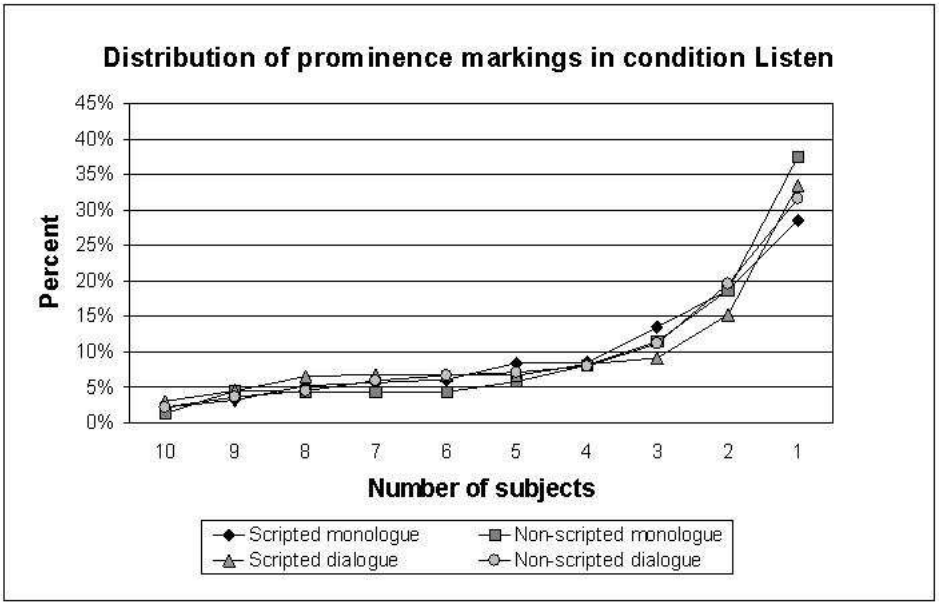

Figure 6.6: Proportion of boundary annotations per subject in condition Listen.

COMPARE	Monologue	Dialogue
Scripted	0.56	0.64
Non-scripted	0.53	0.54

Table 6.7: Inter-annotator agreement for prominence annotation (majority agreement) between conditions Read and Listen.

general low, very similar across the speaking styles (scripted monologue (κ=0.56), non-scripted monologue (κ=0.53), non-scripted dialogue (κ=0.54)), and only in the scripted dialogue there is a slightly higher level of agreement (κ=0.64). Please note that these κ values concern only the agreement across conditions for the agreement within conditions, and they should not be directly compared with the κ values for the inter-annotator agreement within conditions. For a more elaborate discussion, see chapter 5, section 5.2.2

The fact that the κ values are very low indicates that the positions where subjects after all did agree in condition Read differed from those where subjects did agree in condition Listen. Thus, the low inter-annotator agreement here indicates that information carried by the speech signal has influenced the subjects in condition Listen so that they annotate differently from the subjects in condtion Read. This is clearly different from the boundary annotation task, where the level of agreement across conditions is high and

thus indicates a lower impact of the prosody. We return to this issue in the discussion in section 6.7.

To sum up this section on inter-annotator agreement, we can state that the level of agreement in general in both conditions Read and Listen is very low in the prominence annotation task. In both conditions Read and Listen, the level of agreement for the monologue is lower than that for the dialogue. In addition, this proportional difference between the speaking styles is stable both across conditions and across full and reduced data sets. It was also shown that a higher similarity between the subjects regarding the annotation frequencies increased the κ values, and this indicates that the subjects to some extent agree on a core set of prominent words, even though the general agreement is low due to "extra" annotations of prominent words.

Concerning the agreement across conditions the figures were very low compared with the boundary annotation task. Thus, we can suspect that the impact of prosody was stronger in the prominence marking task than in the boundary marking task.

Now that we have a picture of the agreement in the prominence annotations, we can turn to examining the linguistic and their acoustic context.

6.4 Phrase Level Properties of the Prominence Annotations

Turning to a closer inspection of specific features of the boundary annotations, we start with the examination of the phrasal properties. The phrase context is the first specific linguistic context examined in order to obtain a more detailed picture of where the subjects actually agree on a prominence marking. The set of data we use is the words annotated as prominent by the majority of subjects, i.e. positions where at least 6 subjects agree on a prominence marking. In the examination of the phrase context we inspect the same features as those in the examination of the phrase context in the boundary annotation study:

- The phrase depth for the prominent words in each speaking style.
- The phrase context of the prominent words in each speaking style.

In addition we briefly recapitulate the pictures of the general phrase depth in the speaking styles, as well as the general phrase frequencies. However, for a detailed description the reader is referred to chapter 5, section 5.3.

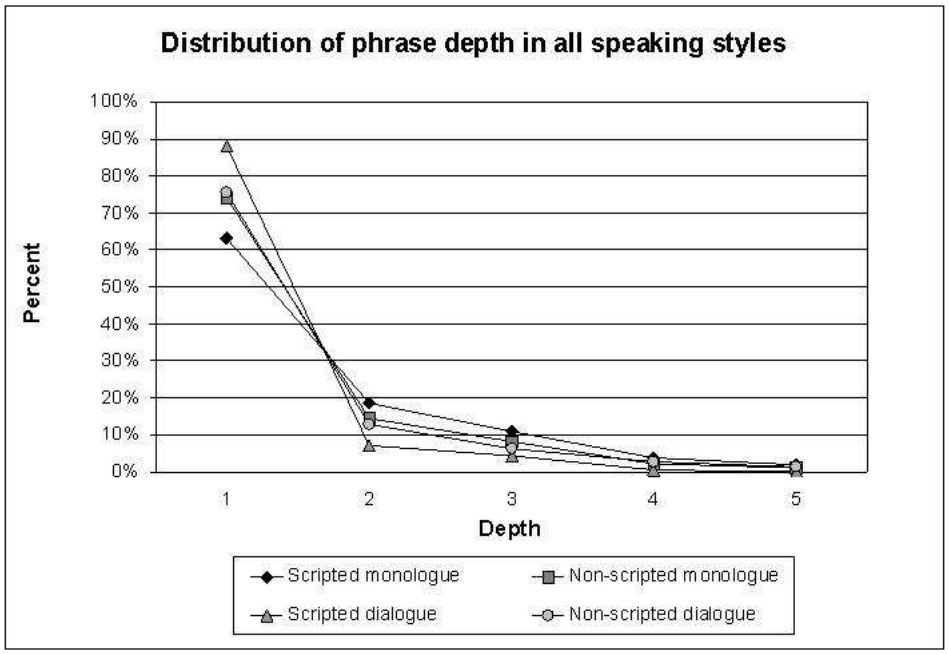

Figure 6.7: Phrase depth distribution in the four speaking styles.

6.4.1 The Phrase Depth of the Prominent Words

To remind the reader of the general distribution of the phrase depth in the speaking styles, the general phrase frequencies from chapter 5 are recapitulated in figure 6.7. In this section the figure serves only as a reference point, and is not discussed in detail. However, a more elaborate comment is found in chapter 5, section 5.3.1.

In the boundary annotation the position in front of a non-embedded phrase was favoured over the position in front of a more deeply embedded phrase, i.e. the position in front of the beginning of a phrase was favoured over the position deeper inside a phrase. This is what we might expect, since in general boundaries come at breaks, and not inside units. However, prominent units differ in boundaries. Since prominent new information, and also sentence accent, in general come in the later part of the sentence, we could expect to find prominent words in more embedded positions more often. Thus, we might get a different pattern in the prominence annotation task than in the boundary annotation task. In a similar way to the boundary annotation task, we measure the degree of embeddedness with the shallow parse. Figure 6.8, showing the distribution of phrase depth for the words marked as prominent in condition Read, shows that the picture differs from the boundary annotation.

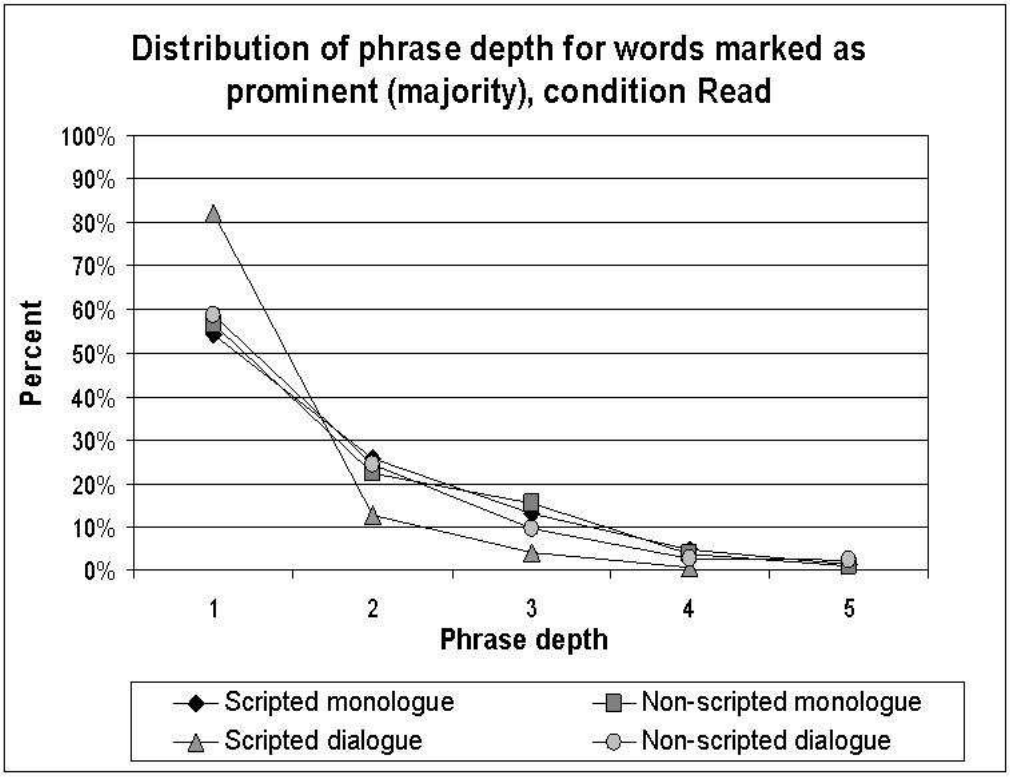

Figure 6.8: The phrase depth of words annotated as prominent, condition Read.

In figure 6.8 it is shown that all speaking styles have a slightly higher proportion of prominent words in more embedded positions than is the case with the general distribution of words in phrases. The only speaking style which shows a specific profile is the scripted dialogue which has a greater proportion of prominent words in less embedded phrases than the other three speaking styles, all three of which show very similar curves.

Since the phrase depth for the words annotated as prominent in condition Listen was very similar to the one for condition Read, we do not show a separate figure for condition Listen. Instead we state that the phrase depth properties of words annotated as prominent are nearly identical across conditions Read and Listen.

Compared with what was the case for the boundaries there is a higher proportion of words in the context after boundaries in less embedded phrases. This means that to some extent boundary annotations are favoured at the beginnings of phrases, while prominence annotations, more often than boundary annotations, are found at more embedded positions.

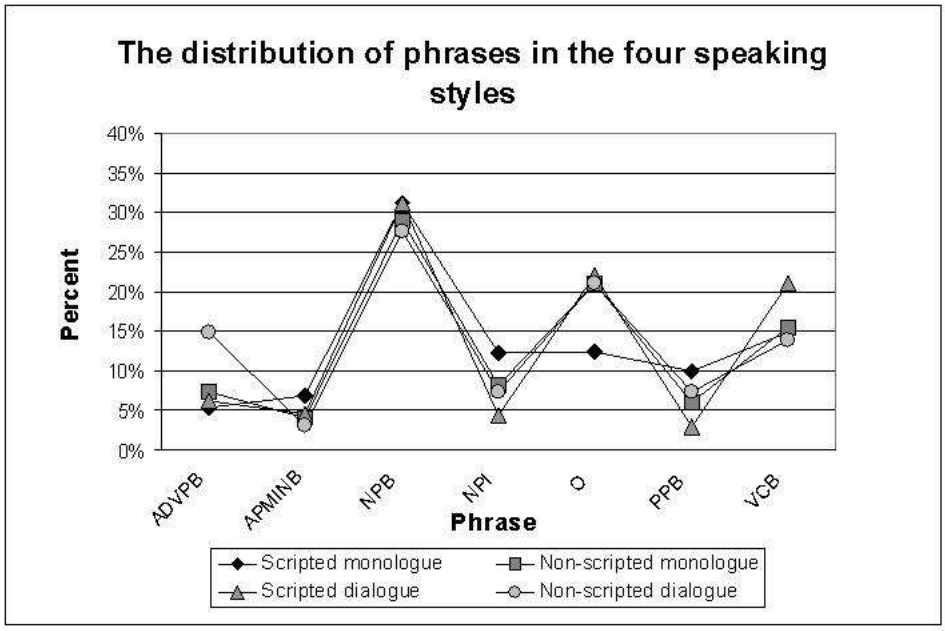

Figure 6.9: The distribution of phrases in the four speaking styles.

6.4.2 The Phrasal Context of Prominence Annotations

In what phrases do we find the prominent words in the different speaking styles? Are there differences between the speaking styles within conditions and/or across conditions? Are some specific phrases marked more often as prominent than other phrase categories? These are the questions to be answered in this section. As reference point the figure of the general phrase distribution from chapter 5 is repeated in figure 6.9.

Figure 6.9 shows that the four speaking styles have a fairly uniform distribution of the phrases, the most notable exception is the lower frequency of words outside the scope of any phrase (i.e. for example conjunctions) in the scripted monologue. For a more detailed comment on the general phrase distribution please see chapter 5, section 5.3.2.

In which phrases are the words annotated as prominent located? Figure 6.10 shows the phrase distribution of words marked as prominent by the majority of subjects in condition Read.

All phrases shown in the figure have reached at least 5% in some speaking style. Compared to the general phrase frequencies there is one salient divergence: VCI (verb cluster inside) is added to the phrase frequencies for the words annotated as prominent. Thus, VCI is clearly more frequent as a prominent word than in general in the speaking styles.

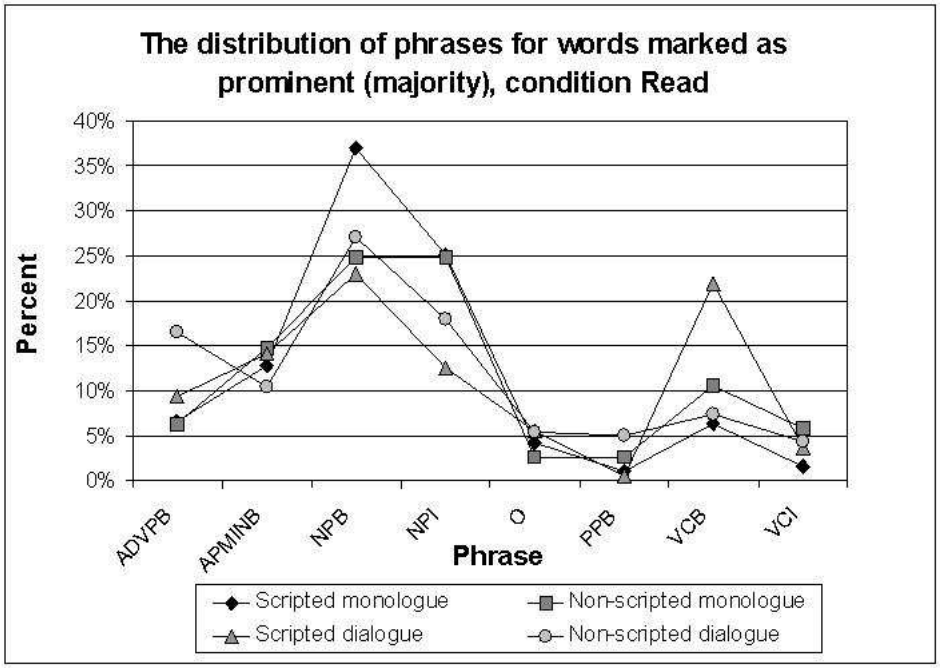

Figure 6.10: Phrase context of the prominence annotations, condition Read.

Inspecting figure 6.10 we note that each of the four speaking styles has a specific profile, i.e. there is no clear pattern for monologue, dialogue, scripted or non-scripted speech. However, even though the proportions differ across speaking styles, the peaks are the same. Thus, in all four speaking styles noun phrases constitute an important category for prominent words, both NPB (at the beginning of noun phrases) and NPI (inside noun phrases).

At four points, one for each speaking style, departures from the general tendency are found. First, the non-scripted dialogue has a high proportion of adverb phrases, but this divergence is also found in the general distribution in figure 6.9. That means that this divergence might be an effect of more adverb phrases in general in the non-scripted monologue and not a preference by the subjects to annotate adverb phrases as prominent in non-scripted dialogue.

Inspecting the noun phrases (NPB) the scripted monologue singles out as having a higher proportion of words labelled with NPA annotated as prominent than the other speaking styles do. In this case it seems that the subjects prefer to annotate these phrases as prominent, since the general distribution in figure 6.9 indicates about the same frequency of NPB for scripted monologue as for the other speaking styles.

The proportion of verb clusters (VCB) marked as prominent is higher in the scripted dialogue than in any other speaking style. Also in this case the scripted dialogue in general contains a higher proportion of VCB, which means that this might be an effect of the general distribution rather than a marking preference by the subjects.

The annotation of prominence is concentrated to the noun phrases, the adverb phrases, the adjective phrases, and to some extent to the verb clusters. Compared to the general distribution in figure 6.9 the categories noun phrases, adverb phrases and adjective phrases are proportionately more frequent among the prominent words than in the general distribution.

In all speaking styles the noun phrases (NPB and NPI together) constitute the largest category of words annotated as prominent. The highest proportion is found in scripted monologue, then in non-scripted monologue, in non-scripted dialogue and finally in scripted dialogue. The proportion of prominent noun phrases follows the order of the speaking styles based on NQ and NVQ.

We shall now examine the phrase distribution of words marked as prominent in condition Listen. The proportion of the phrases of words annotated as prominent in condition Listen is shown in figure 6.11.

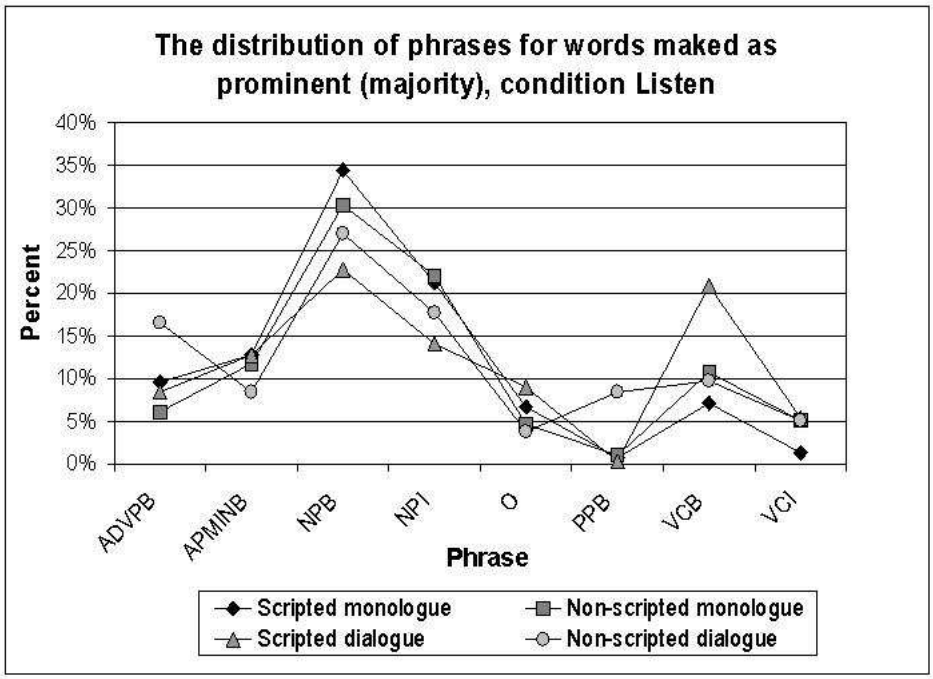

Figure 6.11: Phrase context of the prominence annotations, condition Listen.

In condition Listen the word marked as prominent are distributed across the phrases in proportions similar to what was found in condition Read, however, with some exceptions.

The most notable one is the higher proportion of noun phrases (NPB) in the non-scripted monologue.

This means that the subjects in condition Listen in general annotate as prominent words from the same set of phrasal categories as the subjects in condition Read. There are slight differences, but the most prominent feature is the similarity across conditions.

6.4.3 Summary of the Phrase Level Properties

In both condition Read and condition Listen the prominent words in the four speaking styles are found in similar phrase contexts: the most common phrase category among the words annotated as prominent is the noun phrase (NPB and NPI). The other large phrase categories among the words annotated as prominent are adverb phrases, adjective phrases and verb clusters. This means that in both condition Read and condition Listen the prominent words come from similar phrasal categories, but since the κ values indicated such a low level of agreement across conditions, we can assume that they came from different instances of these categories.

In the boundary annotation task, the phrase distribution across speaking styles showed very small differences. An examination of the boundary annotations in the context of parts-of-speech showed clear differences across speaking styles. Therefore, an analysis of the part-of-speech context of the words annotated as prominent was also made. However, the results did not change the picture from what was already found in the phrase context, i.e. words annotated as prominent were mainly nouns, adverbs, adjectives and verbs. Thus, instead of changing the picture, as in the boundary annotation task, the part-of-speech analysis in the prominence marking task confirmed the picture from the phrase context. Therefore, since the part of speech context does not add anything new, it is not reported here in detail.

The analysis of the grammatical context did not indicate very great differences, neither between speaking styles nor across conditions. However, the κ analysis indicated a substantial great difference in the set of words annotated as prominent across conditions Read and Listen, indicating that prosody causes a difference in the subjects' annotation of prominence.

To get a fuller picture of the prosodic properties of the words annotated as prominent, we make an acoustic analysis of those words.

6.5 Prominence Annotations and Focal Accent

Trying to identify the difference between condition Read and condition Listen, the primary question would be "what makes the listeners classify other words as prominent than the readers?". Thus, the ideal would be an examination of the acoustics of the words

which are preferred as prominent in condition Listen compared with the acoustics of the words that are not preferred as prominent in condition Listen, but preferred in condition Read. However, since our materials do not have this extensive acoustic mark-up, but only a mark-up of the words that are annotated as prominent by the majority of subjects in condition Listen, this question is not considered. Instead, we work along the lines indicated by the κ analysis and the study of the phrasal features, i.e. that something in the acoustics caused the subjects to consider a certain set of words as prominent, and deal with the questions "What acoustic properties do words annotated as prominent by the majority of subjects in condition Listen have?" and "Are there differences across the speaking styles in the acoustic properties in the words annotated as prominent by the majority of subjects?". Thus, all results referring to acoustic features of prominent words are based on the words annotated as prominent by the majority of subjects in condition Listen, and since a complete acoustic analysis of all data was not carried out, prominent words from condition Read are disregarded.

In the examination of the prosodic and acoustic features we first study some acoustic correlates to prominence: average pitch, average standard deviation in pitch, duration and intensity. Each of the features was measured on the words annotated as prominent in the four speaking styles in condition Listen. Then we examine to what extent words annotated as prominent could be classified as focal, and touch on other factors that might cause a word to be annotated as prominent.

6.5.1 Acoustic Properties of the Prominent Words

In this section we examine some acoustic features often mentioned as correlates to accent and prominence. The features are F0, duration and intensity. Each of these features is measured on the whole word annotated as prominent. In other words, the average F0 is measured for the word, as well as the duration and the average intensity, In addition the standard deviation in the F0 is measured for each word annotated as prominent. All measurements were done in Praat (Boersma and Weenink, 1996) by the author.

Do we find differences in these acoustic features which can be related to the speaking styles? The investigation of the phrasal context did not yield any clear patterns for the speaking styles as was the case in the boundary annotation task. How will the picture develop concerning the acoustic features?

Since F0 is related to the voice, while a pause in our definition is related to silence, we can expect that for this study differences across speakers will colour the picture more than was the case for the boundaries. Therefore we start by examining closer the three speakers who are constant over three speaking styles, and then relate the tendencies in these three speaking styles to the fourth speaking style. First we examine speakers A, B and C (and for the moment disregard speaker D) in the speaking styles scripted and non-scripted monologue and non-scripted dialogue. Afterwards the results are related to

the rest of the data (the scripted dialogue). The first acoustic feature considered is the F0.

Figure 6.12 shows the average F0 measured for all words in the speaking styles scripted and non-scripted monologue and scripted dialogue. The black diamonds show the average figures for all three speakers in each speaking style, squares represent speaker A (female), the triangles speaker B (female) and the circles speaker C (male). These average F0 measurements are used as a reference point for the F0 values of the words annotated as prominent (shown in figure 6.13). The mean value is computed as the mean for the speakers, which means that e.g. more data collected from one speaker does not influence the mean value for all speakers.

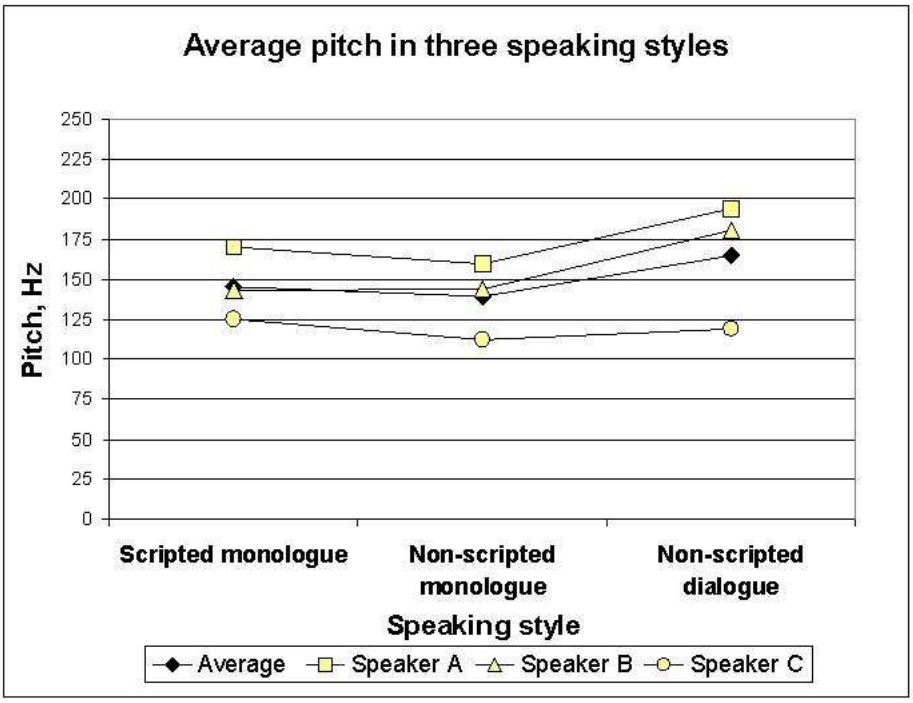

Figure 6.12: The mean figures for pitch in three speaking styles.

Figure 6.12 indicates differences in the average F0 between the speaking styles. The average figures for each speaking style (diamonds) indicate a lower F0 for the non-scripted monologue (139 Hz) than for the scripted monologue (145 Hz). The highest mean is found in the non-scripted dialogue (165 Hz). The three speakers follow this tendency with some exceptions: Speaker B does not show much difference between the scripted and non-scripted monologue, and speaker C does not have a high value for non-scripted dialogue but is rather on a level with the scripted monologue.

Turning to the F0 values for the words annotated as prominent, rendered in figure 6.13 we find that there is a similar over-all pattern, but with a general higher mean. The

average mean for the prominent words in the scripted monologue is 169 Hz, in the non-scripted monologue 159 Hz and in the non-scripted dialogues 202 Hz.

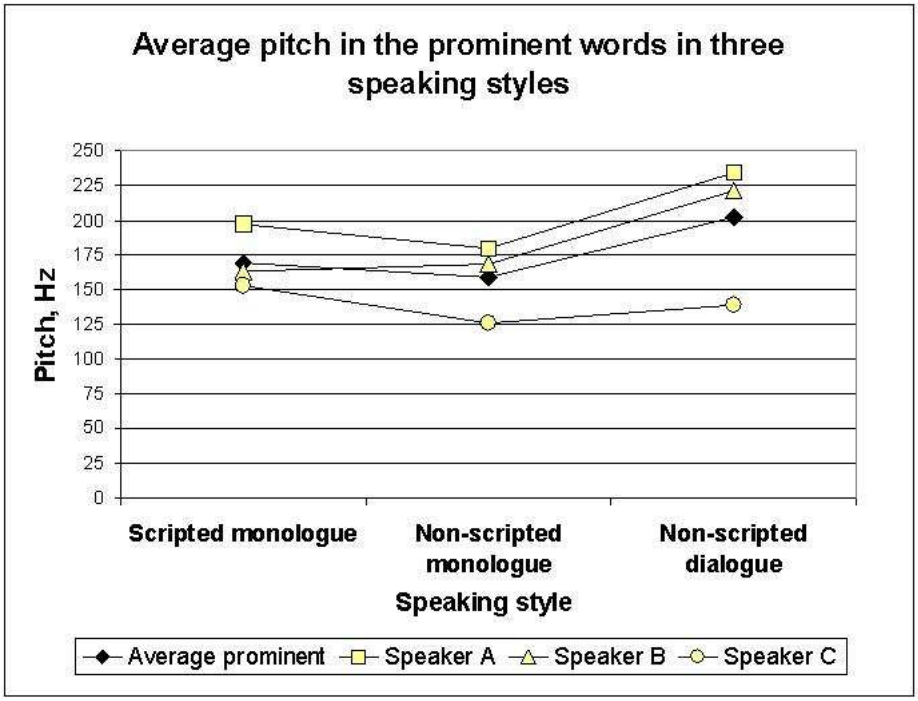

Figure 6.13: The mean figures for pitch in prominent words in three speaking styles.

Also in this case the speakers' mean follows the average mean, with the same exceptions as for the speaking styles' general mean (shown in figure 6.12: There is no difference between scripted and non-scripted monologue for speaker B and no great rise in the non-scripted dialogue for speaker C. Thus, for all three speaking styles and all three speakers it is generally valid that the words annotated as prominent on average have a higher mean F0 than the average for the speaking styles. In addition the interval of difference is similar for all the speakers and all the speaking styles (15 – 28 Hz higher), with the exception of speakers A and B in the non-scripted dialogue where the difference between their average mean pitch and the mean pitch in the prominent words is 44 Hz for speaker A and 40 Hz for speaker B.

The lack of difference between scripted and non-scripted monologue in the case of speaker B might be due to the fact that speaker B has a trained voice and has been working as e.g. speaker on television. Thus, the lack of difference between scripted and non-scripted monologue in the case of speaker B might be due to the fact that the situation of text-based speech is not as unfamiliar to speaker B as it is to speakers A and C, and that speaker B, furthermore, when reading professionally consciously strives to be natural.

Both speaker A and speaker C shows a clear reduction of the average F0 in the non-scripted monologue compared with the scripted one, and using a t-test, in both cases

this difference is significant ($p<.0001$). In the case of the increase in the average F0 in the dialogues compared with the monologues, the difference is significant by comparison with both monologues for speaker A and B ($p<.001$). For speaker C the difference was significant only by comparison with the non-scripted monologue ($p<.05$) and not by comparison with the scripted one.

Where significant differences concerning F0 were found between speaking styles, these differences also tended to be supported by differences in the average measurements for standard deviation of F0 and average intensity in the prominent words. Thus, the relatively low F0 in the non-scripted monologue was combined with a lower average of standard deviation for F0 and a lower average intensity in the prominent words. A full record of the three measures in all three speaking styles is shown in figure 6.14.

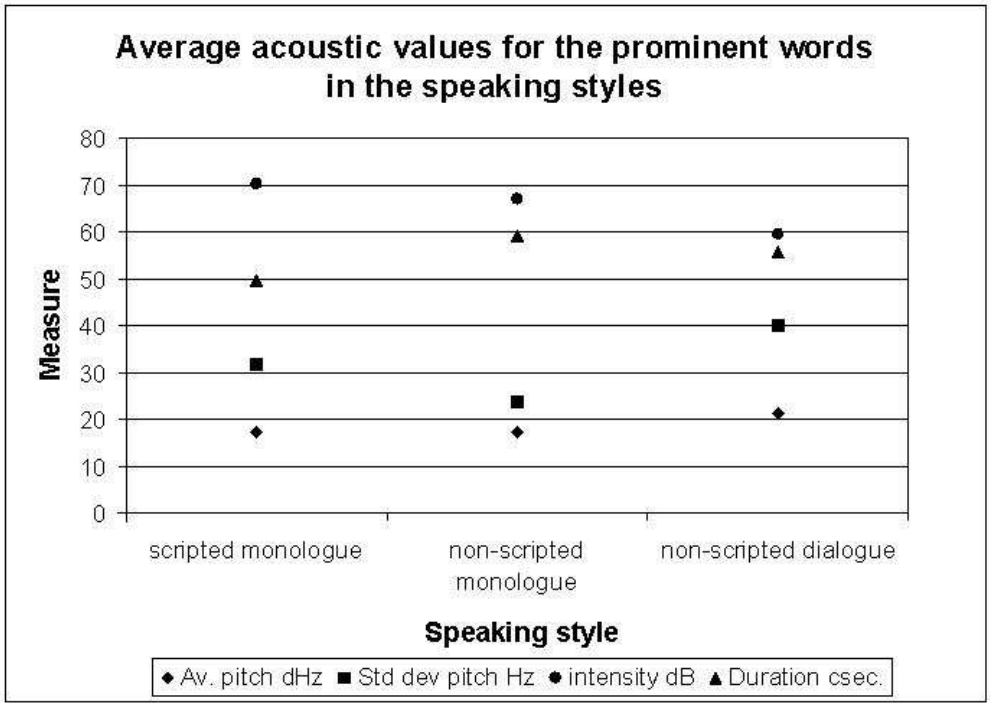

Figure 6.14: Average acoustic measures in three speaking styles.

Please note the difference in units of measurement between the measurements in figure 6.14. Average F0 is shown in deciHertz (thus, 17 deciHertz means 170 Hertz), while an average standard deviation of F0 is shown in Hertz. Average intensity is shown in dB, and average duration is presented in centiseconds(60 centiseconds means 600 milliseconds). This conversion was made in order to make all the measurements fit into one diagram, so that a combined picture of all three speaking styles could be presented.

All the differences in the average measurements in figure 6.14 are significant (at least $p<.001$) except the F0 difference between scripted and non-scripted monologue and the difference in duration between non-scripted monologue and non-scripted dialogue.

In addition, the difference in duration between scripted monologue and non-scripted dialogue is barely significant with $p<.05$.

The average F0 in the different speaking styles (figures 6.12 and 6.13) varies between speakers. This is also the case in the rest of the acoustic measurements, and the differences between speaking styles are not significant for all the individual speakers to the same extent as the averages presented in figure 6.14 are. However, since in most cases the tendencies for all the individual speakers pointed in the same direction as the averages in figure 6.14, we consider it fair to account only for the averages.

The acoustic measurements give a rather specific acoustic profile for the prominent words in each of the three speaking styles. The picture that evolves is that the F0 in the prominent words is on average lower and less varied (lower standard deviation of F0) in the non-scripted monologue than in the scripted one. In addition the prominent words in the non-scripted monologue are of a longer duration and a lower intensity than those in the scripted monologue.

In the non-scripted dialogue the words annotated as prominent are on average characterized by a higher F0, a higher standard deviation of F0 and a lower intensity than those in both monologues. Furthermore the duration value lies between the values for the scripted and the non-scripted monologues.

It is not possible to compare the scripted dialogue directly with the other speaking styles, since there are different speakers in this style from the three other ones. However, to give a rough picture of the relationship of the scripted dialogue to the three other speaking styles, we have included scripted monologue (as well as speaker D) in an overview of all four speaking styles in figure 6.15. In the scripted dialogue, the duration of the prominent words turned out to surpass by far the three other speaking styles, therefore the duration value of the scripted dialogue appears in an insert in the figure.

Thus, even though no differences were indicated regarding the kind of words annotated as prominent, i.e. phrasal categories and parts of speech, there seem to be some differences in the acoustic realisation. Since traces of such differences are found also in the average measurements of the pitch for the whole speaking styles, we consider these findings to be characteristic of all the speaking styles, not just to the prominent words in the speaking styles.

6.5.2 Prominence and Focality

Are the words annotated as prominent in general focal in terms of having an F0 contour that could be classified as focal? In case we find prominent words which not could be classified as focal, what factors can have caused a majority of subjects to annotate these words as prominent? These are the primary questions treated in this subsection.

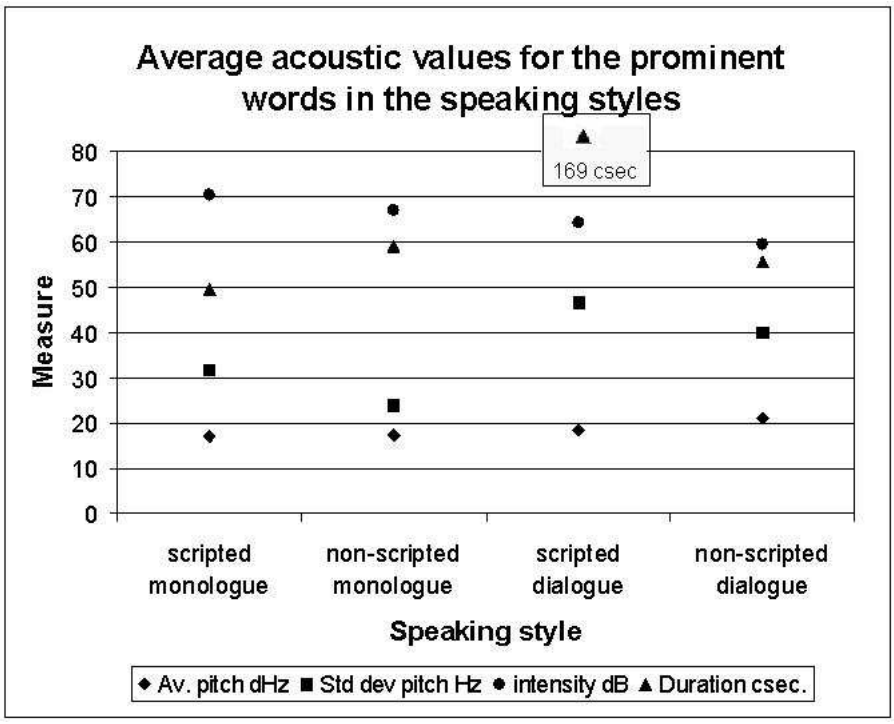

Figure 6.15: Average acoustic measures in all speaking styles.

How great a proportion of the words annotated as prominent are actually focal in terms of a focal accent gesture? Table 6.8 shows the number of words, the proportion of words annnotated as prominent and the proportion of words classified as focal for each speaking style in condition Listen. A rather similar proportion of words are annotated as prominent by the majority of subjects in all the speaking styles. In scripted monologue the figure is 13%, and in non-scripted monologue 9%. The dialogues lie on 14% (scripted) and 11% (non-scripted). Of the prominent words 69% are classified as focal in the scripted monologue, whereas the figure is 77% in the non-scripted monologue. In the scripted dialogue 79% of the prominent words are focal, and in the non-scripted dialogue the figure is 73%. Thus, most of the words annotated as prominent are also characterised by a focal accent, however, we also find a large proportion of words that lacks such accent. Which are these words?

Words annotated as prominent which do not have a correlate in focal accent seem in many cases to become prominent by means of connection to some other prominent word, or to some specific context. Table 6.9 shows the categories found in our data, and also the proportion of prominent words in each category.

Category 1. contains sequences such as "...hon vet inte..." ("...she knows not..."). The underlined word is annotated as prominent, but it is not acoustically emphasised. However,

	Scripted monologue		Non-scripted monologue		Scripted dialogue		Non-scripted dialogue	
# of words	2858		2084		2212		2248	
	N	*%*	*N*	*%*	*N*	*%*	*N*	*%*
prominences	380	*13*	197	*9*	320	*14*	240	*11*
focal	263	*69*	152	*77*	254	*79*	174	*73*
non-focal	117	*31*	44	*22*	66	*21*	65	*27*

Table 6.8: The words annotated as prominent which have a focal accent.

the preceding word, "vet", is focal, and it thus seems that the subjects have interpreted the verb and the negation together as one element.

Category 2. resembles category 1., the difference is that the words are more distant from each other. an example of a sequence in this category is "...inte använda vatten till disk och dusch..." ("...not to use water for dish washing and showering..."). In this example "vatten", "disk" and "dusch" could be said to be closely related, and all three are annotated as prominent. However, just "vatten" is focal, "disk" is acoustically very weak and "dusch" is clause final with a pitch movement which was classified as a high boundary tone. The tendency to annotate all three words as prominent might instead be due to the fact that the concept of water is important in that sequence of discourse, and the word "vatten" is emphasised, therefore also "disk" and possibly "dusch" are coloured with importance and thus annotated as prominent.

Category 3. contains clause initial words, often in the position after a pause and articulated with a higher F0 and intensity than words inside the clause. Thus, since these words can be said to have a rather specific status in the clause, it is not strange to find some of them annotated as prominent.

Category 4. contains clause final words.

Category 5. phrase final words.

Category 6. contains proper names, which often seem to be annotated as prominent even though they are not focal. In some cases one name in a sequence of a first name and a surname is acoustically emphasised and the other is not, although both are annotated as prominent.

Category 7. contains sequences from the news headings at the beginning of the read radio broadcasts. In these passages the news announcer lists the headings in a very specific style, typical of Swedish radio broadcasts. These headings have some similarity to the material in category 8.

	Scripted monologue		**Non-scripted monologue**		**Scripted dialogue**		**Non-scripted dialogue**	
	N	%	*N*	%	*N*	%	*N*	%
1. Adjacent other prominent word	*18*	**15**	*1*	**2**	6	**8,5**	*13*	**20**
2. Related to other prominent word	*18*	**15**	*10*	**21**	5	**7**	*3*	**4**
3. Clause initial	*9*	**7**	*1*	**2**	5	**7**	*6*	**9**
4. Clause final	*23*	**19**	*14*	**30**	37	**51**	*12*	**17**
5. Phrase final	*11*	**9**	*10*	**21**	5	**7**	*8*	**12**
6. Proper name	*19*	**15**	*0*	**0**	0	**0**	*0*	**0**
7. In heading	*8*	**7**	*0*	**0**	0	**0**	*0*	**0**
8. In list	*4*	**3**	*0*	**0**	4	**5,5**	*0*	**0**
9. Other	*13*	**11**	*11*	**23**	10	**14**	*25*	**38**
Total number of cues	*125*		*47*		*72*		*65*	
Total of non-focal words	*117*		*44*		*66*		*65*	

Table 6.9: Categories of non-focal words annotated as prominent.

Category 8. contains list structures. Also in this case a number of elements are presented in sequence, a structure which in itself seems to give the annotators the impression of prominence. An example from category 8., lists, is the following sequence: "...beskriva rött, kärlek eller tandvärk..." ("describe red, love or tooth-ache..."). In this particular case only "tandvärk" is focal, but also "rött" and "kärlek" are annotated as prominent by the majority of subjects.

Thus, in most cases the words annotated as prominent which do not have a clear acoustic emphasis are found at positions where the structure of the message supports prominence, or in relation to other prominent words or expressions which seem to be able to colour the acoustically less prominent word with an air of importance and prominence.

Even though part of the words annotated as prominent do not have a correlate in focal accent, the majority of them are classified as focal. Thus, in our data most prominent words are focal, and in the scripted styles most non-focal prominent words have other clear cues to prominence. However, in the case of non-scripted styles, the cues for the non-focal prominent words are not so clear. To some extent they can be classified as related to other prominent words, but a great proportion falls into the category "Other". One reason for this might be that we have examined fairly simple grammatical correlates, but in the non-scripted styles the correlates might be more complex and/or of pragmatic nature, the consequence of which might be that they do not fit in into the above categories. We leave this issue for now, but shall return to it in the discussion in section 6.7.

6.6 The Boundary and Pause Context of the Prominence Annotations

It remains to investigate one feature of the prominences: their relationship to discourse boundaries and their relationship to pauses. Many researchers have pointed out the tendency for prominent information to come at the end of a sentence, both in work on information structure, where the prominent information interpreted as e.g. "new", and in work on prosody, where the prominent information is to interpret as acoustically prominent. Is this something we can see in the data in terms of prominences in the context before a boundary position? To study this we have extracted all words annotated as prominent and examined i) how great a proportion is located to the context before a discourse boundary? and ii) how great a proportion is located to the context before a silent pause? Moreover, regarding the relationship between prominences and boundaries we should also compare conditions Read and Listen, but in the case of pauses we inspect only condition Listen. In section 6.6.1 we account for the relationship between prominence and boundaries and in 6.6.2 for the relationship between prominence and pauses.

6.6.1 Boundary Marking Context of the Prominence Annotations

In order to investigate the relationship between prominence and boundaries we have computed precision and recall for the prominent words and the boundary annotations, i.e. we check on one hand how great a proportion of the words annotated as prominent which is located to the context before a boundary annotation (Precision), and on the other hand how often boundary annotations have a preceding word annotated as prominent (Recall). For a closer description of the formula for Precision – Recall, see chapter 5, section 5.5.

Figure 6.16 offers a picture of the relationship between prominence markings and boundary annotations expressed through precision and recall for all four speaking styles in both conditions Read and Listen. The scripted monologue is represented by black diamonds, the non-scripted monologue by dark grey squares, the scripted dialogue by grey triangles and the non-scripted dialogue by light grey circles. The arrows point from condition Read to condition Listen.

Figure 6.16 shows that in general the overlap between prominence annotations and boundary annotations is slight, and the difference between conditions Read and Listen is rather slight. Lets us start with examining condition Read, i.e. the symbols away from which the arrows are pointing. In scripted monologue the precision is low (22%) meaning that the proportion of prominence annotations that coincide with a boundary annotation is low. The recall is slightly higher (32%), meaning that the proportion of boundary

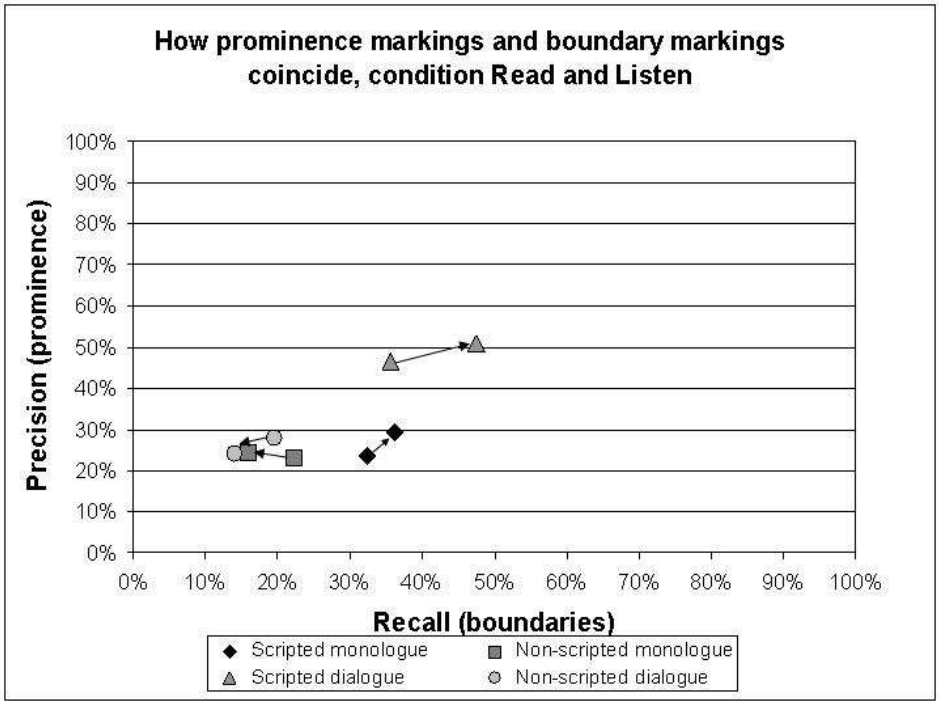

Figure 6.16: The relationship between prominence and boundary annotations in conditions Read and Listen.

annotations that occur in the context after a boundary annotation is slightly greater. In the non-scripted monologue the precision is equally low (21%), and in addition also the precision (22%). This indicates a small overlap between prominence annotations and boundary annotations. Scripted dialogue has the highest precision (46%) and in addition also a fairly high recall (37%). Thus, the greatest overlap between prominence annotations and boundary annotations is found in the scripted dialogue. The non-scripted dialogue lies very closely to non-scripted monologue with a prominence on 29% and a recall on 20%.

Examining condition Listen, at the arrowheads we do in general not find any great differences. However, in the two scripted speaking styles both recall and precision increase, whereas in the two non-scripted styles the recall is reduced. In other words, with access to the speech signal the overlap between prominent words and boundary annotations (meaning a boundary immediately after a prominent word) increases in the scripted speaking styles, while the proportion of prominent words immediately followed by a boundary annotation is reduced in the non-scripted speaking styles.

The generally higher values for the scripted speaking styles might be an effect of a higher proportion of grammatically complete sentences in the scripted speaking styles, while

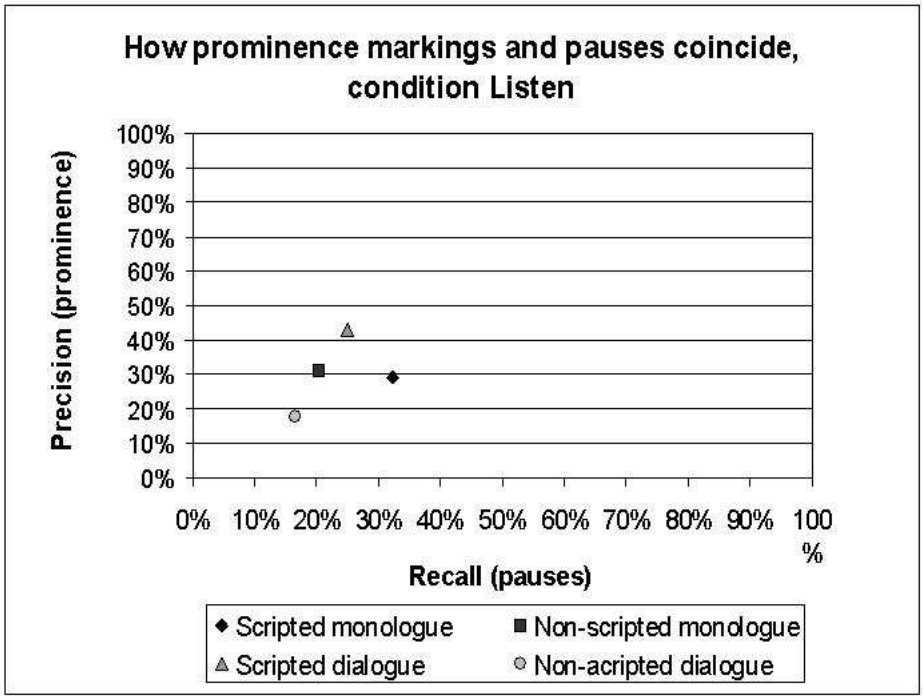

Figure 6.17: The relationship between prominence annotations and pauses in condition Listen.

the difference in condition Listen might indicate that the speech signal accentuated the tendency already present in condition Read. In other words, the scripted sentence structure becomes even more visible with access to prosodic information.

6.6.2 Pause Context of the Prominence Annotations

In the previous section we studied the relationship between words and boundaries annotated as prominent, although a boundary is not equivalent to a pause. Therefore we also examine the relationship between prominences and pauses. Also in this case we account for the relationship with Precision and Recall, precision here meaning the proportion of the prominent words located to the context before a pause and recall meaning the proportion of pauses located to the context after a prominence.

In figure 6.17 the relationship between prominence and pauses is shown.

Figure 6.17 reveals great similarities to the precision – recall relationships between prominences and pauses, but there are some differences. In the case of scripted monologue the picture is very similar to the prominence – boundary relationships in figure 6.16. Also the non-scripted monologue and the non-scripted dialogue are very similar. However, the

scripted dialogue has a substantially poorer recall in the relationship between prominences and pauses compared with prominences and boundaries, thus, the prominence was followed by a pause in many fewer cases than followed by a boundary.

To sum up the relationship between prominence and boundaries and prominence and pauses: it seems that a relationship between prominent words and boundaries or pauses is more clearly present in the scripted speaking styles than in the non-scripted ones. Perhaps this might be connected to information structure issues in scripted versus non-scripted discourse. However, since we have not made a deep analysis of the information structure, it is not possible to draw any conclusions about this, Nevertheless, we shall return to this issue in the general discussion in chapter 8.

6.7 Discussion of the Prominence Annotation Task

In the study of prominence, the picture turned out to differ greatly in a number of dimensions compared with the boundary annotations. The general discussion of these differences is left to chapter 8, however, some features will still be mentioned here since they are too striking to leave out.

The inter-annotator agreement within conditions was low, indicating a rather high level of disagreement between subjects where to annotate prominence. A study of the relationship between the prominence sites and the majority annotations in table 6.1 shows that there is a great difference, and the majority sites constitute about a quarter or a fifth of the total of prominence sites. This relationships is also clear in the inter-annotator agreement figures. Thus, the level of disagreement in the prominence marking task was much higher than in the boundary marking task.

We can think of a number of reasons for this low level of agreement. Firstly, the task was very unfamiliar. Subjects feel rather confident in inserting punctuation since this is something very similar to what they do when they write. However, inserting prominence is not a very familiar task, and this might have given rise to some of the problems.

The problem with unfamiliarity with the task might have been minimised with more detailed guidelines, but then the annotations would also have been more directed towards a certain kind of prominence. Our aim was to explore prominence in discourse, and therefore this was not the way we wanted to go. In fact, we were a bit worried that our instructions indicating that prominence often is signalled by acoustic focus would be too leading. However, when the results were examined, this turned out not to be the case.

In the instructions the subjects were briefed to mark one or more words as prominent, and at first we thought that the low level of agreement could be due to some subjects preferring to mark just the head of a phrase, while other subjects marked e.g. the whole phrase. Thus, we inspected a subset of the corpus and cancelled all sequences, leaving just the head word in the sequence marked as prominent. This did not substantially

influence the inter-annotator agreement, which is why we decided to skip this editing of the data and use the majority set of the original annotations.

In contrast to the boundary annotation task, the level of agreement across conditions was also low. This means that while the majority annotations in condition Read and condition Listen to a large extent pointed out the same positions in condition Read and condition Listen in the boundary annotation task, the relationship was the opposite in the boundary annotation task. In the boundary annotation task the majority set in condition Read was different to the majority set in condition Listen. This indicates that the speech signal had influenced the subjects in condition Listen to annotate differently from the subjects in condition Read.

Concerning the investigation of the linguistic features of the words annotated as prominent there were, in contrast to the boundary marking task, only marginal differences across speaking styles and moreover, the pattern did not change between the phrasal and the part of speech features. This means that all the prominent words came from the same linguistic set of candidates irrespective of speaking style. These results are in line with results reported by e.g. Gårding (1967).

We now turn to the prosodic features and the acoustic correlates in the speaking styles. We would like to stress that the acoustic measures in this study are crude, but the findings in our data are in line with what other researchers have reported for Swedish. The prominent words have on average a higher F0 than the average F0 for the speaking style in general. The differences across speaking styles seem to be related to both the dimension scripted – non-scripted and the dimension monologue – dialogue. We start by discussing the three speaking styles where the speakers are constant, and then relate our discussion to the scripted dialogue.

The prominent words in the scripted monologue have on average both a higher F0, a higher standard deviation of F0, and are of a shorter duration and a higher intensity than the non-scripted monologue. Thus, the scripted monologue is faster, louder, and with a higher pitch. This can be due to the fact that in the scripted monologue the text is present and the speaker can thus rush through the text with a higher tempo and stronger intensity, which might also cause a rise in the F0. We thus believe that to a large extent this difference is related to physiological features like subglottal pressure. Since the message is present in advance and the speaker can plan where to breathe, the subglottal pressure can be kept high all through the message. In the non-scripted monologue the speaker has to plan and is thus not able to let the breathing be subordinated to the speaking in the same way.

In the non-scripted dialogue the F0 of the prominent words is higher than in the monologues, especially by the female speakers. In addition there is a higher standard deviation of F0 in the prominent words, a lower intensity and a difference in duration between scripted and non-scripted monologue. Thus, the dialogue is faster than the non-scripted monologue and in addition more quiet with a higher and more varied pitch than both monologues. The rise in tempo might come from the interactive aspect, i.e. the speaker

does not have all the time to himself but has to speak in interaction with another speaker. Also the higher and more varied pitch might be a result of the interaction. However, since we have not studied the interaction in detail, we cannot conclude anything about this.

Relating the acoustic features of the prominent words in the scripted and non-scripted monologues and the non-scripted dialogue to the scripted dialogue, it immediately becomes clear that the duration of the prominent words in the scripted dialogue is far longer than in any other speaking style. There is not much point in comparing the pitch value to the other styles, since the speakers are not the same ones in the scripted dialogue, however, it is appropriate to note that the speakers in the scripted dialogue varied their voices to a much greater extent than the other speakers, with e.g. whispering, hissing and screaming etc. The standard divergence of F0 is also comparably high, indicating a wide variation in the pitch.

Our hypothesis about the scripted dialogue which is acted, compared with the three other speaking styles, is that to a large extent the scripted dialogue has used acoustic features in order to make speech as clear as possible. Thus, the extremely long duration and the highly varied pitch are two such means. In addition, in the scripted dialogue a high proportion of the pauses correlated to boundaries. We will leave this issue for now, but we shall return to it and elaborate on it in the general discussion in chapter 8.

Regarding the focal accent, in general it seems that many words which are not focal but still annotated as prominent are emphasised by means of their connection to other words. These words are in turn acoustically emphasised and form, together with the non-focal word, e.g. a contrast pair or a multi-word unit. In other cases, non-focal words seem to achieve prominence because of their position in specific structural contexts, such as e.g. lists or headings. Thus, the prominence of a closely related word, or a structural context, seems to have the power to colour a less emphasised word with prominence.

The annotation of non-focal words as prominent is more frequent in the scripted conditions, and in particular these non-focal prominent words are spread over a wider range of structural features on the text level (lists, headers) than in the non-scripted styles. This can indicate that a more text-like sentence structure allows a wider range of features to give words prominence in form of e.g. more elaborate syntactic constructions.

In the monologues, a greater proportion of non-focal prominent words classified as connected to another prominent word are found compared with the dialogues. If the prominence of non-focal words is related to the information structure, then there might be differences in the distribution of information structure features across different types of discourses in the same way as there are differences in lexicogrammatical patterns captured by e.g. phrase properties, part of speech distribution and measures like NQ.

To conclude the study of prominence we can state that in many cases it turned out to show features complementary to the study of boundaries.

We now proceed to the last study in this work; the study of intentions in the form of questions.

Chapter 7

Questions, an Example of Segment Intention

In the previous two chapters we have studied the interaction between the string of words and the prosody in relation to the discourse segments. In many discourse theoretical frameworks, e.g. (Grosz and Sidner, 1986), as well as in dialogue coding, e.g. (Carletta *et al.*, 1997) these segments are considered to be motivated by a certain intention, for instance the intention to ask for information or to give information. If the intention is to seek information the segment prototypically has the form of a question.

7.1 Introduction to the Study of Questions

How is the string of words on one hand and the prosody on the other hand influencing the subjects' classification of segments with regard to a specific intention? To investigate this we have carried out a more detailed study of the subjects' annotations of question marks. The question mark conventionally expresses the intention to request information. The question is thus the intention with which we work, and the section of discourse ending with a question mark is the unit we use for the analysis.

The materials in this study are the same as those that were used in the studies of boundaries and prominence. Thus we have the same four speaking styles (scripted and non-scripted monologue and dialogue). The question annotations is a subset of the annotations in the boundary marking task and are thus annotated either solely on the basis of the string of words (condition Read) or on the basis of both the string of words and the speech signal (condition Listen). As it was for the studies of boundaries and prominence, our aim is to investigate to what extent subjects use the string of words on one hand and the speech signal on the other when annotating segments as questions.

There are a number of reasons for choosing questions as the object for this study of segment intention. One reason is that questions are unambiguously signalled by the subjects with a graphic symbol – the question mark – in the annotations, which makes the positions easy to find. Even though the questions are clearly signalled by the punctuation, they come in different syntactic forms. In this study we distinguish between verb initial questions, question word questions and questions with a declarative form. In addition questions are accompanied by intonational features. A specific question intonation contour does not seem to be present, see e.g. Bolinger (1989), Bruce (1998),House (2003), however, examples of prosodic cues to questions are given by e.g. Bolinger (1989).

Thus, questions are signalled by many means, but what influences a subject towards classifying a segment as a question or not? Examining the subjects' annotations of questions across conditions Read and Listen will give us a picture of the distribution of labour between the syntax and the prosody in cueing subjects as to what might be a question and what might not.

Since we use the subjects' annotations we get an opportunity to study subjects' agreement regarding both the general annotations in the different speaking styles and the subjects' agreement concerning different question types. Thus, do subjects agree more with regard to one form of questions than to another, or in a particular speaking style?

The present study of questions focuses primarily on similarities and differences across conditions Read and Listen, and does not specifically deal with phrase or part-of-speech frequencies or acoustic measurements.

The rest of the chapter has the following structure: in section 7.2 the experiment procedure is described in more detail, and in section 7.3 the results are presented. Section 7.4 shows the distribution of question types, and lastly the results are discussed in section 7.6.

7.2 Method for Investigating the Question Segments

In order to account for the relative variations between subjects' agreement and different types of questions, the data was divided into three sets, and in addition the question segments were classified according to three different types. In section 7.2.1 the sets of data are described and in section 7.2.2 the question types.

7.2.1 Classification of Data

In the study we have used only the annotations of question marks in the data, i.e. all other kinds of boundary annotations were discarded in the analysis, i.e. equivalent to "no marking". To enable us to study how the distribution of question annotations

varied across different levels of annotator agreement, three different sets of data were constructed:

- All question annotations, "set 1".
- All positions where at least 2 subjects agreed on a question mark, "set 2".
- All positions where the majority agreed on a question mark, "set 3".

Set 1 (all annotations) and set 3 (majority annotations) are parallel to the sets used in the studies of boundaries and prominence, but set 2 (at least 2 subjects) is new for this study. The reason for extending the analysis with set 2 is that we have fewer data in the question annotation task than in the full boundary annotation task, and extending the materials with this set of data allows us to i) use more of the data and ii) more closely follow tendencies from a low level of agreement to a majority agreement.

Set 1 (all annotations) is used for the general measurement of inter-annotator agreement, which shows how well the annotators in general agreed in the annotation of questions. In this case each annotation stands for one subject's annotation, and agreement is computed within conditions Read and Listen.

Set 1 is, however, also used in a normalized form where each position where a subject has annotated a question mark is taken as a question position. Thus, this normalized set 1 shows the total number of question positions marked by at least one subject.

Set 2 (at least 2 subjects) was constructed so that we could make use of as much data as possible. Set 2 is normalized, meaning that a position where at least two subjects have annotated a question mark is considered a question position. Since these data are normalized to question positions (and not question annotations) it is possible to compute the level of inter-annotator agreement across conditions Read and Listen based on this set.

Set 3 (majority annotations) is used in parallel with set two in order to show the extent to which the majority annotations differed from the annotations made by at least 2 subjects. Set 3 is normalized, and the positions where the majority have annotated a question mark are considered question positions. Thus, it is possible to compute the level of inter-annotator agreement across conditions Read and Listen also in this set.

Using these three sets of data we can study the general level of agreement in the annotation of question marks (set 1), the distribution of question positions annotated by at least 2 subjects in the different speaking styles (set 2) and the distribution of question positions on which the majority of the annotators agreed (set 3). In sets 2 and 3 it is possible to compute inter-annotator agreement across conditions.

7.2.2 Question Types

In order to study the distribution of different question types, the questions were classified into three categories according to question form:

- Verb initial questions: Tog du den? (Took you it?, *Did you take it?*)
- Question word questions: Vad tog du? (What took you?, *What did you take?*)
- Declarative questions: Du tog den? (You took it?, *You took it?*)

In addition the questions with a declarative form were divided into two classes: 1) declarative questions containing an adverbial or discourse particle ("nog", "väl", "då" (eng. "probably", "I suppose", "then")) and 2) declarative questions not containing such adverbial or discourse particles. In the examples below questions without (example 7.1) and with (example 7.2) sentence adverbials are shown.

(7.1) Du tog den? (You took it?, *You took it?*)

(7.2) Du tog väl den? (You took probably it?, *You took it, I suppose?*)

Thus, in both conditions Read or Listen we have three different sets of data, set 1 (at least 1 subject), set 2 (at least 2 subjects) and set 3 (majority), and in addition the questions are classified into three categories where the questions of declarative form contain two sub-categories.

7.3 Inter-Annotator Agreement for the Annotation of Questions

The topic for this section is the inter-annotator agreement for the annotation of questions in the four speaking styles within and across conditions Read and Listen. In the same way as in the studies of boundaries and prominence, the level of inter-annotator agreement is expressed using κ-values. However, in order to secure a fuller picture of these figures we start by examining the number of question positions in the different speaking styles. Our initial overview of question positions is based on the normalized set 1, i.e. the number of positions where at least one subject has annotated a question mark. The distribution is shown in figure 7.1.

Considering that figure 7.1 shows the maximum number of question positions, i.e. including positions marked by only one subject, the data is very sparse in the monologues

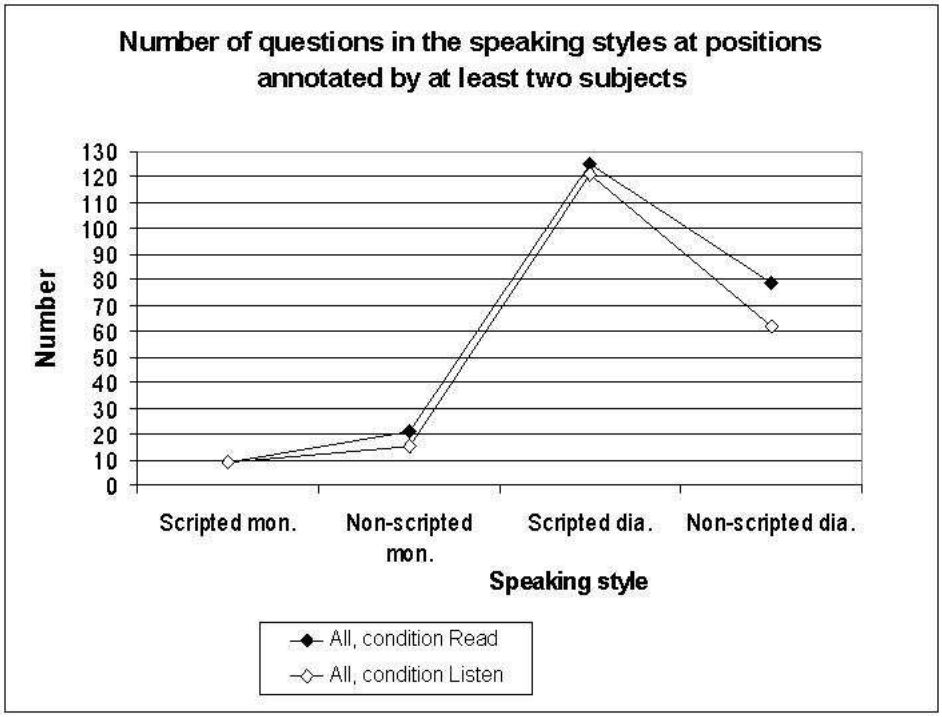

Figure 7.1: Number of all the question positions in the speaking styles.

(10-20 positions), but more frequent in the dialogues (60-120 positions). The annotation of questions is therefore very unevenly distributed over the four different speaking styles, clearly mirroring the difference between the more narrative monologues and the goal oriented dialogues. Furthermore, there are slightly fewer positions in condition Listen, but the difference is obvious only in the non-scripted dialogues.

In the rest of this section on inter-annotator agreement, we present the figures for both the monologues and the dialogues, even though the data for the monologues are very sparse. However, in the subsequent analyses of question types, the monologues are discarded and only the dialogues, where we have more data, are considered.

7.3.1 Inter-Annotator Agreement Within Conditions for the Annotation of Questions

Turning to the figures for inter-annotator agreement we ask: how well do the subjects agree on the annotation of question marks within conditions Read and Listen? To obtain a picture of this, the general level of inter-annotator agreement for the subjects' annotations of questions was computed using the κ-statistics. The κ-statistics are based

on set 1 (each subject's annotations of a question mark), i.e. we compute the level of agreement between the subjects for the question positions shown in figure 7.1. The result is shown in table 7.1.

	Scripted Monologue	Non-scripted monologue	Scripted Dialogue	Non-scripted dialogue
Condition Read	0.82	0.56	0.73	0.63
Condition Listen	0.75	0.58	0.74	0.69

Table 7.1: Inter-annotator agreement for question annotation within conditions Read and Listen.

The level of inter-annotator agreement in the scripted monologue is very high, especially in condition Read, while the non-scripted monologue has a substantially lower level of agreement with similar figures for both conditions Read and Listen. However, please note that the actual number of question mark annotations in the monologues was very low, which means that these figures are less reliable.

In the dialogues the scripted dialogue has similar κ-values in condition Read (κ=0.73) and Listen (κ=0.74), while the non-scripted dialogue ranks lower with κ=0.63 in condition Read and κ=0.69 in condition Listen.

The difference in the level of inter-annotator agreement between conditions Read and Listen is very slight in the scripted dialogue but a little higher in the non-scripted one. Thus it is possible that the extra information in the speech signal has influenced the subjects more in the non-scripted dialogue than in the scripted one. In the scripted monologue we find a lower level of agreement in condition Listen than in condition Read, but the data is too sparse to hazard any general remark.

Relating the inter-annotator agreement in table 7.1 to the number of positions shown in figure 7.1, we find that a lower number of positions in condition Listen is accompanied by a rise in inter-annotator agreement in condition Listen (in all the speaking styles except the scripted monologue). In the non-scripted monologue and the scripted dialogue the differences are, however, slight to non-existent, but in the non-scripted dialogue the difference is clearer. This means that at least in the non-scripted dialogue the access to the speech signal might have influenced the subjects towards annotating fewer question positions, but the level of agreement was higher.

7.3.2 Inter-Annotator Agreement Across Conditions for the Annotation of Questions

How well does the annotation of questions agree across conditions Read and Listen? Do subjects annotate about the same set of positions as questions in both conditions Read and Listen, or do the question positions differ across conditions? In order to investigate

this, the level of inter-annotator agreement was computed between the set of questions annotated in condition Read and the set of questions annotated in condition Listen. For the measuring of inter-annotator agreement we use set 2 (at least 2 subjects) and set 3 (majority annotations), and these two sets are also used throughout the rest of the question study. The method for computing the level of inter-annotator agreement across conditions is described in chapter 5, section 5.2.2.

A high level of inter-annotator agreement across conditions indicates that subjects annotate the same sets of positions as questions in both condition Read and condition Listen. The lower the level of agreement, the more the sets of question positions in the two conditions differ. A high level of inter-annotator agreement in set 2 means that there is an agreement between condition Read and Listen in general. A high level of inter-annotator agreement for set 3 means that at positions where a majority of subjects agreed in condition Read, the majority of subjects also agreed in condition Listen. If there is a higher level of agreement in set 2 than in set 3, this means that the majority annotations diverge more across conditions than the annotations made by fewer subjects. If the opposite is the case, it means that the fewer the subjects that agree inside conditions, the lower the level of agreement for these positions across conditions.

Table 7.2 shows the inter-annotator agreement across condition Read and Listen expressed with κ-values for both set 2 (at least 2 subjects) and set 3 (majority).

	Scripted Monologue	Non-scripted monologue	Scripted Dialogue	Non-scripted dialogue
At least 2 subjects (set 2)	0.93	0.91	0.77	0.73
Majority (set 3)	0.86	0.66	0.90	0.94

Table 7.2: Comparison of inter-annotator agreement between conditions Read and Listen, questions.

In the monologues, there are higher κ-values in set two than in set three, i.e. using set 2 (at least 2 subjects), we find a high degree of overlap between conditions Read and Listen concerning positions annotated as questions. However, examining set 3 (majority) the overlap between condition Read and Listen decreased. This means that there is a generally high agreement between condition Read and Listen concerning where the question positions are located, while the overlap declines when the data is reduced to the positions where the majority of subjects agree. However, the differences in the monologues are based on a small set of data (about 7 positions in the scripted monologue and 7-12 in the non-scripted monologue), which means that very small differences in the subjects' annotations become very visible in the κ-values. In other words, since the data is so sparse, the figures are not fully reliable.

Turning to the dialogues where we have more data, we find that the result is the opposite. The κ values are higher in set 3 (majority) than in set 2 (more than 2 subjects). This means that there is a core set of question positions on which the subjects agree across

conditions Read and Listen. When annotations made by a minority of the subjects are filtered away, the κ values increase. Thus, where subjects agree inside conditions (majority) is also where subjects agree across conditions (high κ-values). This indicates that the access to the speech signal does not greatly influence the subjects in general towards annotating differently in condition Listen than in condition Read. However, since the level of agreement is lower in set 2, single subjects might be more influenced in single positions.

To sum up, the level of inter-annotator agreement regarding the question positions is rather high, and the prosody does not seem to influence the annotation of question positions to any great extent. However, this holds for questions in general but does not hold for specific question types, the topic for next section.

7.4 Distribution of Question Types in the Different Speaking Styles

In this section we make a more detailed study of the question types. In this proportional analysis of question types we restrict the materials to the dialogues where the data is less sparse. The analysis of question types is based on set 2 (at least 2 subjects) and set 3 (majority) in both conditions Read and Listen. However, before presenting the proportions of question types we examine the number of question positions in set 2 (at least 2 subjects) and set 3 (majority) in all four speaking styles.

In figure 7.2 the number of question positions in the set 2 (squares) and the set 3 (triangles) are shown for both conditions Read (filled symbols) and Listen (empty symbols). As a comparison the number of all question positions in set 1 (diamonds, dotted line) is recapitulated from figure 7.1.

Not very surprisingly, the number of question positions decreases from set 1 (at least 1 subject) to set 3 (majority), i.e. there are more question positions in set 1 (all) than in set 3 (majority). In the scripted monologue the difference between the different sets of data is in general slight, but there is also a very low total number of question positions in this speaking style. The same holds for the non-scripted monologue.

The scripted dialogue has the highest number of question positions as well as the greatest difference between conditions Read and Listen (in set 2, squares) with about 81 positions in condition Read and nearly 100 in condition Listen. In the non-scripted dialogue the positions in each of the three sets are fewer than in the corresponding sets in the scripted dialogue. Furthermore, the difference between conditions Read and Listen which is present in set 1 (diamonds) has vanished where there is a higher level of annotator agreement in set 2 (squares) and 3 (triangles).

In general the difference in the number of question positions between conditions Read and Listen is minimal. How then are the question types distributed over the sets of

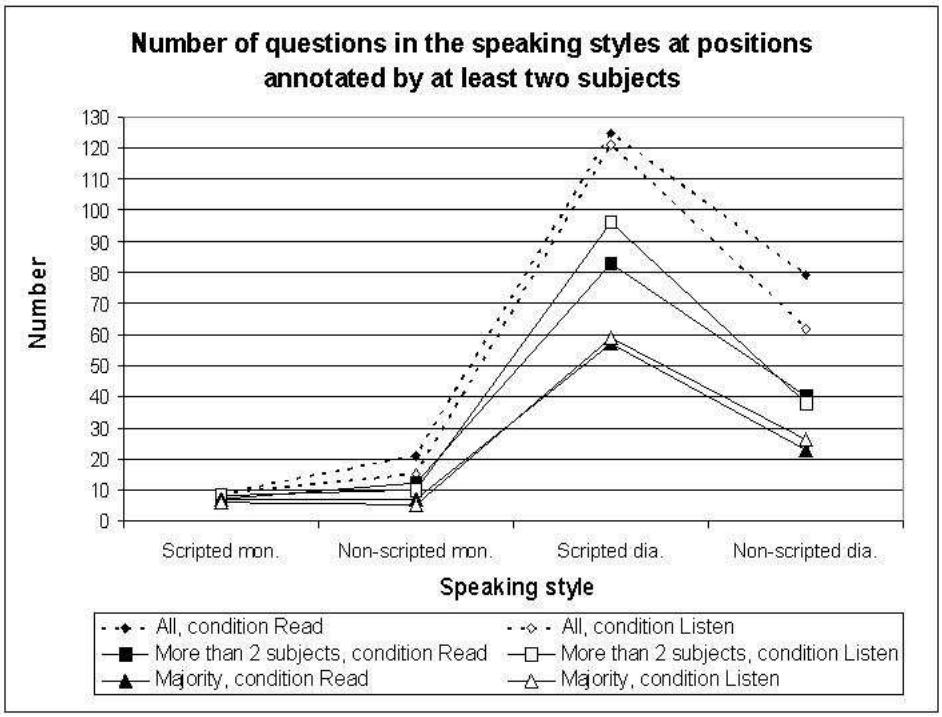

Figure 7.2: Number of questions in each speaking style, conditions Read and Listen.

data? Are they evenly distributed in all the cases, or is there a predominance of one type of questions in e.g. set 3 (majority) compared with set 2? In addition, are the annotations of one specific question type influenced more by the subjects having access to the speech signal than the annotations of another question type? We examine these questions separately for both types of dialogue, starting with the scripted dialogue.

7.4.1 Question Types in the Scripted Dialogue

The scripted dialogue is the speaking style with the highest number of question positions. In figure 7.3 the proportions of the three question types are shown for the scripted dialogue for both set 2 (at least 2 subjects, diamonds) and set 3 (majority, squares) in both conditions Read (filled symbols) and Listen (empty symbols).

Upon examining the three question types in figure 7.3, the distribution of the three question types in set 2 (diamonds) is quite even. Verb initial questions constitute about 40%, question word questions about 29% and declarative questions about 31%. There is no difference between condition Read and Listen.

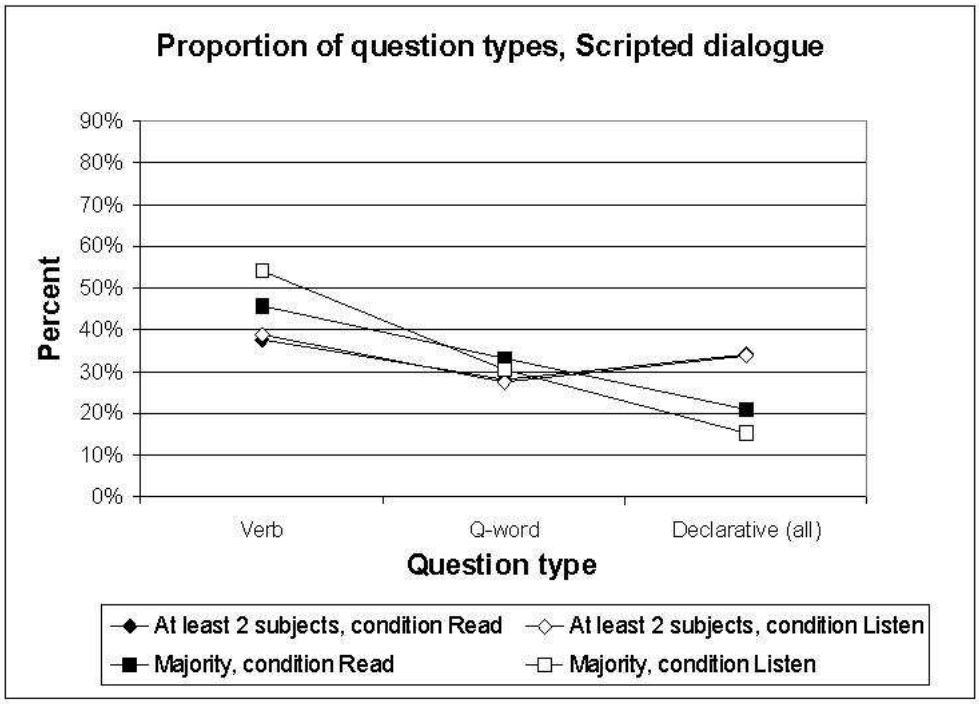

Figure 7.3: Proportion of question types in scripted dialogue.

The majority set (squares) shows a slightly different pattern. The proportion of the verb initial question has increased (45% in condition Read and 55% in condition Listen) while the proportion of the question word questions is about the same compared with set 2 (30%). The proportion of verb initial questions has, however, declined (20% in condition Read and 15% in condition Listen). In set 3 there is a greater difference between condition Read and condition Listen than was found in set 2, and the trend with a greater proportion of verb initial questions and a smaller proportion of questions of a declarative form is clearer in condition Listen as compared with condition Read. Thus, the positions where subjects agree differ slightly between condition Read and condition Listen, condition Read to a greater extent supporting questions of a declarative form.

The differences between set 2 (at least 2 subjects) and set 3 (majority) indicates that there is a lower level of agreement between the subjects about the questions with declarative forms since this category decreases in the majority set. In addition, the proportion of verb initial questions increases in condition Listen.

The results from the scripted dialogue thus indicate a higher and more stable level of agreement for questions which have some syntactic or lexical features that signal the question mode. In addition, the results indicate that access to the speech signal does

not greatly influence the subjects' annotations. Bearing these observations in mind we now turn to the non-scripted dialogue.

7.4.2 Question Types in the Non-scripted Dialogue

For the non-scripted dialogue, please bear in mind that the data is sparser than in the scripted one. The proportions of the question types in the non-scripted dialogue are presented in figure 7.4 along the same principles as for the scripted dialogue. Set 2 (at least 2 subjects) is represented by diamonds and set 3 (majority) by squares. Condition Read is shown with filled symbols, and Listen with empty symbols.

An examination of the distribution in the non-scripted dialogue in figure 7.4 clearly shows that the proportions of the question types differ from those in the scripted dialogue, the non-scripted dialogue having in general a higher proportion of questions with a declarative form than the scripted dialogue does.

In set 2 (at least 2 subjects) the proportions of both verb initial questions (about 14%) and question word questions (about 17%) are low, while the proportion of declarative questions is fairly high (about 69%). There is hardly any difference between condition Read and condition Listen.

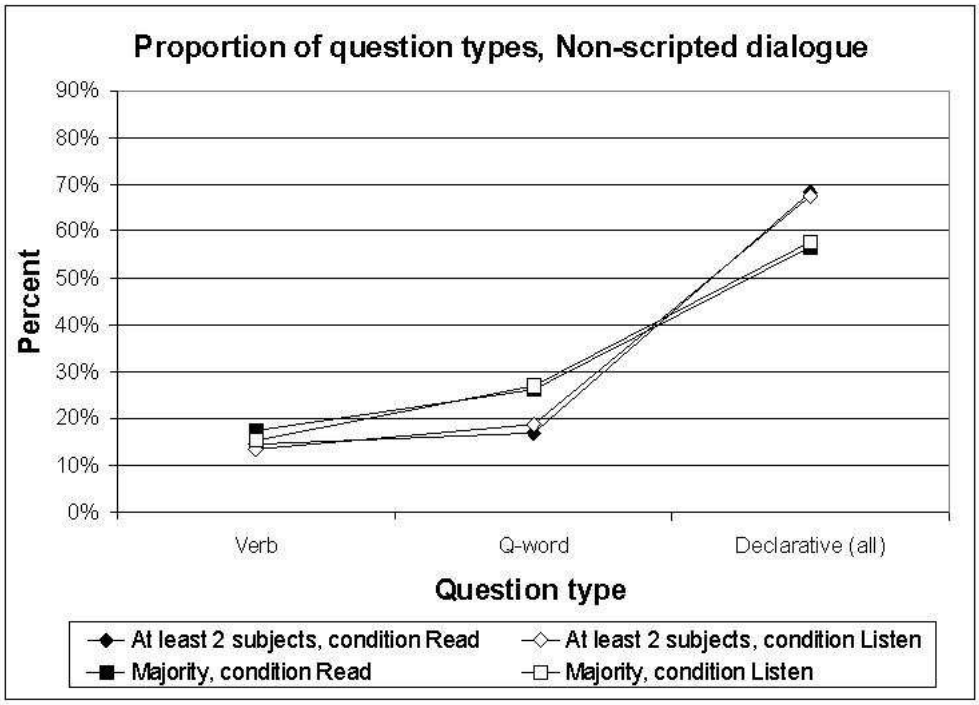

Figure 7.4: Proportion of question types in non-scripted dialogue.

Set 3 (majority) shows the same trend with a high proportion of declarative questions (54%) and lower proportions of verb initial questions (18%) and question word questions (28%). Also in the majority set the differences between condition Read and condition Listen are marginal.

Comparing the scripted and the non-scripted dialogues we see that the distribution of the three question types differs between the speaking styles, the non-scripted dialogue containing a larger proportion of declarative questions. However, the proportion in this question type is declining in set 3 (majority) compared with set 2 (at least 2 subjects) in both the scripted and the non-scripted dialogues, indicating that subjects disagree more about this question type than about the other two types.

7.4.3 Differencs by Questions of Declarative Form

The questions of declarative form do not form a homogeneous category. Some of these questions are of a purely declarative form, while others contain an adverbial or discourse particle which modifies the declarative question and indicates its status as a question. In figure 7.5, the proportion of the questions not containing such an adverbial is shown across the scripted and non-scripted dialogues as well as across the 2 sets, (at least 2 subjects and majority), and the figure indicates that the vaguer type decreases as the level of agreement rises.

Figure 7.5 shows each set in both condition Read and Listen, with bars for both scripted (light bar) and non-scripted dialogues (dark bar).

In set 2, condition Read (the pair of bars on the very left), in the scripted dialogue there is a substantial proportion of declarative questions not containing an adverbial, most of the declarative questions lacking such adverbials. In the non-scripted dialogue the proportion is smaller than in the scripted one but still high in comparison with the other sets.

In set 2, condition Listen (the pair of bars second from left) the proportion of the questions containing an adverbial has decreased. The proportion is still higher in the scripted dialogue, but in both scripted and non-scripted dialogue it is proportionally lower than in condition Read. Thus, it seems that the speech signal in both scripted and non-scripted dialogue has influenced the subjects towards sorting out questions of the vaguer type ("pure" declarative questions) and towards favouring declarative questions containing adverbials.

In set 3, condition Read (the pair of bars third from left), the proportion of questions without adverbials again decreases in the scripted dialogue compared with set 2. Moreover, in the scripted dialogue the proportion of questions without adverbials again decreases in condition Listen in set 3 compared with condition Read. In other words, the lower the proportion of purely declarative questions, the higher the level of agreement between subjects. This steady decrease in the proportion of questions without adverbials is not

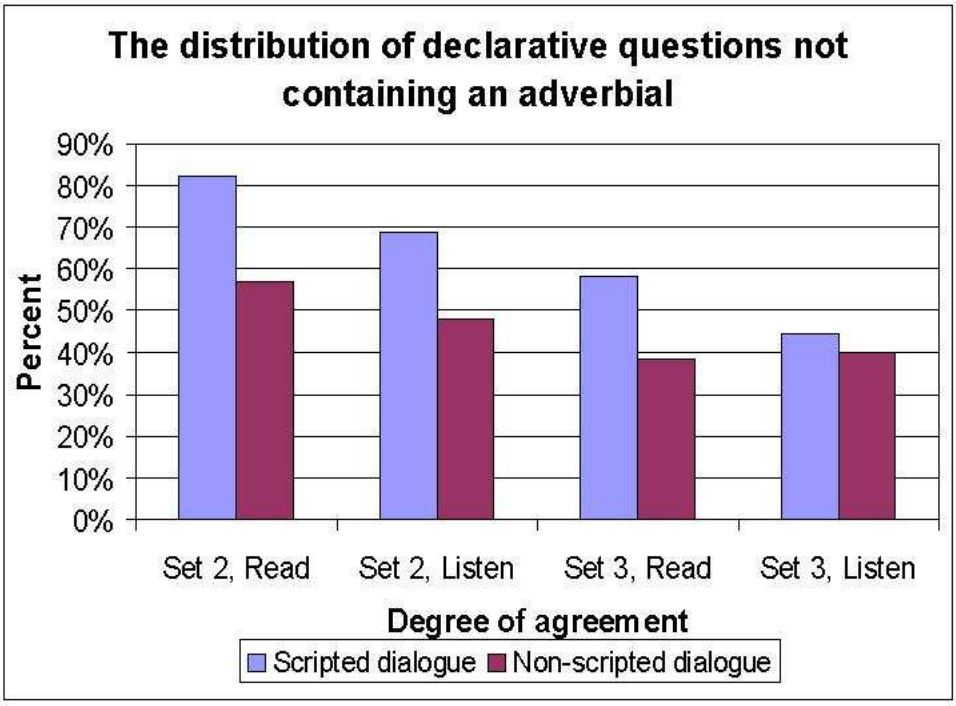

Figure 7.5: Proportion of declarative questions containing adverbials.

present in the non-scripted dialogue. However, in the non-scripted dialogue the initial proportion of the declarative questions containing adverbials was higher, and in set 3, condition Listen (the pair of bars on the very right), the proportion in scripted and non-scripted dialogue is about the same.

So, in scripted dialogue, there is a tendency that an increase in the level of annotator agreement (set 3, majority, compared with set 2) goes together with a decrease of the questions with a vaguer surface form (i.e. pure declaratives) and access to the speech signal (condition Listen). In non-scripted dialogue this tendency is present only within set 2. However, it should be noted that the data is sparse.

7.5 Points of Disagreement Between Conditions Read and Listen in the Annotation of Questions

In the scripted dialogue, the proportion of verb initial questions increases in condition Listen. Actually, in absolute numbers there is also a higher number of verb initial questions in condition Listen than in condition Read. In the scripted dialogue, something in

the speech signal made the subjects more prone to annotate an utterance with the syntactic form of a verb initial question as a question in condition Listen than in condition Read. In addition the proportion of questions of a declarative form tended to decrease in condition Listen, thus, also here the speech signal seems to have some influence on the subjects' annotations of questions. In order to find out more specifically what might have influenced the subjects' annotations, we examined all positions of disagreement between conditions Read and Listen, i.e. where a majority in condition Read had annotated a question while only a minority or no-one in condition Listen had done this, and vice versa. Since the level of agreement between condition Read and condition Listen was high, the data was very sparse, but the findings were still interesting.

In the scripted dialogue (set 3, majority) there were no verb initial questions in condition Read which were not marked in condition Listen. In condition Listen, there were 4 questions which were not marked as questions in condition Read.

In general, having access to the speech signal seemed to cue the subjects to segment the string of words differently from the subjects who had access only to the transcripts. In most cases this resulted in shorter segments and extra inserted questions in condition Listen, while the annotators in condition Read instead used a maximum reading of the segment. An example of such a passage for a question word question is shown in 7.3.

(7.3) hur vet ni det [1] att ni inte är överflödig [2]
How do you know that [1] that you not are superfluous [2]

The point of disagreement between condition Read and condition Listen lies at [1]; in condition Listen a question mark is annotated here by the majority of subjects but not in condition Read. In both condition Read and Listen, however, a question mark is annotated at position 2. There is a pause at position [1] (0.680 sec) and at position [2] (0.2 sec). It is clearly possible to regard the string of words in the example as one single question, but also as two separate questions. In condition Read, where the subjects had access to only the string of words, the majority of subjects chose to interpret the string of words as one question. In condition Listen, where the subjects had access to the speech signal, they chose instead to interpret the string of words as two questions. In this particular case, the "det" in front of [1] is annotated as prominent in condition Listen, and the F0 ends on a high boundary tone before the pause. However, in most other cases such a F0-rise is not present, only a pause. We take this to support the view that the pause alone influenced the subjects in condition Listen towards splitting the string of words into two questions.

Another example is shown in 7.4.

(7.4) menar ni nyss [1] så sa jag endast ja [2]
mean you just a moment ago [1] then said I only yes [2]
If you mean just a moment ago I said only yes

In example 7.4 the disagreement is about position [1]; in condition Listen the majority of subjects have annotated a question mark, while this is not the case in condition Read. Also in this example it is possible to regard the verb initial conditional as one single unit or as two units. At position [1] there is a pause (1.110 sec), and the word "nyss" is annotated as prominent by all subjects in both conditions Read and Listen. At position [2] there is also a pause, but no question mark was annotated in either condition Read or Listen. However, in both conditions this is marked as a boundary position. In the example, position [1] is annotated as a question in condition Listen but not in condition Read, thus the string of words is interpreted as containing a question in condition Listen, but not in condition Read. The only point of difference is the pause. If the string of words is regarded as two units, the first one gives a strong impression of a question because it is a verb initial. If the string of words is interpreted as one, the impression of question disappears since the second part of the segment does not continue to support the question form. This type of difference in interpretation of segment length is found in all cases where we find verb-initial questions and question word questions marked in condition Listen but not in condition Read. In the case of declaratives the picture is, however, slightly different.

In contrast to the verb initial questions, the proportion of the questions of declarative form was in general lower in condition Listen than in condition Read. Thus, in the first category the speech signal helped to increase the number, while it helped to decrease the number in the latter category. We have seen that the additional information about the segmenting might have influenced the first category, but what might have influenced the second one?

In the case of declarative form questions, the differences between the majority sets in conditions Read and Listen were very slight, and thus it was not possible to use those on their own. Instead we investigated whether the phenomena found in these very small sets (3 occurrences each in conditions Read and Listen) were supported in the sets consisting of annotations by at least 2 subjects (set 2).

Inspecting what kind of declarative questions were present in condition Read, but not in condition Listen, we found a fairly high proportion of feedback expressions of the type "jaa" ("yees"). These were made by subjects not having access to the speech signal classified as questions, but if they had access to acoustic information the question classification disappeared. In addition there were differences in the actual content of the sets of declarative questions in condition Read compared with in condition Listen in the non-scripted dialogue, i.e. even though the proportions are the same across conditions (see figure 7.4) the content were different. An example of a passage from the non-scripted dialogue is shown in 7.5.

(7.5) när du är i höjd med den här bukten/ ja [1] / så längre ut till...
when you are in level with this here bay/ yes [1]/ then longer out to...
When you are level with this bay/ yes/ then a bit further to....

The point of disagreement is at position [1] where the subjects in condition Read have annotated a question mark while those in condition Listen have not. Here we suggest an explanation similar to how we explained the boundary marking in condition Read as being more frequent than the one in condition Listen in chapter 5. In condition Read rather short and silent feedback contributions are more salient in the transcripts than in the speech. In other words, in condition Read where subjects annotate on the basis of the transcripts alone, the reduced contributions are more visible and subjects then tend to treat them more in isolation, and sometimes classify the expressions as e.g. questions. However, in condition Listen such feedback expressions are not as salient, because they revert to their "normal" reduced form, and subjects do not mark them as independent utterances any more. In other words, the acoustic information reduces the salience of the expressions, thus limiting their classification of the segment as a question. To some extent this might be a result of the task.

In the rest of the declarative questions in set 2 (at least 2 subjects) no specific tendencies could be found, largely because of the paucity of data. The remainder of the questions of declarative form that differed between conditions were all very vague, and since we have only few examples annotated by few subjects, it is not possible to determine what influenced this small group of annotators. What is possible to say is that there was variation. In some cases the concept of question is genuinely vague, meaning that some subjects interpret a specific segment as a question while others do not.

7.6 Discussion of the Study of Questions

In general there is a high level of inter-annotator agreement across conditions Read and Listen in the annotation of questions. This indicate that subjects to a large extent make use of the string of words in the annotation. The syntactically and lexically clearly signalled question types are more stable across the two sets of data (set 2 and 3), i.e. the annotators agree more on these clearer forms of questions. The access to the speech signal has given rise to new annotations of questions, mostly segments with a syntactic question form, and also limited the annotations of questions, mostly questions with a less clear question form.

The types of questions are unevenly distributed over the two speaking styles scripted and non-scripted dialogue. There is a higher proportion of questions with a declarative form in the non-scripted dialogue than in the scripted one.

In general there is lower level of agreement in the declarative form questions, i.e. the proportion of questions with a declarative form decreases in set 3 compared with set 2. This indicates that the annotation of the declarative questions is more subjective than the annotations of the other question categories.

Just what is annotated as a question seems to be influenced not only by the immediate form of the clause – syntactic, lexical or acoustic – but also by the context in which

the segment is interpreted. In example 7.3 we saw an example of how a narrow scope introduced one more question, while example 7.4 shows how a wider scope cancelled the question form of an included segment. By analogy, feedback expressions were annotated as questions in cases where we could assume that they make a more salient and independent impression (i.e. in the transcripts), while this question status disappeared when they became less salient and were integrated into the flow of the dialogue (i.e. with access to the actual speech signal).

In relation to the examples of difference in segment range between condition Read and condition Listen in the scripted dialogue, the wider range – longer segment – was found in condition Read. Thus, without access to the speech signal the subjects annotated the longest match in these cases. As a rule, the longest match preference is used in finite state parsing techniques (Abney, 1992), and this preference was by Abney (1992) also shown to hold for subjects in a study on English. Thus, in cases like this one, where there is a choice between classifying a segment as two shorter ones or one longer one, it is possible the longest match is preferred also in Swedish.

We have not made any more differentiated measurements of the prominent words in the questions, even though the properties of the prominence marking for an item in a segment might well be a factor influencing whether the segment is interpreted as a question or not. However, in our data the main acoustic factor which influenced the subjects annotation was the speech signals' ability to segment the discourse, and thus set a frame for the context of interpretation for the segment. Thus, a long match preference based on the string of words affects not only the interpretations within phrases of phrasal attachments, but might also affect the interpretation of the intention underlying the word sequence.

Even though we could see that the question types are distributed differently across the two speaking styles, scripted and non-scripted dialogue, we are not able to see any differences connected to the impact of the acoustic information in the two speaking styles. For example, in our data there are more cases where a verb initial question or a question word question are annotated only in condition Listen in the scripted dialogue, and similarly there are more cases where a feedback expression is annotated as a question in condition Read in the non-scripted dialogue. However, since the materials are not very comprehensive we cannot say whether these differences really are related to the scriptedness – non-scriptedness.

Chapter 8

Discussion

WHAT is the relationship between discourse type, prosody and segmenting, how do these factors interact, and how can this interaction be expressed within a discourse theory? The object for this chapter is to relate the results from our three studies to each other in order to suggest answers to these questions.

We briefly recapitulate the research questions for this thesis from chapter 2:

- In a spoken language message, what does prosody contribute and what does the string of words contribute to the discourse structure in terms of "boundaries" and "prominent parts" in the the discourse segments?
- Does the interaction between the string of words and the prosody differ across speaking styles?
- How can we model differences between speaking styles including differences in the relationship between the strings of words and the prosody within a discourse theory?

In the previous chapters we have related factors from the string of words e.g. part-of-speech features, and from the prosody, e.g. pauses, to annotations of boundaries, prominences and intentions in the different speaking styles. However, we have not yet touched upon the interaction between the string of words and the prosody across the different speaking styles. In addition, we have not yet discussed how these differences between the speaking styles can be expressed within discourse. To model the interaction between the string of words and the prosody within a discourse theory implies integrating prosody into a discourse theory. This might become a fairly abstract manipulation of features, and therefore we elaborate on what this requires and what it includes.

In our view, integrating prosody into discourse theory means establishing a link between discourse theory and prosody. If discourse theory describes discourse structure, and if

prosody is central to the establishment of the discourse structure in spoken language, then discourse theory should also be able to model contributions of prosody. Otherwise, the discourse theory misses important points of the discourse it aims to describe.

The strategy in our work is to relate features of prosody and features of style to the aspects of boundaries and prominences in the units of discourse. Thus, in our study we relate the prosodic features of phrasing and focal accent to the speaking style properties of phrasal and part-of-speech features at boundary and prominence annotations. These boundary and prominence features are then related to a theoretical description of the units of discourse within the framework of Grosz and Sidner (1986).

The framework of Grosz and Sidner (1986) assumes that each discourse segment consists of three components: linguistic structure, attentional state and intentional structure. Tentatively we have related the aspect of boundaries to the linguistic structure and the aspect of prominence to the attentional state. Thus, variations in the boundary marking (signalled through the string of words and the prosody) could be interpreted as variations in the signalling of the linguistic structure. Variations in the prominence marking (signalled through the string of words and the prosody) could be interpreted as variations in the signalling of the attentional state. The actual variations related to the string of words and the prosody are studied across the speaking styles.

Variations are studied also across conditions and across tasks. If the boundary marking changes across conditions Read and Listen, we could express this as the linguistic structure being more affected by, and thus more dependent upon, the prosody. If the opposite is the case, i.e. there is a very slight difference across conditions Read and Listen, we could interpret this as the linguistic structure being more dependent on the string of words. If the annotation of prominence changes across conditions Read and Listen, we could interpret this as the attentional state being dependent on the prosody. If the opposite is the case we could interpret this as the attentional state being more dependent on the string of words.

In order to study the interaction between the string of words and the prosody we have selected a small number of specific linguistic and prosodic features. However, the number of variables to be compared is still rather high, therefore for the sake of clarity we recapitulate all the features studied. First, in order to study the impact of the string of words and the prosody on the boundary and prominence annotations we distinguish between *Task*, *Condition* and *Style*.

- **Task** Segmenting: the boundaries and the prominences as two aspects of a segment. This is mirrored in the two studies "Boundaries" and "Prominences".
- **Condition** The impact of prosody: annotations in each task made either on the basis of transcripts alone (Read) or on the basis of transcripts together with the speech signal (Listen).

- **Style** The impact of the string of words: all annotations in each task studied in four different speaking styles assumed to be structurally different.

Using these three variables we can investigate whether there are any differences between the tasks (the boundary and prominence annotations) which seem to have been influenced by condition (prosody) or style (speaking styles). For example, within the task boundary annotation there was a high level of agreement across conditions Read and Listen indicating that to a great extent the subjects relied on the string of words in the boundary annotations. However, there was also a difference in the level of agreement within conditions across styles, showing a higher level of agreement in the scripted styles than in the non-scripted ones. Thus, we can study some specific types of differences across tasks, conditions and styles, and relate specific linguistic and prosodic features to a theoretical account for discourse segments.

In order to investigate more closely the type of differences, we have studied a number of selected *linguistic* and *prosodic* features in both tasks for all four speaking styles. The linguistic features are studied in both conditions Read and Listen but the prosodic features only in condition Listen.

- Linguistic characteristics of the *boundary* annotations: Phrasal context, part-of-speech context.
- Prosodic characteristics of the *boundary* annotations: prosodic phrasing (pausing).
- Linguistic characteristics of the *prominence* annotations: Phrasal context, part-of-speech context.
- Prosodic characteristics of the *prominence* annotations: focal accent (F0, duration, intensity).

Given the linguistic and prosodic features we can describe in more detail the kind of difference found across speaking styles with regard to boundaries and prominences. For example, in the boundary annotations there were differences in the agreement between speaking styles. In addition there are differences in the part-of-speech context of boundaries as well as in the pausal pattern across speaking styles.

In the rest of this chapter we discuss the relationships between the variables described above and relate the annotations to the theoretical concept of discourse segments suggested by Grosz and Sidner (1986). In this chapter we will argue that:

- The contributions of the string of words and the prosody regarding boundaries and prominences vary across speaking styles. This issue is addressed in section 8.1.
- The interaction between the string of words and the prosody is different across speaking styles. The topic is discussed in more detail in section 8.2.

The third research question was how prosody could be integrated into an account of discourse structure. In the last section of this chapter, section 8.3, a way of accounting for the results from the studies in this thesis within the framework of Grosz and Sidner (1986) is suggested.

The annotation task was not without difficulties for the subjects, and therefore issues related to the annotations are discussed relative to the discussion of the inter-annotator agreement in section 8.1. We argue that even though the tasks were difficult for the subjects and did pose some problems, the results are useful.

8.1 The Annotation Task and the Inter-Annotator Agreement

In this section the annotations as well as the inter-annotator agreement are discussed. The main point is to state what characterized the majority annotations across tasks, conditions and speaking styles, thus forming a base for further discussion of the relationship between boundaries and prominences in the speaking styles.

The claim we make in this section is that:

- The contributions of the string of words and the prosody (i.e. condition read and Listen) regarding boundaries and prominences vary across tasks and speaking styles.

Before turning to the discussion of the annotation task, we return to the example of the scripted and non-scripted monologues from chapter 1. However, in order to show the what the annotated transcripts look like, the example is shown with the majority annotations of boundaries and prominences from conditions Read and Listen inserted.

The subjects' annotations are coded as follows: a boundary in condition Read is indicated by an ! (exclamation mark) and in condition Listen by a £ (liber). A prominence annotation in condition Read is indicated with [ ] (square brackets around the word) and in condition Listen ( ) is used. In addition -<focal> attached to a word means that it has a focal accent while <pause> indicates a pause in that position.

Example 8.1 shows an excerpt from the scripted monologue (DN article read aloud), while example 8.2 shows the non-scripted monologue (Dn article retold).

(8.1) <pause> !£ ([arten])-<focal> [människa] skaffar sig en ([religion])-<focal> <pause> !£ det är en ([naturlig]) följd av att vi till skillnad från [övriga] arter har ett ([språk])-<focal> <pause> !£

English transliteration:

Species-DEF-SG human gets self a religion. This is a natural consequence of that we to difference from other species have a language.

English translation:
The human species acquires a religion. This is a natural consequence of the fact that unlike other species we have a language.

(8.2) <pause> !£ det gör det gjorde att människan kunde ehm [kommunicera] med (varandra)-<focal> och kunna [samordna] sina krafter till att till exempel jaga ([villebråd]) !£

English transliteration:
...this does this did that man-DEF-SG could-FIN ehm communicate with one-another-PL and be-able-to-INF coordinate their forces to to for instance hunt game...

English translation:
...this does this did mean that man could ehm communicate with one another and be able to coordinate their forces in order to hunt game...

The excerpts show that the boundary marking is carried out in a very similar way across conditions Read and Listen. In the scripted monologue (example 8.1) the overlap between pauses and boundaries is considerable, but in the non-scripted monologue (example 8.2) only one of the two boundaries is annotated at a pause position. In the scripted monologue, all prominences annotated in condition Listen are also annotated in condition Read. In condition Read, where subjects did not have access to the speech signal, there are also additional annotations in this excerpt. A similar pattern is found in the non-scripted monologue.

Inspecting the prominence pattern in the scripted monologue, there is no great difference between conditions Read and Listen. There are two more words annotated as prominent in condition Read, but reading the two versions aloud there is no great difference between them in the impression of the structure of the excerpt. However, inspecting the non-scripted monologue, there is a rather interesting difference in the prominence annotations across conditions Read and Listen. In condition Read, the subjects have annotated the two words "kommunicera" and "samordna" as prominent, while they have ignored "varandra". "Varandra" is, however, annotated as prominent in condition Listen. In both conditions Read and Listen the final word "villebråd" is annotated as prominent. Reading these two proposals (i.e. from Read and Listen) aloud, two rather different versions concerning the structure of the message evolve.

Starting with the prominence pattern suggested in condition Read, the words "kommunicera" and "samordna" would be interpreted as parallel, both sharing the modal verb

"kunna", which, however, in front of "kommunicera" is infinite and in front of "samordna" is finite. Thus, the string of words seems to be taken to signal a parallel relationship between "kommunicera" and "samordna", which gives an impression of coherence to the string of words...*kunde ehm kommunicera med varandra och kunna samordna sina krafter... (could ehm communicate with one-another and be-able-to coordinate their forces).* This contrasts with the impression resulting from the prominence pattern in condition Listen, where the word "varandra" which is focal is annotated as prominent. Reading the version from condition Listen, with a prominence on "varandra" the existing parallel between "kommunicera" and "samordna" seems to be cancelled. Thus, a prominence on "varandra" removes the interpretation of "och" as clearly coordinating "kunna kommunicera" and "kunde samordna". Furthermore it removes much of the impression of a problematic mix of an infinite and a finite verb. The impression of the message thus becomes different in condition Listen from in condition Read. The realization of the prominence seems to influence the segmenting in the non-scripted style to a higher extent than in the scripted style.

The impression of segmenting in the non-scripted monologue does not seem to be strong enough to influence the boundary marking since no majority boundary is located in front of "och". However, even though this is not the case it gives a picture of how the impact of prominence can vary across speaking styles.

The above example highlights an important result from our study: the agreement regarding boundaries and the disagreement regarding prominences. Moreover, it highlights the result of a disagreement concerning prominences. In the example from the scripted monologue there does not seem to be any alternative to the interpretation of the internal coherence between condition Read and condition Listen. One reason for this might be the clear syntactic structure. However, in the non-scripted monologue the string of words seems to be able to give rise to such a difference, and the prosody might be assumed to play a prominent role in conveying a specific interpretation of the message. One reason for this might be the less well defined syntactic structure, i.e. the less clear boundaries in the non-scripted styles.

Let us sum up the observations from the examples 8.1 and 8.2. The differences between the boundary annotations across conditions Read and Listen are slight in both speaking styles. The differences across conditions Read and Listen concerning the prominence annotations are substantial in both speaking styles, and also of a different kind in the scripted monologue than in the non-scripted monologue. The non-scripted monologue seems to introduce a greater impression of change between condition Read and Listen compared with condition Read. Thus, the prominence seems to be able to influence the segmenting more in the non-scripted speech than in the scripted speech.

8.1.1 The Annotation Task

In this section we recapitulate some of the issues touched upon in connection with the annotation tasks. Although the tasks were not without their problems, we nevertheless argue that the majority annotations we used did serve the purpose of marking general boundaries and prominences.

The annotations were difficult and time-consuming to carry out for the annotators. This was pointed out by a number of subjects. The method of using punctuation can be discussed, since we can see three possible problems with this: i) by implication the annotations would be inserted on a sentence and clause level, and in discourse segmenting we might want larger segments, ii) the original written text is better suited to punctuation, and therefore the application to the original spoken discourse is unreasonable and iii) making use of the familiar punctuation in the boundary annotation task and pairing this with the completely unfamiliar task of annotating prominences introduce a clear difference in the familiarity of the task between boundary and prominence annotation. We discuss each of these three objections separately. In addition we discuss the annotation frequency in both annotation tasks, pointing out the average length of the segments.

Concerning the first issue, we use the majority annotations in our analysis, which Swerts (1997) calls "strong boundaries". This procedure could be assumed to filter out many boundary annotations on a lower level. Initially our aim with using punctuation was to make a difference between paragraph markings, sentence markings and markings below the sentence level. However, on examining the data it became clear that the subjects used punctuation in very different ways, and that only a few of the subjects annotated paragraph boundaries. This led to the use of punctuation sites as described in chapter 4. Moreover, to give stability to our results we decided to use only the majority annotations. This screened out less well defined boundaries between e.g. phrases and kept stronger boundaries and sentence and paragraph boundaries.

We agree that there is an inherent effect concerning the differences in the suitability of punctuation applied to originally written or original spoken discourse. However, since we wanted to avoid lengthy guidelines for how to perform discourse segmentation, we found the procedure of punctuation attractive. Moreover, the difference in applicability might also capture some of the structural differences between scripted and non-scripted speech, i.e. since we know that the punctuation suits the scripted styles, a difference between scripted and non-scripted styles might indicate a difference in the structures, making the punctuation less ideal for the non-scripted styles. Since we are studying how such differences influence segmenting, and are not aiming to develop the best set of boundary markers independent of style, we believe that a difference in applicability might also yield interesting results. Thus, in view of these considerations we feel that the use of punctuation was a useful choice.

In both annotation tasks we tried to influence the subjects as little as possible in our instructions while still providing as clear instructions as possible. The boundary marking

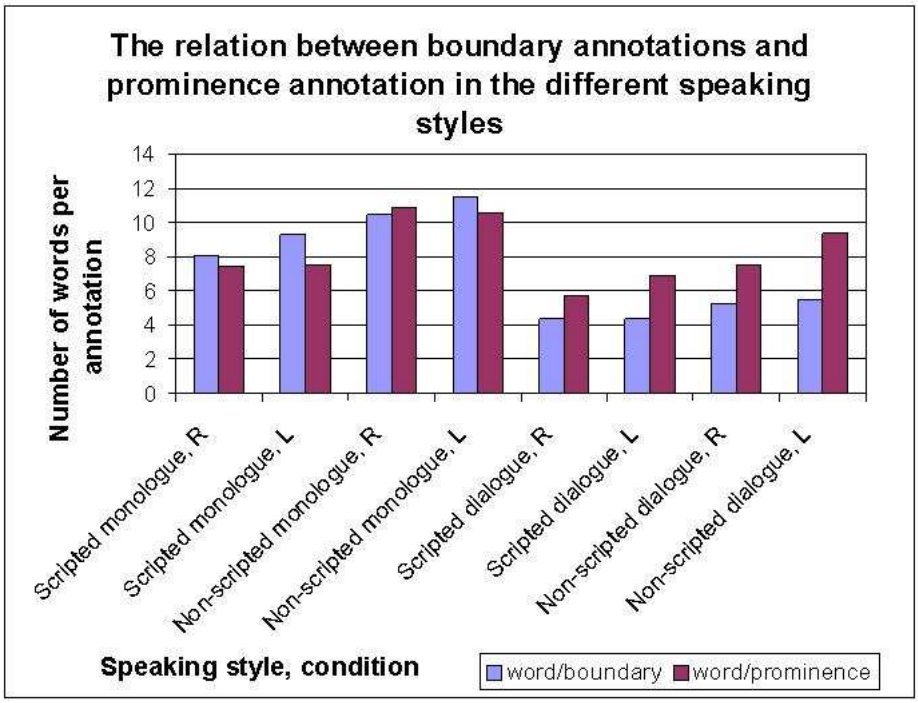

Figure 8.1: The average annotation frequency for boundaries and prominences.

task was in some way familiar, but the prominence marking in the task was entirely new for all the subjects. Initially we feared that our instructions might be too detailed, since we clearly stated that prominence is often correlated to acoustic prominence. Moreover, for readability the transcripts had retained punctuation in the prominence annotation task. Upon examining the results and finding a wide variation in the prominence annotations, we find that the subjects annotated in very different ways which suggests that they were not influenced by the instructions.

The annotation frequency differed greatly across subjects in the prominence annotation task, and to a certain extent also in the boundary annotation task. However, comparing the average frequency for the majority boundaries and the average frequency for the majority prominences, the measurements converge. Figure 8.1 accounts for the average annotation frequencies for boundaries and prominences in conditions Read and listen, all speaking styles.

Figure 8.1 shows the average annotation frequency for boundaries and prominences. For each speaking style two pairs of bars are shown, the first showing words per boundary and words per prominence in condition Read (with an R after the name of the speaking style), and the second pair of bars showing the same measurements in condition Listen (with a L after the name of the speaking style).

In the monologues the figures for boundaries and prominences are very close, indicating that each annotated segment contains on average one word which is annotated as prominent. On average the segments contain about 8 words, and approximately one of them is annotated as prominent. Thus, these segments are similar to sentences containing one sentence accent. The non-scripted monologue shows very similar figures. The similarity is greater in the monologues than in the dialogues. In the dialogues the segments are on average 5 words, and about every 6th to 9th word is annotated as prominent. This indicates shorter segments and also more of a mismatch between segments and prominences than was the case in the dialogues. This might be related to a tendency by the subjects to annotate boundary at speaker change. In small segments, containing e.g. one short feedback expression, there is no prominence annotation. In other words, the number of segments is higher in relation to the prominences since some segments are "empty" of prominence because they consist of very reduced feedback expressions.

The indication that the boundary annotation frequency converges with the prominence annotation frequency supports the view that a relevant segment contains a prominent part. This is not news, but it lends more weight to our annotation data. In addition, applying the method of Swerts (1997), we can assume that the boundaries are strong boundaries. Thus, it was concluded that the annotations suit the purposes of the studies.

8.1.2 The Inter-Annotator Agreement

In this section we discuss the relationship between the inter-annotator agreement in the different tasks, conditions and speaking styles. The section largely consists of a recapitulation of the inter-annotator agreement across tasks, conditions and speaking styles, but we also expound on the issue of how the graphic representation might have influenced the annotators.

In particular we want to stress:

- The boundary annotation task was characterized by a rather high level of agreement in the scripted styles but a lower one in the non-scripted styles, as well as a high level of agreement across conditions Read and Listen.
- The prominence annotation task was characterized by a low level of agreement in general across speaking styles and low level of agreement across conditions Read and Listen.

The annotations of boundaries and prominences show quite different properties. In table 8.1 a compilation of the inter-annotator agreements expressed with κ from both tasks is offered. The first column shows speaking style, the second shows the inter-annotator agreement within condition Read, the third shows the inter-annotator agreement within condition Listen and the fourth column shows the comparison of the majority annotations across conditions Read and Listen. Thus, since the fourth column compares only

the majority annotations, the figures differ from the figures for column two and three where the total agreement within conditions is shown. The top half of the table shows the boundary annotation task and the lower half shows the prominence annotation task.

BOUNDARIES	READ	LISTEN	COMPARE
Scripted monologue	0.73	0.79	0.87
Non-scripted monologue	0.59	0.58	0.71
Scripted dialogue	0.78	0.70	0.85
Non-scripted dialogue	0.63	0.56	0.77
PROMINENCES	READ	LISTEN	COMPARE
Scripted monologue	0.33	0.35	0.56
Non-scripted monologue	0.37	0.36	0.53
Scripted dialogue	0.46	0.42	0.64
Non-scripted dialogue	0.43	0.39	0.54

Table 8.1: Inter-Annotator agreement for boundaries and prominences.

The level of inter-annotator agreement in the boundary annotation task is rather high compared with that in the prominence annotation task. In addition, the inter-annotator agreement in the boundary annotation task differs across speaking styles. We suggest that this is interpreted in terms of the clarity of boundaries. Referring back to the issue of applicability of punctuation, let us consider what applicability of punctuation implies. Punctuation marks are normally inserted between paragraphs, sentences and clauses. If the boundaries are clearly signalled syntactically, as in written text, it is easy to apply the punctuation marks. In cases when the boundaries are more vaguely signalled, then it is also harder see clearly where to insert the punctuation marks. Consequently, the punctuation task becomes harder. Interpreting the results from the boundary annotation task in the light of clarity of the boundaries, the differences in inter-annotator agreement across speaking styles suggest that the boundaries in the scripted styles were indicated more clearly, whereas the non-scripted styles were characterized by less clear boundaries.

One interesting feature is that the inter-annotator agreement in condition Listen in all speaking styles except the scripted monologue dropped, specifically in the dialogues. Does this indicate that the prosody makes everything more difficult? We believe that the answer is no. Instead we believe that this is a function of the prosody counteracting annotations inserted purely on the basis of the graphic representation. One clear example discussed in chapter 5 was the case of annotations in front of feedback expressions. We elaborate on this in order to clarify how a lower level of agreement might indicate a less automatic annotation. In condition Read the annotations of very short speaker contributions, such as very reduced feedback expressions as independent segments were more frequent than in condition Listen. We believe that these expressions were perceived as more distinct in condition Read when the expressions occurred on a separate line. In condition Listen, where the subjects in addition to the transcript heard the reduced form, the expressions were no longer so consistently interpreted as independent

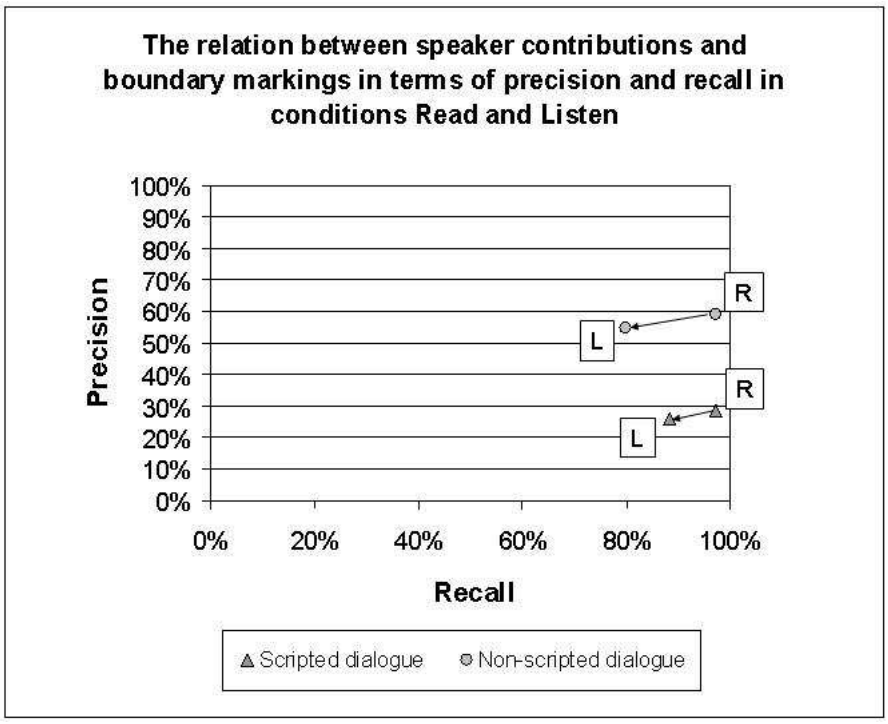

Figure 8.2: The relationship between speaker contribution and boundary annotations.

segments. This lowers the level of agreement in the dialogues. The actual difference in boundary annotations of speaker contributions is shown in figure 8.2, which accounts for the relationship between boundaries and speaker change in conditions Read and Listen.

In figure 8.2 the proportion of boundary markings at speaker change is indicated with Recall, while speaker change at boundary annotation is shown as Precision. The symbol at the end of the arrow indicates condition Read, while the symbol at the head of the arrow indicates condition Listen. Thus, there is a clear reduction in recall between condition Read and condition Listen. In both cases nearly 100% of all speaker changes in condition Read were annotated as boundaries, while in condition Listen this figure dropped considerably. The figure shows the difference in majority boundary annotations. This means that some subjects might still annotate these positions as boundaries, thus introducing a higher level of disagreement within condition Listen. This is an indication that the speech signal might have added information about the status of the contributions in the dialogues, thus paradoxically causing the level of agreement to drop. In condition Read on the other hand, the extra prosodic information was not present and all subjects could then follow the "text annotation strategy", which meant that the subjects could then follow cues in the graphic presentation. This effect indicates the strength of the "new line" cue and was not present in the monologues.

In contrast to the boundary annotation task, the level of agreement in the prominence marking task is low across both conditions and all speaking styles. Thus, the prominence annotation seems to be influenced less by style than was the case in the boundary annotation. The same relationship with a lower level of agreement for prominences than for boundaries is reported by Heldner (2001). Relating to the discussion of the applicability of the punctuation marks, this higher level of agreement on boundaries might be due to the fact that no speaking style had an advantage in the prominence annotation task, but we do not believe that this is the case. We believe instead that this effect is related to the linguistic properties of the prominent words, an issue we return to in section 8.1.3.

Furthermore, we believe that the low figures in the prominence annotation task are due partly to the subjects' differing interpretations of the annotation task. The annotation profiles show that the annotation frequency varied greatly across subjects and speaking styles, leading to a very low level of agreement. Again, we return to this issue in the discussion of the linguistic context of the annotations in section 8.1.3.

The level of agreement across conditions in the prominence marking task is substantially lower than the level of agreement across conditions in the boundary annotation task. This indicates that the annotators were influenced by the prosody in the annotation of prominence to a greater extent than in the annotation of boundaries. However, if the subjects were influenced by the speech signal, could we not expect a higher level of inter-annotator agreement in condition Listen than in condition Read? We believe that the absence of such an effect is due to the subjects' different annotation strategies which lowers the general level of inter-annotator agreement. However, the low level of agreement concerning the majority annotation in table 8.1 (column COMPARE) indicates that the prosody makes a difference to the annotations.

8.1.3 The Relationship Between the Annotations and the Linguistic Features

We now proceed to discuss the relationship between the majority annotations and the linguistic features. In this section we recapitulate the characteristics of the linguistic context of the boundary and prominence annotations. In the boundary annotation task there was a difference in the linguistic context of the boundary annotations across the speaking styles. This was, however, not the case in the prominence annotation task. Moreover, in none of the annotation tasks was there any substantial difference in the linguistic context of the annotation across conditions Read and Listen.

We have already suggested that this might indicate that our four speaking styles can be characterized as having different part-of-speech profiles for their strong boundaries. Below we list some typical properties of the context following a majority boundary in each speaking style:

- Scripted monologue: a high proportion of nouns and conjunctions
- Non-scripted monologue: a high proportion of conjunctions and a medium proportion of pronouns.
- Scripted dialogue: a high proportion of pronouns and a medium proportion of interjections (feedback).
- Non-scripted dialogue: a high proportion of interjections (feedback) and a medium proportion of conjunctions and adverbs.

This indicates that the scripted monologue could be assumed to have many boundaries in front of a sentence which begins with a noun subject, and also boundaries in front of coordinated and subordinated clauses. The high proportion of pronouns could be assumed to contain formal subjects or anaphoric pronouns.

The non-scripted monologue indicates a more narrative style with a very high proportion of boundaries in front of conjunctions. The high number of what could be noun subjects is thus not present in the non-scripted monologue. Moreover, boundaries are located in front of pronouns, which also here can be assumed to be formal subjects or anaphoric expressions.

The scripted dialogue has a very high proportion of boundaries in front of pronouns. An examination indicated that many of the pronouns were of the form "you" and "I". We believe that this is clearly related to the task. The conversation is between a newly deceased soul and S:t Peter, and they try to work out of what sex the soul had been when it was alive. In many cases the lines start as follows: "I don't know how I felt..." or "You have to remember...". The feedback expressions make up a large portion of the boundaries, however, not as large as in the non-scripted dialogue.

In the non-scripted dialogue the highest proportion of boundary annotations occurs in front of feedback expressions. In the second place come conjunctions and adverbs. The feedback expressions and the conjunctions are presumed to be related to the boundary marking, but we believe that the adverbs are related more to the task. In the present Map Task dialogues many instructions have the form of adverbs, such as "upwards, to the north", thus, it seems that the adverbs are related more to the task than to the speaking style.

Taken together, this indicates some similarities between the scripted styles – more nouns at segment beginnings – and the non-scripted styles – more function words at the segment beginnings. The dialogues are characterized both by a large overlap between the boundary positions and the positions of speaker change, and in addition clearer traces of the task are present in the dialogues than in the monologues.

Relating the clear differences in the boundary contexts across speaking styles to the marker hypothesis (Green, 1979), we might say that each of our four speaking styles is characterized by a different set of preferred markers at strong boundaries. Considering

the very slight differences across the conditions, in terms of the high level of agreement between condition Read and condition Listen, the boundaries seem to have been communicated primarily by the string of words. Thus, in the boundary annotation the subjects used specific sets of words and this cue was not often overridden by the prosody.

The opposite picture is emerging in the case of the prominence annotations. In all four speaking styles the same set of linguistic features were present at the prominent words across both speaking styles and conditions. However, the level of agreement across conditions Read and Listen was low, indicating that even though e.g. nouns were popular as prominent words in both conditions, it was not the same nouns which were annotated in condition Read as in condition Listen. The low level of agreement across conditions Read and Listen, together with the same set of candidates in both conditions indicates that subjects have been guided by the speech signal when annotating prominence much more than when annotating boundaries. Thus, the prosody in the speech signal seems to have been a stronger cue as to which words should be annotated as prominent than the information structure in the string of words. In other words, the cue from the string of words was often overridden by the prosody.

Upon examining the linguistic features, the conclusion is that boundaries in our data are located in front of words belonging to parts of speech, and that boundary annotations only to a lesser extent depend on the prosody. In addition the context set of the boundary annotations varies across speaking styles. The prominent words were also found in one particular set of candidates, nouns, and this set did not vary across speaking styles. In addition, the prominence annotations seem to be more sensitive to prosody than was the case in the boundary annotations.

8.1.4 The Relationship Between the Annotations and the Prosodic Features

In this section we restate the clearest results regarding the prosodic context of the boundary and prominence annotations. Since we are dealing with the context in the speech signal, only annotations from condition Listen are considered, since the annotations in condition Read could not be influenced by these features. In addition to the annotation context we discuss differences across the speaking styles based on the acoustic measurements.

In particular we stress that the pausing and prominence patterns differ across the speaking styles:

- The pausing patterns differ across speaking styles. In the scripted styles the boundaries and the prominences coincide with the pauses to a greater extent than in the non-scripted styles.

- The prominence patterns in terms of focal accent differ across speaking styles. In the non-scripted styles, a higher proportion of the prominences has focal accent than is the case in the scripted styles.
- In the scripted styles, where boundaries are more clearly signalled through the string of words, those boundaries are supported by pauses to a greater extent than in the non-scripted styles. On the contrary, the non-scripted styles, with less well defined boundaries signalled through the string of words, seem to have more clearly signalled prominences.

Based on the acoustic measurements reported in chapter 6, the picture that emerges is of four rather different speaking styles, where the differences are related to the different speaking situations, e.g. that more planning is required in non-scripted speech than in scripted. We describe all four styles below, leaving the scripted dialogue to the last since it is slightly different.

- The scripted monologue has a high articulation rate and rather long pauses which to a high degree overlap the boundaries. This presumably stems from the fact that the speakers did not have to plan the speech and that breathing and pausing could be done according to the script.
- The non-scripted monologue has a lower articulation rate, with longer pauses and a higher proportion of pauses which do not correlate with a boundary. We interpret this as a sign of the planning in this speaking style.
- The non-scripted dialogue is slightly faster than the non-scripted monologue, with shorter pauses and a lower proportion of boundaries which do not correlate to a pause. We interpret this as a result of the expected interaction, i.e. in a dialogue it is not one speaker's turn all the time. In a structured task like the Map Task dialogue, if the speaker whose turn it is does not speak, the other dialogue participant is likely to utter something, e.g. a question, thus reducing the silent interval.
- If the scripted dialogue was to conform to the above pattern relating non-scripted monologue and dialogue, it would have a higher articulation rate than the scripted monologue and also have shorter pauses. In addition it would have fewer pauses which are not coinciding with boundaries. However, the scripted dialogue deviates from this pattern. It does have shorter pauses (although these might have been edited since it is a radio play), but the proportion of pauses corresponding to a boundary is about the same as in the scripted monologue. Moreover, the articulation rate is substantially slower. We believe that these features are due to an intention to make an extremely clear communication in the acted speech. The clarity is produced by the pauses being long enough to be obvious, but not disturbingly long (possibly as a result of editing), and by making the pauses to a great extent correlate to linguistic boundaries, and by pronouncing the words carefully and rather slowly.

The findings concerning the scripted and non-scripted monologue and the non-scripted dialogue are fully in line with what is reported for non-professional reading and spontaneous monologue by Strangert (1993) and for non-professional reading and spontaneous dialogue by Gustafson-Čapková and Megyesi (2001, 2002) and Megyesi and Gustafson-Čapková (2001, 2002). In all these cases the difference in planning in relation to the differences in the pausing pattern was stressed, see especially the discussion in Strangert (1993).

Turning to the focal accent, the picture differed across speaking styles. In the scripted monologue a high proportion of the prominent words did not have a correlate in focal accent. However, in the two non-scripted styles as well as in the scripted dialogue the proportion was smaller. One interpretation of this is that the non-scripted styles together with the scripted dialogue in general were clearer in their signalling of prominence than was the scripted monologue. This gives us the interesting outcome that in the scripted styles the pausing was to a greater extent coincidental with the boundaries. Thus the pausing supported the boundaries more in the scripted styles than in the non-scripted ones. In the case of prominence the non-scripted styles did have stronger support from the focal accent than the scripted monologue. We leave this issue for now, but return to it in section 8.2.

When categorizing the prominent words which lack a focal accent we found that the instances were distributed over a range of contexts. In all speaking styles non-focal prominent words are located at the beginnings and the ends of clauses and phrases. This indicates that F0 gestures partly related to phrasing also signal prominence. The phenomena of phrase initial prominences is reported also e.g. by Gårding (1967), Bruce *et al.* (1993) and Strangert (1993) for Swedish and by Stifelman (1995) for English.

One of the more salient differences regarding the contexts of prominences which are not focal is that the non-scripted styles have a high number of words in the category of "other". This indicates that the categories based primarily on criteria related to text structure are not applicable to the non-scripted styles to the same extent.

We have recapitulated and discussed the most prominent results from the boundary and the prominence annotation tasks and now proceed to discussing their interrelationship.

8.2 A Unified View on Discourse Segmenting

In this section we account for the relationships between the boundary and prominence annotations in the four speaking styles investigated in our studies. Thus, we discuss the second claim:

- There are differences in the interaction between the string of words and the prosody across speaking styles. We will argue that this difference in the distribution of labour across speaking styles is determined by a principle of economy and clarity.

In addition we discuss two tendencies; firstly that pauses seem to support other boundary markings and in this way make them clearer, secondly that focal accent seems to establish prominence.

Based on the studies of the boundaries and the prominences, a picture of the differences in the features of majority boundaries and majority prominences across speaking styles has emerged. There are also differences in the impact of the speech signal across the boundary and prominence annotation, the prominence annotations being more sensitive to prosody than the boundary annotations. In this section we put together the different results from the boundary and the prominence aspect into a unified picture of the discourse segment in the four speaking styles.

The inter-annotator agreement figures, the linguistic features and the prosodic features in general indicate differences across speaking styles. Initially we stated that prosodic phrases have two aspects, the boundary and the prominence. We hypothesized that the discourse segments might also consist of both boundaries and prominences, and we further hypothesized that a variation in the structure of the string of words would influence both the aspect of boundaries and the aspect of prominences in the discourse segment. Moreover, we hypothesized a complementary relationship where a less clear boundary marking would require a clearer marking of the prominence. Do our data indicate such a complementary relationship? We argue that the answer is yes.

We have selected different speaking styles as examples of different structures. Our results from the boundary annotation task indicate that the boundary marking could be classified as clearer in the scripted speaking styles since there was a higher level of inter-annotator agreement in the scripted styles than in the non-scripted ones. Moreover, the feature of prosodic phrasing that we have studied, pausing, is supporting the structure given by the string of words rather than establishing boundaries by own virtue. This is indicated by the high recall of pauses at boundaries in the scripted styles and the low recall in the non-scripted styles in figure 5.12. In other words, in the scripted styles the pausing pattern supports the (syntactic) boundaries, while in the non-scripted styles the boundaries signalled through the string of words are less well defined, and to a great extent the pauses are also located at non-boundaries. The variation in the clarity of the boundary signalling, as mirrored by the annotations in our study, seems to be primarily related to the organization of the strings of words, since the high level of agreement across conditions Read and Listen indicates that the speech signal does not greatly influence the annotations. How then do the boundaries relate to the prominences?

In the case of prominences there were also differences across speaking styles. One of the most visible differences across speaking styles was that the proportion of prominences with a correlate in focal accent varied across speaking styles, with a high proportion of non-focal prominent words in the scripted monologue and a lower proportion in the non-scripted styles and the scripted dialogue. Thus, we might say that the non-scripted monologue was the clearest one in this respect, since it contained the lowest proportion of non-focal prominences.

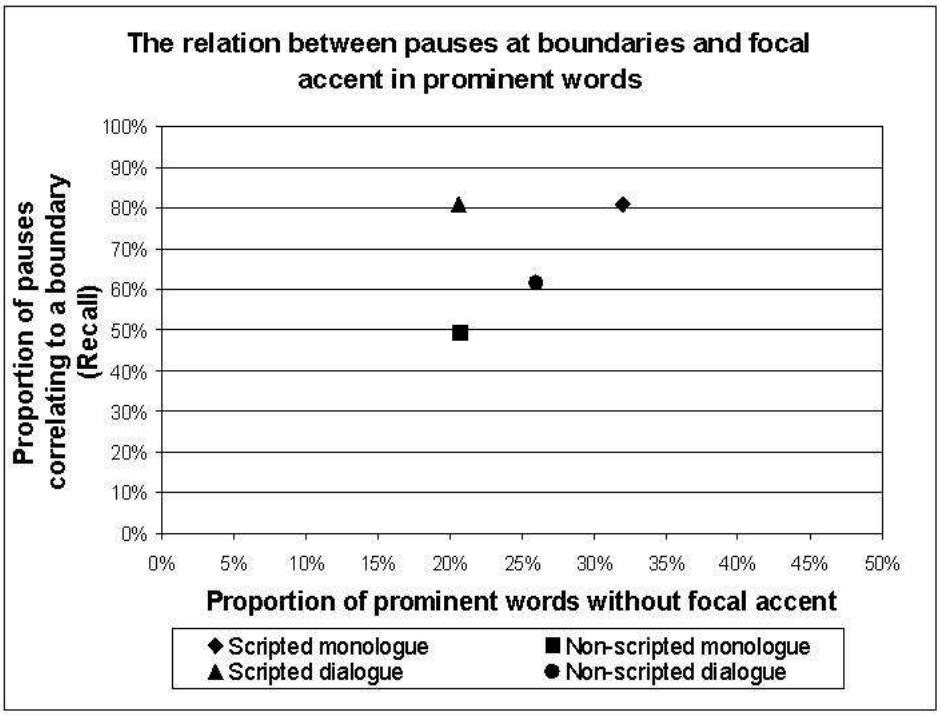

Figure 8.3: The relationship between boundaries and prominence in the speaking styles.

Let us relate the prosodic aspects of prominences and boundaries, focal accent and pauses to each other for each speaking style, i.e. for each different structure of discourse. As a measurement of the boundaries we select the recall figures for boundaries and pauses, i.e. the proportion of pauses correlating to a boundary. We assume that a higher proportion of pauses coinciding with a boundary annotation implies clearer boundaries. As a measurement of the prominences we select the percentage proportion of prominences not correlating to a focal accent. We assume that a higher proportion of prominences which also have a focal accent implies clearer prominences. Figure 8.3 shows the relationship between these two measurements in all four speaking styles.

The figure shows that a higher proportion of pauses corresponding to a boundary goes together with a higher proportion of prominent words without focal accent. Thus, when the pauses are clear boundary signals (high proportion of pauses corresponding to a boundary) and not present at other positions, the content – the prominent word – is signalled by more variation. When the importance of the pauses as boundary signals decreases, i.e. when many pauses are at positions other than boundaries, the proportion of prominent words correlating with focal accent is greater. This means that when the boundaries as signalled by the string of words become less clear, the prosodic signal of the prominence becomes more clear.

The scripted monologue has a clear organization of the string of words, and the pauses to a high degree support this (syntactic) organization (high recall). However, it also has a high degree of prominent words without focal accent. Slightly lower in recall comes the non-scripted dialogue, but the proportion of non-focal prominences has also decreased. The lowest recall is found in the non-scripted monologue, where the string of words conveys the least well defined boundary markings. However, this speaking style also shows the lowest proportion of non-focal prominences. These figures indicate a complementary relationship between boundaries and prominences in our data, i.e. when the boundaries are clear, the prominences can be less clearly signalled. However, there is one speaking style which goes against this principle: the scripted dialogue, i.e. the acted dialogue.

We believe that this relationship between boundaries and prominence in the speaking styles could be explained in terms of clarity and economy. We assume that a higher number of coinciding cues is clearer, e.g. a boundary indicated by both a conjunction and a pause is clearer than a boundary indicated only by a conjunction. Following the idea that a segment consists of both boundaries and prominence (content), we can think about a segment signalled to some extent by the boundaries and to some extent by the prominence. The segment is signalled clearly enough when the combination of the boundary signals and the prominence signals achieve a specific threshold level. An extremely clearly signalled segment consists of both clearly signalled boundaries and a clearly signalled prominence.

Both the clarity of the string of words and the clarity of the speech signal are included in the aspect of clarity. We have already suggested that a more scripted style contains boundaries more clearly signalled through the string of words. When this clear boundary signal is supported also by the pausing, the need for clarity in the prominence signalling seems to decrease. In our data, in the scripted monologue a high proportion of the pauses correlates with the boundaries. The boundaries are thus very clearly signalled, but the scripted monologue also has a high proportion of prominences not correlating to a focal accent. So, it seems that since there is a high degree of clarity in the boundary signalling, the requirements on the clarity of the prominence are not so strict, for economical reasons. In our data, this relationship is indicated in the scripted monologue, the non-scripted monologue and the non-scripted dialogue.

The acted speech has a high recall together with a low proportion of prominences not correlating to a focal accent. We interpret this as indicating that the acting goes against the principle of economy and favours only the principle of clarity. We find it plausible that such a deviation is found in the acted speech. In acted speech there is a demand for clarity, and in addition, the interaction in the dialogue per se is not the point, but the form of the interaction is.

To sum up: our data indicates that there might be a complementary relationship between the aspect of boundary signalling and the aspect of prominence signalling in different speaking styles. The complementary relationship seems to depend on the type of di-

scourse. When the boundaries are more clearly signalled through the string of words, the prominence is less clearly signalled through the focal accent (scripted monologue). When the boundaries are less clearly signalled through the string of words (non-scripted style), they are more clearly signalled through the focal accent. Thus, the type of discourse works in agreement with the prosody in an economic signalling of the segmental structure.

The data in the study reported by Swerts and Geluykens (1994) reveals a similar relationship between F0 maximum on topical noun phrases and pause length. Swerts and Geluykens (1994) report that the subject who had the lowest proportion of F0 maximum on topical noun phrases gave the clearest boundary signals in terms of pausing. In the two other cases, where subjects had a higher proportion of F0 maximum at topical noun phrases, the pauses at segment boundaries were substantially shorter. Thus, the results reported by Swerts and Geluykens (1994) support the view that there might be a complementary relationship between the rhythm in terms of pauses and melodic features in terms of F0 in relation to prominent phrases.

Our results indicate yet another complementary relationship. The boundary markings were not affected very much by the absence of or access to the speech signal, thus, it seems that the pausing in most cases supports the boundaries signalled by the string of words rather than establishes new boundaries. In the case of prominences the opposite was the case, i.e. the difference was substantial across conditions, thus, it seems that the single words – with their meanings – are playing the role of prominent elements rather than establishing prominence by their own virtue. Regarding the prominence, the primary work seems to be done by the acoustic prominence, and the single words support what the acoustic prominence indicates and gives it reference.

Let us elaborate on this in relation to what Bruce (1998) mentions about word recognition. In the continuous speech signal where the "blank spaces" between the words are absent, we perceive the peaks of the words in terms of prominent parts as being important targets to aim for. We can imagine that the more separated objects like words are with e.g. blank spaces, the less important specific peaks inside the objects become. Thus, when the parts between the words which are withheld becomes clearer with the help of e.g. extremely clear boundaries, the accentuation does not have to be that clear since the withheld part – the boundary aspect – has increased.

If this mechanism is at work on the level of word recognition, it is not implausible to assume that it might be at work also on higher levels in the discourse. Thus, if there are less clear boundaries in the discourse units we might need to aim for clearer prominent peaks. However, the discourse segments are not signalled solely by either the string of words or the prosody, but both aspects have to be included in the notion of clear or less clear signals. If we assume that clear syntactic boundaries are one way to make segment boundaries clear, then speaking styles with clear syntactic boundaries would be less dependent on the prominent peaks, while the opposite would be the case for less clearly signalled boundaries.

The results from the studies in this thesis could thus be interpreted as follows: in the scripted styles the segmenting relies more on the aspect of boundaries, while in the case of non-scripted speech the segmenting is more influenced by the prominence. Thus, in the scripted styles the shaping of discourse segments can rely to a fairly great extent on the boundary marking as signalled by the string of words because of a clearer sentence structure. The non-scripted styles are more dependent on the prominences for forming discourse segments since the boundary signalling sentence structure is less clear. Since the non-scripted styles are more dependent on the prominences they are also more dependent on the speech signal. This mirrors the rather trivial fact that spontaneous speech is more dependent on the speech signal than a written text.

The differences between a focus on the boundaries and a focus on the prominences in relation to segmenting which are indicated in our data might be related to more general differences between the role of boundaries and the role of the cohesion in between. Selkirk 2000 has suggested a general tendency to a demarcative or a cohesive strategy in phrasing, and she suggests that a language has either a demarcative or a cohesive strategy. Hansson (2003) reported results indicating that Swedish spontaneous speech uses a cohesive strategy. Our results could be interpreted as giving some support to the results reported by Hansson (2003), but such an interpretation also indicates that strategies might differ across speaking styles. In other words, demarcation and a greater dependence on the boundaries are related to a written syntax, while cohesiveness and a greater dependence on prominence are related to a spoken language syntax. The studies carried out in this thesis and the studies carried out by Selkirk (2000) and Hansson (2003) differ greatly in their approach. Nevertheless, the issue is interesting enough to pursue further.

The results concerning the segmenting in different speaking styles might have a bearing also on computational approaches to discourse processing. If the strategies towards the segmenting differ across speaking styles, segmenting strategies in computational approaches should perhaps also differ depending on style. For instance, a demarcative parsing strategy might be more feasible in scripted styles, while in non-scripted styles a prominence determined parsing strategy would be more appropriate.

Let us relate the above issues to the initial question behind this thesis. In the first chapter the idea is put forward that some motifs would be better communicated by the boundaries than by the content, and other motifs would be better communicated by the content than by the boundaries. Regarding discourse segments, the data used in the studies in this thesis indicate that to some extent this is the case. Thus, returning to the metaphor of a motif, we interpret the segments in the scripted styles as communicated to a great degree through the boundaries, while the segments in the non-scripted styles are communicated to a great degree through the content. In the case of the acted dialogue the principle of economy was violated, and in order to achieve maximum clarity both boundaries and content were clearly communicated.

8.2.1 The Intentions Behind the Segments

In the study of discourse intentions in the form of questions, the relationship to the discourse segment is of another nature. We recapitalate some of the basic prerequisites and findings from the study of the questions.

Questions are in some cases related to a specific question intonation, but many researchers have pointed out that a specific question intonation is not always present (Bruce, 1998), (Hirschberg, 2000), (House, 2003). Thus, it seems that we cannot rely solely on prosody as a cue to what is a question. So, what kind of cues are used in subjects' assessments of specific segments as being, or not being, questions?

Our results regarding the question segments indicate a high level of agreement in the annotations across conditions Read and Listen, but with some differences across question types. In general, subjects did not seem to be greatly influenced by having access to the speech signal. They were, however, more influenced in some specific cases than in others. There are some specific types of questions on which subjects agree more; the question word questions and the verb initial questions, i.e. questions of a clearer syntactic form. However, in the case of questions of a less well defined form the level of disagreement was higher across conditions Read and Listen. We remind the reader that the data was too sparse for us to use the monologues, and therefore we use only the dialogues. This means that we cannot conclude anything about the difference between interactive and non-interactive styles.

In our results, there were above all two factors that seem to have influenced the annotations of the segments as questions or not; firstly the scope of the segment and secondly the status of feedback expressions.

In the scripted dialogue, there were cases in condition Read where a possible site for a question mark was ignored, and the segment continued to a later point. In these cases the segment was not interpreted as a question, since the segment as a whole did not support such an interpretation. In condition Listen a pause might be present, thus supporting the segmenting of the string into two separate parts. The resulting segment then had to be interpreted, and a single segment, possibly in the form of a question, seems to have forced an interpretation of the segment as a question. As an example we recapitulate the example from chapter 7 in example 8.3.

(8.3) menar ni nyss [1] så sa jag endast ja [2]
mean you just a moment ago [1] then said I only yes [2]
If you mean just a moment ago I said only yes

The position indicated with a [1] is annotated with a question mark in condition Listen, but not in condition Read. Position [2] is not annotated as a question in any condition. There are silent pauses at both positions [1] and [2].

In the case of the non-scripted dialogues some very short contributions were interpreted as questions. In this case we assume a pattern similar to the one which was present in the boundary annotations in front of speaker change. Subjects seem to be more inclined to regard single words as separate segments without access to the speech signal, i.e. the written transcripts seem to make these words stand out more, but when the subjects hear the very reduced word, it loses the status of a separate segment. In other words, when subjects heard the very reduced form of the contributions they no longer interpreted them as a question, but as a dialogue regulating feedback instead. An example from the non-scripted dialogue, recapitulated from chapter 7, is given in example 8.4.

(8.4) när du är i höjd med den här bukten/ ja [1] / så längre ut till...
when you are in level with this here bay/ yes [1]/ then longer out to...
When you are level with this bay/ yes/ then a bit further to....

The annotation is similar to the previous example. The [1] represents the position where subjects in condition Read, but not in condition Listen, have annotated a question mark. There is no pause at the position.

In addition to the above suggestions concerning the difference in annotations, the F0 of the first speaker (uttering när du är...) might indicate a continuation of the contribution before and after the feedback expression. It thus forms one single segment through a coherent F0. In such a context a question would not be applicable. However, this last issue concerning the F0 contour across contributions was not pursued in the present study.

In both cases it seems plausible that the annotation is not made on the basis of a given segment, but that the interpretation of the segments as being or not being questions depends on the interpretation of the segmenting. We can hypothesize that the high level of agreement regarding the questions of a clearer form (i.e. verb initial questions or question word questions) is based on not only the clear question forms in the segments, but also to some degree on the segment context, e.g. whether the subsequent context contains an answer or not. Thus, it is possible that not only the question form is clearer in these cases, but also the question context. We cannot decide on this since we have not made an extensive analysis of all the contents of the speaking styles, but only of the passages where subjects have annotated a question. However, example 8.3 indicates that there were cases where a potential question form was ignored in condition Read.

The example in 8.3 shows a verb initial conditional sentence, and is thus not a question. This is also how we could assume the utterance to be interpreted in condition Read, since it was not marked with a question mark by the annotators. However, in condition Listen the the segment “menar ni nyss” was interpreted as a question. The reason for this interpretation could be the long pause after “nyss”, signalling that the segment “menar ni nyss” would be interpreted as an independent segment. Interpreted in this way, “menar ni nyss” is in this form a perfect verb initial question. Thus, even though the agreement between Read and Listen concerning the boundary annotation was high,

there were cases like this where the prosody clearly influenced the segmenting, and, as it seems, consequently also the interpretation of the resulting segments.

The preference (in example 8.3, condition Read) for regarding the whole sentence as one segment in condition Read could also be expressed in the form of discourse intentions as described by Grosz and Sidner (1986). In Grosz and Sidner's (1986) discourse theory intentions have to be satisfied, i.e. they are viewed as goals which have to be achieved. Thus, a request needs an answer. In the case of 8.3, the potential question form of "menar ni nyss" is clear, but the request is rather unclear. Moreover, the subsequent "så sa jag bara ja" does not provide an answer to "menar ni nyss", and the possible request does not show any sign of being satisfied. This is another way to show that the question interpretation is abandoned in condition Read. However, in condition Listen the question interpretation is forced by a long pause, leading to an interpretation of the sentence as consisting of two separate segments. In this case the preference was to interpret the string "menar ni nyss" as a question.

In our data, the intention in the form of questions seems to be rather clearly communicated by the form of the string of words since the level of agreement across conditions Read and Listen is high. However, in some cases – as e.g. in the case described above – there are also other factors influencing the interpretations, such as the scope and status of the segment as well as the extent to which the context supports the interpretation. Thus, our data indicates that the interpretation of a "question" is in some cases influenced by the context which might vary depending on the segmenting. We conclude that there seems to be an interesting relationship between the segment intention as communicated by the verbal message and the segmenting of this message. To communicate an intention thus includes the communication of both the words and the segmenting. This supports the view by Grosz and Sidner (1986) that the intention is an important feature in the shaping of the discourse. However, in order to make any further claims we would need to carry out a more extensive study of all the materials, not only the annotated materials included in this study.

Based on our materials it seems that a clear question form is less dependent on the speech signal. This is indicated by the higher level of agreement across condition Read and Listen concerning the question word questions compared with both the verb initial questions and the questions of declarative form. However, there are also indications that the interpretation of a segment as a question is dependent not only on the form, but also on the context of interpretation, i.e. the speech signal might disambiguate the segment as a question simply by separating it from another segment and thus force a separate interpretation of the potential question segment. From this we conclude that the segmenting can be an important factor in conveying an underlying intention.

8.3 The Relationship of the Studies of Boundaries and Prominences to Discourse Theory

In this section we suggest a method of linking the findings from our studies to discourse theory and thus express the differences across speaking styles regarding the segmenting within Grosz and Sidner's (1986) discourse theory. The claim made is that:

- The relationship between the boundaries and the prominence in the different speaking styles (and by implication the linguistic and prosodic features which we have studied) can be captured within a discourse theory. We suggest that such a relationship exists between our data and the discourse theory of Grosz and Sidner (1986).

We selected the discourse theory by Grosz and Sidner (1986) since this framework suited our data best. The experiment was designed using features from each of the three components constituting the discourse structure: the linguistic structure, the attentional state and the intentional structure. The linguistic structure is said to be the structure which is segmented into discourse segments, and therefore we related our study of boundaries to this structure. The attentional state is related to the salience of elements at certain points of discourse, and we parallelled this with our study of prominence. The intentional structure relates to the intentions behind the utterances, and in our study it is mirrored by the small study of questions.

In section 8.2, we argued for the view that the scripted styles were characterized by clearer boundaries which to a great extent were supported by the pausing. The non-scripted styles were characterized by less well defined boundaries, and the boundary markings and the pauses were coinciding to a lesser extent. Regarding the prominence the picture was the opposite. In the scripted styles the prominences were to a lesser extent coinciding with a focal accent, while in the non-scripted styles such a coincidence was present in a higher proportion. We now link these results to a discourse theory.

Transferred to discourse terms, we suggest that the differences regarding boundaries and prominences across the speaking styles mean that in the scripted styles more emphasis is placed on the linguistic structure, while the attentional state is more emphasized in the non-scripted styles. We explain our suggestion using the scripted and the non-scripted monologues as examples.

We have argued that our results indicate a difference between scripted and non-scripted monologue in the clarity of on one hand boundaries and on the other hand prominence. In the case of scripted monologue, the aspect of boundaries is more clearly conveyed through the string of words, while the prosody in terms of pausing supports the structure signalled through the string of words. The aspect of prominence is less clearly signalled, since there was a higher proportion of prominent words not correlating with focal accent. In our view, what defines a discourse segment is in scripted monologue more

closely related to the boundary marking and thus to a heavier workload on the linguistic structure. Regarding the non-scripted monologue, where the prominence cues communicated through focal accent were clearer, there is a heavier workload related to the attentional state.

Regarding the relationship of the prosodic features, in the experiment we have already related prosodic phrasing to the linguistic structure and focal accent to the attentional state. Based on our results we can describe our findings as follows: the acoustic feature of pausing seems to support primarily the linguistic structure, while the aspect of focal accent seems to influence the attentional state to a high degree.

We would like to to stress that in all styles the discourse segment consists of all three components, and it is not the case that e.g. written text might lack attentional state. Our suggestion concerns the relative workload, which features in one structure or another, might have in different types of discourses.

In the view of Grosz and Sidner (1986), the intentional structure determines the relationships between the discourse segments. We see nothing in our study of questions that argues against this view; however, in order to establish a clearer relationship a more elaborate study of the full data is needed.

The relationship between on one hand the interaction of prosodic features and discourse types and on the other hand the discourse theory of Grosz and Sidner (1986) is a proposal that needs more data and further research. We have only related a small number of features to a part of the theory, the one closely related to discourse segments, and we have left out the hierarchical discourse structure. Since this hierarchical structure is determined by the intentional structure, a closer study of the hierarchical relationships demands a more elaborate study of the intentional structure than the study carried out in this thesis. Thus, a range of issues remains to be pursued.

While interpreted within a discourse framework, the differences in our study have been related primarily to the dimension of scriptedness, i.e. to the differences between scripted and non-scripted styles. Thus, the dimension of interaction, i.e. the difference between monologue and dialogue, is not present to any great extent. We interpret these results in two ways. Firstly, the unified approach to monologue and dialogue in Grosz and Sidner's (1986) discourse theory might mean that differences between monologue and dialogue are not highly visible in their framework. In addition, the feature of interaction is perhaps not particularly well captured on a discourse segment level. Secondly, one of the more visible discourse level differences between interactive and non-interactive speech might lie primarily in the nature of relationships between segments instead of in the segments themselves. Thus, making a record of the types of relationships, such as e.g. elaboration, contrast, and concession might perhaps reveal more about the differences between monologue and dialogue, since they might be viewed as a kind of relationship profile for the speaking styles in the same ways as the parts-of-speech was on the word level. Thus, the difference between monologue and dialogue might be expressed as a difference in argumentative and interactive relationships between the segments rather

than by segment properties. We cannot conclude much about this in this study, but we find it an interesting issue to pursue.

Chapter 9

Conclusions

IN this thesis the discourse segments in four speaking styles have been investigated from both the aspect of boundaries and the aspect of prominence. The aim of the studies was to find out if the contribution of either of the two aspects differed across speaking styles. In addition both aspects were studied from a lexicogrammatical as well as a prosodic perspective. The results have been related to discourse theory, thus expressing the differences across speaking styles in terms of discourse level differences.

On the basis of the results from the studies in this thesis we conclude that the discourse segments are signalled through an interaction of the string of words and the prosody, and that this interaction varies across speaking styles. A clear boundary marking in the string of words can be supported by the pausing pattern, but in the case of such a clear boundary marking the prominence marking in terms of focal accent might be less clearly signalled. In contrast, a clear prominence marking in terms of focal accent can be combined with a less well definied boundary marking regarding both the string of words and the pausing.

The interaction of the aspect of boundaries and the aspect of prominence varied across speaking styles. In three of the four speaking styles (the scripted monologue, the non-scripted monologue and the non-scripted dialogue) the interaction was determined by a balance between the principles of clarity and economy. In the fourth speaking style (the scripted dialogue) the relationshipship between boundaries and prominences was determined only by the clarity.

The discourse segments are discourse level units and it should be possible to model them within a discourse theory. We propose to account for the variation in the properties of the discourse segments in terms of a variation in discourse level features within the discourse theory of Grosz and Sidner (1986). Thus, the differences across speaking styles in the contributions of the aspect of boundaries and the aspect of prominences in the discourse segment were related to differences in the contributions of the discourse components "linguistic structure" and "attentional state".

Returning to the initial question, whether some motifs can be more clearly expressed through the boundaries while other ones can be more clearly expressed by what is between the boundaries, it seems that this is the case for discourse segments. In the scripted styles the segments seem to be more clearly defined by the boundaries, while in the non-scripted styles there is a greater emphasis on the prominences. However, this holds for the "natural" speaking styles. In the case of acted speech the balance is changed, and also the impression of "naturalness". We find the conclusions interesting but would like to stress that in order to draw general conclusions we would need more data, and we would also need to investigate further linguistic and prosodic features.

If speaking styles in general do show this relationshipship of balance between boundaries and prominences, this is something which might be important in computational processing of the different speaking styles. In cases of read aloud speech a more demarcative segmenting strategy might be more feasible, while in more spontaneous speaking styles a prominence-focused or cohesive segmenting strategy might be more feasible. It remains to examine whether such properties would also have a bearing on parsing strategies in the different speaking styles.

Regarding the third discourse component, the intentional structure, our study did not yield any clear results. In some cases we saw that the interpretation of a segment depended on the segmentation, which in turn was influenced by the context of both the string of words and the prosody. However, a more elaborate study related to the intentional structure would be needed in order to make the relationship between boundaries, prominences and intentions more visible. In our view this question is closely related to the issue of discourse relations. In a similar way, another analysis would be needed regarding the difference in the dimension of interaction, and we suggest that a study of discourse relations would produce clearer results.

It would be an interesting continuation of the study in this thesis to study the actual discourse relations between the segments in different speaking styles. Do such relationships differ across speaking styles, of what kind are they, and what is their relationship to an underlying intention? Are discourse relations some kind of discourse level anaphora? We have left out many aspects of discourse structure in the present study, but we consider nevertheless the relationship between prosody and discourse structure suggested in this thesis to be a valid base for further investigations.

The suggestions made in this thesis are based on an empirical investigation of linguistic and prosodic features in different speaking styles. We believe that our suggestions do model our set of data, however, it should be stressed that this set of data is sparse. In addition the number of linguistic and prosodic features investigated is low. In order for us to draw more general conclusions, a more comprehensive study covering more data would be needed. Another type of future work is thus to develop a larger set of data, covering a higher number of speaking styles as well as a higher number of linguistic and prosodic features. Since the development of acoustically and prosodically annotated corpora is time-consuming, a procedure which allows different analyses to be added over

a period of time would be the most suitable. To conform to a stable format and to add annotations on different tiers would in our opinion be an ideal way to go forward.

Thus, much work remains to do, and a large number of questions are left to pursue.

References

Abney, S. 1992. Prosodic Structure, Performance Structure and Phrase Structure. In *Proceedings of Speech and Natural Language Workshop*, pp. 425–428, San Mateo, CA. Morgan Kaufmann Publishers.

Abney, S. 1995. Chunks and Dependencies. In Cole, J., Green, G. M., and Morgan, J. L., editors, *Computational Linguistics and the Foundations of Linguistic Theory*, pp. 145–164. CSLI.

Ahrenberg, L., Dahlbäck, N., and Jönsson, A. 1995. Coding Schemes for Studies of Natural Language Dialogue. In *Working Notes from AAAI Spring Symposium*, Stanford.

Allen, J. 1983. Recognizing intentions from natural language utterances. In Brady, M. and Berwick, R. C., editors, *Computational Models of Discourse*, pp. 107–133. MIT Press.

Allen, J. and Core, M. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. Technical report, Columbia University, New York.

Allen, J. and Perrault, C. 1980. Analysing Intention in Utterances. *Artificial Intelligence*, **15**(3):143–178.

Allwood, J. 1995. An activity based approach to pragmatics. Technical Report Technical Report, Gothenburg Papers in Theoretical Linguistics (GPTL) 75, Department of Linguistics.

Arim, E., Costa, F., and Freitas, T. 2003. An empirical account of the relation between discourse structure and pauses in Portuguese. In Mettouchi, A. and Ferré, G., editors, *Proceedings of IP2003, Prosodic Interfaces*, Nantes, France.

Arons, B. 1994. Pitch-Based Emphasis Detection for Segmenting Speech Recordings. In *Proceedings of the International Conference on Spoken Language Processing*, pp. 1931–1934.

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

Austin, J. L. 1962, 1975. *How to Do Things with Words.* Harward University Press, 2:nd edition.

Ayers, G. 1994. Discourse Functions of Pitch Range in Spontaneous and Read Speech. Technical Report 44, Ohio State Unuversity.

Ayers, G., Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D., and Touati, P. 1995. Modelling Intonation in Dialogue. In *Proceedings of the XIIIth International Congress on Phonetic Sciences (ICPhS)*, pp. 278–281, Stockholm.

Bachenko, J. and Fitzpatrick, E. 1990. A Computational Grammar of Dicourse-neutral Prosodic Phrasing in English. *Computational Linguistics*, **16**:155–170.

Biber, D. 1988. *Variation across speech and writing.* Cambridge University Press.

Blackburn, P. and Bos, J. forthcoming. *Representation and inference for Natural language: A First Course in Computational Semantics.* CSLI Press, Stanford.

Boersma, P. and Weenink, D. 1996. PRAAT, a system for doing phonetics by computer. Technical Report 132, Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam.

Bolinger, D. 1972. Accent is predictable (if you're a mind-reader). *Language*, **48**(3):633–644.

Bolinger, D. 1989. *Intonation and its Uses.* Edward Arnold.

Brants, T. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, Washington, USA.

Bruce, B. 1975. Generation as Social Action. In *Proceedings of Theoretical Issues in Natural Language Processing-I*, pp. 64–67, Cambridge, MA.

Bruce, G. 1977. *Swedish Word Accent in Sentence Perspective.* Ph.D. thesis, Travaux de l'Institute de Linguistique, Lund University. CWK Gleerup.

Bruce, G. 1982. Textual Aspects of Prosody in Swedish. *Phonetica*, **39**:274–287.

Bruce, G. 1995. Modelling Swedish Intonation for Read and Spontaneous Speech. In *Proceedings of International Congress on Phonetic Sciences*, volume 2, pp. 28–35, Stockholm.

Bruce, G. 1998. *Allmän och svensk prosodi.* Number 16 in Praktisk lingvistik. Department of Linguistics, Lund University.

Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D., and Touati, P. 1996. On the Analysis of Prosody in Interaction. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing Prosody*, pp. 43–59. Springer, New York.

Bruce, G., Granström, B., Gustafson, K., and House, D. 1993. Phrasing strategies in prosodic parsing and speech synthesis. In *Proceedings of Eurospeech '93*, volume 2, pp. 1205–1208, Berlin.

Carletta, J., Isard, A., Isard, S., Kowto, J., Doherty-Sneddon, G., and Anderson, A. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, **23**(1):13–31.

Chafe, W. L. 1982, 1993. Integration and Involvement in Speaking, Writing and Oral Literature. In Tannen, D., editor, *Spoken and Written Language: Exploring Orality and Literacy*, number 9 in Advances in Discourse Processes, pp. 35–53. Ablex Publishing Corporation.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurements*, **20**(1):37–46.

Cohen, P. and Levesque, H. 1991. Confirmations and Joint Action. In *Proceedings IJCAI-91*, pp. 951–957.

Cohen, P. and Perrault, C. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*, **3**(3):422–440.

Cutler, A., Dahan, D., and van Donselaar, W. 1997. Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, **40**(2):141–201.

di Eugenio, B. and Glass, M. 2004. The Kappa Statistic: A Second Look. *Computational Linguistics*, **30**(1):95– 101.

di Eugenio, B., Jordan, P. W., Thomason, R. H., and Moore, J. D. 2000. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human Computer Studies,*, **53**(6):1017– 1076.

van Donzel, M. 1999. *Prosodic Aspects of Information Structure in Discourse*. Ph.D. thesis, Netherlands Graduate School of Linguistics, Holland Academic Graphics.

Eckert, M. and Strube, M. 2000. Dialogue Acts, Synchronising Units and Anaphora Resolution. *Journal of Semantics*, **17**(1):51–89.

Einarsson, J. 1978. *Talad och skriven svenska*. Lundastudier i svensk språkvetenskap. Studentlitteratur, Lund.

Ejerhed, E., Källgren, G., Wennstedt, O., and Åström, M. 1992. The Linguistic Annotation System of the Stockholm-UmeåProject. Technical report, Department of General Linguistics, University of Umeå.

El Emam, K. 1999. Benchmarking Kappa: Interrater Agreement in Software Process Assessment. *Empirical Software Engineering*, **4**(2):113 – 133.

Enkvist, N. E. 1974. Några textlingvistiska grundfrågor. In Teleman, U. and Hultman, T. G., editors, *Språket i bruk*. LiberLäromedel, Lund. Gleerups.

Fant, G. and Kruckenberg, A. 1989. Preliminaries to the Study of Swedish Prose Reading and Reading Style. In *STL-QPSR:2*, pp. 1–83, Dept. of Speech, Music and Hearing, KTH, Sweden.

Fant, G., Kruckenberg, A., and Liljencrants, J. 2000. Acoustic-phonetic Analysis of Prominence in Swedish. In Botinis, A., editor, *Intonation: Analysis, Modelling and Technology*, pp. 55–86. Kluwer Academic Publishers.

Ferrara, K. 2001. Intonation in Discourse Markers – The Case of Anyway. In Wennerstrom, A., editor, *The Music of Everyday Speech*, pp. 117–130. Oxford University Press.

Fraser, B. 1990. An Approach to Discourse Markers. *Journal of Pragmatics*, **14**:383–395.

Frege, G. 1882. Ueber Sinn und Bedeutung, (Translated by Herbert Feigl: On Sense and Nominatum.). *Zeitschrift für Philosophie und Philosophische Kritik.*, **100**. (Translation Printed in Readings in Philosophical analysis, Feigl and Sellars eds., 1949).

Gårding, E. 1967. Prosodiska drag i spontant och uppläst tal. In Holm, G., editor, *Svenskt talspråk*, pp. 40–86, Stockholm. Almqvist & Wiksell.

Goldman-Eisler, F. 1972. Pauses, Clauses, Sentences. *Language and Speech*, **15**(2):103–113.

Green, T. R. G. 1979. The Necessity of Syntax Markers: Two Experiments with Artificial Languages. *Journal of Verbal Learning and Verbal Behavior*, **18**(4):481–496.

Grimsrud, B. 2002. *Själen och Sankte Per.* Theatre play. Broadcasted by Sveriges Radio, Radioteatern.

Grosjean, F., Grosjean, L., and Lane, H. 1979. The Patterns of Silence: Performance Structures in Sentence Production. *Cognitive Psychology*, **11**(1):58–81.

Grosz, B. J. and Hirschberg, J. 1992. Some Intonational Characteristics of Disocurse Structure. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 429–432.

Grosz, B. J., Joshi, A. K., and Weinstein, S. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, **21**(2):203–225.

Grosz, B. J. and Sidner, C. 1986. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, **12**(3):175–204.

Gumperz, J. J., Kaltman, H., and O'Connor, M. C. 1984. Cohesion in Spoken and Written Discourse: Ethnic Style and the Transition to Literacy. In Tannen, D., editor, *Coherence in Spoken and Written Discourse*, number 12 in Advances in Discourse Processes, pp. 3–19. Ablex Publishing Corporation.

Gundel, J. 1999. On Different Kinds of Focus. In Bosch, P. and van der Sandt, R., editors, *Focus*, pp. 293–305, Cambridge. Cambridge University Press.

Gustafson-Čapková, S. and Megyesi, B. 2001. A Comparative Study of Pauses in Dialogues and Read Speech. In *Proceedings of Eurospeech 2001*, volume 2, pp. 931–935, Aalborg, Denmark.

Gustafson-Čapková, S. and Megyesi, B. 2002. Silence and Discourse Context in Read Speech and Dialogues in Swedish. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France.

Halliday, M. A. K. and Hasan, R. 1976. *Cohesion in English.* Longman.

Hansson, P. 2003. *Prosodic Phrasing in Spontaneous Swedish.* Ph.D. thesis, Department of Lingustics, Lund University, Lund, Sweden.

Hearst, M. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd. Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pp. 9–16. Association for Computational Linguistics.

Heldner, M. 2001. *Focal Accent – f_0 movements and beyond.* PHONUM, reports in phonetics, Umeå University.

Helgason, P. forthc. Stockholm Corpus of Spontaneous Speech. Department of Linguistics, Stockholm University.

Helgason, P. 2002. *Preaspiration in the Nordic Languages, synchronic and diachronic aspects.* Ph.D. thesis, Stockholm University.

Hellspong, L. and Ledin, P. 1997. *Vägar genom texten.* Studentlitteratur.

Hirschberg, J. 1995. Prosodic and other acoustic cues to speaking style in spontaneous and read speech. In *Proceedings of International Congress on Phonetic Sciences*, volume 2, pp. 36–43.

Hirschberg, J. 2000. A Corpus-based Approach to the Study of Speaking Style. In Horne, M., editor, *Prosody: Theory and Experiment, Studies presented to Gösta Bruce*, pp. 335–350. Kluwer Academic Publisher.

Hirschberg, J. and Litman, D. J. 1993. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, **19**(3):501–530.

Hirschberg, J. and Nakatani, C. 1996. A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pp. 286–293, Santa Cruz.

Hobbs, J. 1985. On the Coherence and Structure of Discourse. Technical Report Technical Report CSLI-85-37, CSLI, Stanford.

Horne, M., Hansson, P., Bruce, G., Frid, J., and Philipson., M. 2001. Cue words and the topic structure of spoken discourse: The case of Swedish MEN 'but'. *Journal of Pragmatics*, **33**:1061–1081.

Horne, M. and Philipson, M. 1995. Developing the Prosodic Component for Swedish Speech Synthesis. In *Proceedings of ESCA EUROSPEECH'95, 4th European Conference on Speech Communication and Thechnology*, pp. 611–614, Madrid, Spain.

House, D. 2003. Final rises in spontaneous Swedish computer-directed questions: incidence and function. In *Proceedings of Speech Prosody 2003*, Noura, Japan.

Kamp, H. and Reyle, U. 1993. *From Discourse to Logic.* Studies in Linguistics and Philosophy. Kluwer Academic Publisher, Dordrecht, Boston, London.

Krippendorf, K. 1980. *Content Analysis: An Introduction to its Methodology.* Sage Publications.

Lambrecht, K. 1994. *Information structure and sentence form.* Cambridge Studies in Linguistics. Cambridge University Press.

Landis, R. J. and Koch, G. G. 1977. The Measurement of Observer Agreement fo Categorial Data. *Biometrics*, **33**:159–174.

Leech, G. and Wilson, A. 1996. *EAGLES Recommendations for the Morphosyntactis Annotation of Corpora.* EAGLES , Instituto di Linguistics Computazionale, Pisa.

Lehiste, I. 1979. Perception of Sentence and Paragraph Boundaries. In Lindblom, B. and Öhman, S., editors, *Frontiers of Speech Communication*, pp. 191–201. Academic Press.

Levow, G.-A. 2004. Prosodic Cues to Discourse Segment Boundaries in Human-Computer Dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pp. 93–96, Boston, Massachusetts.

Litman, L. and Allen, J. 1990. Discourse Processing and Commonsense Plans. In R., C. P., Morgan, J., and Pollack, M., editors, *Intentions in Communication*, pp. 365–388. MIT Press.

Loman, B. 1967. Prosodi och Syntax. In Holm, G., editor, *Svenskt talspråk*, pp. 86–100, Stockholm. Almqvist & Wiksell.

Loman, B. and Jörgensen, N. 1971. *Manual för analys av makrosyntagmer.* Studentlitteratur, Lund.

Mann, W. C. and Thompson, S. A. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text*, **8**(3):243–281.

Marcu, D. 1997. *The Rhetorical Parsing, Summarization and Generetion of Natural Language Texts.* Ph.D. thesis, Department of Computer Science, University of Toronto, Toronto, Canada.

Megyesi, B. 2002. *Data-Driven Syntactic Analysis. Methods and applications for Swedish.* Ph.D. thesis, Department of Speech, Music and Hearing, KTH.

Megyesi, B. and Gustafson-Čapková, S. 2001. Pausing in Dialogues and Read Speech: Speaker's Production and Listeners Interpretation. In *Proceedings of the Workshop on Prosody in Speech Recognition and Understanding*, pp. 107–113, NJ, USA.

Megyesi, B. and Gustafson-Čapková, S. 2002. Production and Perception of Pauses and their Linguistic Context in Read and Spontaneous Speech in Swedish. In *Proceedings of ICSLP 2002 - 7th International Conference on Spoken Language Processing*, pp. 2153–2156, Denver, USA, 16-20 September.

Melin, L. and Lange, S. 1986, 2000. *Att analysera text. Stilanalys med exempel.* Studentlitteratur.

Moser, M. and Moore, J. D. 1995. Investigating Cue Selection and Placement in Tutorial Discourse. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, pp. 130–135, Cambridge, MA.

Moser, M. and Moore, J. D. 1996. Toward a Synthesis of Two Accounts of Discourse Structure. *Computational Linguistics*, **22**(3):409–419.

Norrby, C. 2004. *Samtalsanalys. Så gör vi när vi pratar med varandra.* Studentlitteratur, 2 edition. Lund.

Ostendorf, M. 2000. Prosodic Boundary Detection. In Horne, M., editor, *Prosody: Theory and Experiment, Studies presented to Gösta Bruce*, pp. 263–280. Kluwer Academic Publisher.

Passonneau, R. 1993. Getting and Keeping the Centre of Attention. In Bates, M. and Weischedel, R. M., editors, *Challenges in Natural Language Processing*, pp. 179–227. Cambridge University Press.

Passonneau, R. J. and Litman, D. J. 1997. Discourse Segmentation by Human and Automated Means. *Computational Linguistics*, **23**(1):103–139.

Poesio, M. and Vieira, R. 1998. A Corpus-based Investigation of Definite Description Use. *Computational Linguistics*, **24**(2):183–216.

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

Polanyi, L. 1996. The Linguistic Structure of Discourse. Technical Report CSLI-96-200, CSLI.

Polanyi, L. 2001. The Linguistic Structure of Discourse. In Schiffrin, D., Tannen, D., and Hamilton, H., editors, *The Handbook of Discourse Analysis.* Blackwell Publishers.

Prince, E. 1981. Toward a Taxonomy of Given-New Information. In Cole, P., editor, *Radical Pragmatics*, pp. 223–255. Academic Press, New York.

Samuel, K., Carberry, S., and Vijay-Shanker, K. 1998. Dialogue Act Tagging with Transformation-Based Learning. In *Proceedings of the 17th International conference on Computational Linguistics, CoLing*, pp. 1150–1156.

Schegloff, E. A. 1996. Turn Organization: one intersection of grammar and interaction. In Ochs, E., Schegloff, E., and Thompson, S., editors, *Interaction and Grammar*, Studies in Interactional Sociolinguistics, pp. 52–133. Cambridge University Press.

Searle, J. 1969. *Speech Acts.* Cambridge University Press.

Selkirk, E. 1984. *Phonology and Syntax: The Relation between Sound and Syntax.* MIT Press.

Selkirk, E. 2000. The interaction of constraints on prosodic phrasing. In Horne, M., editor, *Prosody: theory and experiment*, Text Speech and Language Technology, pp. 231–263. Kluwer Academic Publishers.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meeter, M., and van Ess-Dykema, C. 1998. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech*, **32**:127–154.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. 2000. Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. *Language and Speech*, **41**(3-4):439–487.

Sidner, C. 1983. Focusing in the comprehension of definite anaphora. In Brady, M. and Berwick, R. C., editors, *Computational Models of Discourse*, pp. 267–330. MIT Press.

Siegel, S. and Castellan, N. J. 1988. *Nonparametric statistics for the behavioral sciences.* McGraw Hill, Boston.

Sinclair, J. and Coulthardt, R. 1975. *Towards an Analysis of Discourse: The English used by teachers and pupils.* Oxford University Press.

Stede, M. and Umbach, C. 1998. DiMLex: a lexicon of discourse markers for text generation and understanding. In *Proceedings of the 17th International Conference on Computational Linguistics*, pp. 1238–1242.

Stifelman, L. J. 1995. A Discourse Analysis Approach to Structured Speech. In *The AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation*, Stanford University.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meeter, M. 2000. Dialogue Act Modeling for Automatic for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, **26**(3):339–373.

Strangert, E. 1990. Pauses, Syntax and Prosody. In Wiik, K. and Raimo, I., editors, *Nordic Prosody*, pp. 294–305.

Strangert, E. 1992. Prosodic cues to the perception of syntactic boundaries. In *Proceedings of ICSLP 1992*, pp. 1283–1285.

Strangert, E. 1993. Speaking style and pausing. In *PHONUM*, pp. 121–137, Reports from the Departments of Phonetics, University of Umeå.

Strangert, E. and Heldner, M. 1995. Labelling of boundaries and prominences by phonetically experienced and non-experienced transcribers. In *PHONUM*, pp. 85–109, Reports from the Departments of Phonetics, University of Umeå.

Swerts, M. 1997. Prosodic Features and Discourse Boundaries of Different Strength. *Journal of the Acoustical Society of America*, **101**(1):514–521.

Swerts, M. and Geluykens, R. 1994. Prosody as a marker of information flow in spoken discourse. *Language and Speech*, **37**:21–45.

Tannen, D. 1982. The Oral – Literate Continuum in Discourse. In Tannen, D., editor, *Spoken and Written Language: Exploring Orality and Literacy*, Advances in Discourse Processes, pp. 1–17, Norwood, New Jersey. Ablex Publishing Corporation.

Teleman, U. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska.* Lund: Studentlitteratur.

Terken, J. and Hirschberg, J. 1994. Deaccentuation of words representing *given* information: Effects of persistence of grammatical function and surface position. *Language and Speech*, **37**:125–45.

Traum, D. 1999. Speech acts for dialogue agents. In Wooldridge, M. and Rao, A., editors, *Foundations of Rational Agency*, volume 2, pp. 169–201. Kluwer.

Traum, D. and Hinkelman, E. 1992. Conversation acts in task-oriented spoken dialogue. *Intelligence*, **8**:575–599.

Walker, M. A. 1996. Limited Attention and Discourse Structure. *Computational Linguistics*, **22**(2):255–264.

Walker, M. A. 1998. Centering, Anaphora Resolution and Discourse Structure. In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering in Discourse*. Oxford University Press.

Webber, B. L. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, **6**(2):107–135.

Webber, B. and Joshi, A. 1998. Anchoring a Lexicalised Tree-Adjoining Grammar for Discourse. In *COLING/ACL Workshop on Discourse Relations and Discourse Markers*, pp. 86–92, Montreal, Quebec, Canada.

Webber, B., Joshi, A., Knott, A., and Stone, M. 1999. What are Little Trees Made Of? A Structural and Presuppositional Account Using Lexicalised TAG. In *Proceedings of International Workshop on Levels of Representation in Discourse (LORID'99)*, pp. 151–156, Edinburgh, UK.

Webber, B. L., Joshi, A., Stone, M., and Knott, A. 2003. Anaphora and Discourse Structure. *Computational Linguistics*, **29**(4):546–587.

Wells, R. 1960. Nominal and verbal style. In Sebeok, T. A., editor, *Style in language*, pp. 213–220, Massachusetts. The MIT Press.

Wennerstrom, A. 2001. Intonational meaning. In Wennerstrom, A., editor, *The Music of Everyday Speech*, pp. 17–46, Oxford. Oxford University Press.

Appendices

Appendix 1

This appendix contains the subjects' instructions for each experiment task and each condition.

I: The boundary annotation task, condition Read.
II: The boundary annotation task, condition Listen.
III: The prominence annotation task, condition Read.
VI: The prominence annotation task, condition Listen.

In order to save space the instructions are presented in a more compact format than was used in the experiments.

I. Instructions to the subjects, boundary annotation task, condition Read

In this experiment condition we have data from 6 subjects. The task for the subjects is to insert punctuation in transcripts of the monologues and dialogues.

Instruktioner

Sofias undersökning av interpunktion

1 Syfte

Hur visar en talare att vissa delar av det som sägs hänger mer ihop än andra delar? Detta experiment går ut på att undersöka hur läsare/försökspersoner upplever meningsstrukturen i olika texttyper av text. Hela experimentet beräknas ta c:a tre och en halv timme.

2 Beskrivning av uppgiften

Du har fått ett antal texter där punkt, komma och andra skiljetecken fattas. ostrukturerade texter (eller transkriptioner av tal). Din uppgift är att på grundval av orden i texterna, d.v.s. genomläsning av texterna, sätta in interpunktion: , d.v.s. punkt, komma, frågetecken, nytt stycke etc.

3 Beskrivning av materialet

Materialet för denna undersökning består av flera olika typer av textbaserat eller spontant tal, där du har tillgång till nedteckningar av talet, d.v.s. transkriptioner av talet. De transkriptioner som är gjorda av det upplästa talet är mer lika traditionell text, medan det transkriberade spontantalet är mer olikt traditionell text. Materialet kan delas i två typer:
Monolog: Skriftbaserad och spontan monolog
Dialog: Skriftbaserad och spontan dialog
De olika talstilarna är litet olika transkriberade, p.g.a. att förutsättningarna är olika., t.ex. så innehåller det spontana talet oavslutade meningar/meningsfragment och talspråkliga uttryck. Nedan följer en kort beskrivning av de olika taltyperna som du kommer att träffa på, samt en beskrivning av hur de är transkriberade.

3.1 Beskrivning av monologerna

Monologerna är uppdelade i en skriftbaserad och en spontan del. Den skriftbaserade delen består av upplästa nyheter (Dagens Eko) samt en tidningsartikel (Dagens Nyheter), och den spontana delen består av en återberättad nyhetsartikel (Dagens Nyheter). Båda transkriptionerna har hummanden och talspråkliga uttryck införda.

3.2 Beskrivning av dialogerna

Även dialogerna består av en skriftbaserad och en spontan del. Den skriftbaserade dialogen består av ett utsnitt ur en pjäs från Radioteatern ("Själen och Sankte Per" av Beate Grimsrud). Den spontana dialogen är strukturerad i turer, och hummanden och talspråkliga uttryck förts in. Spontandialogen är en s.k. Map Task-dialog, en kartuppgift. Kartuppgifter används ofta för att framkalla spontantal kring en speciell uppgift och går till på detta vis. Två talare sitter så att de inte kan se varandra och de har varsin karta. Den ena talaren är "instruktör" och den andra talaren är "följare". Instruktören har en väg utritad på sin karta, men följarens karta har det inte. Instruktörens uppgift är att ge följaren detaljerade instruktioner om vägen, så att följaren kan rita in vägen exakt på sin karta. För att göra dialogen mer dubbelriktad har kartorna gjorts något olika, så att diskussion ska uppstå om den exakta placeringen och utformningen av olika landmärken på kartan.

3.3 Översikt över de filer som ska annoteras

I tabellen på omstående sida ser du vad de olika transkriptionerna heter. Detta namn återfinner du i filhuvudet på respektive transkription. Du ser också av vilken typ transkriptionen är. Observera att du bara får ett utsnitt ur varje material. Det innebär att dialogerna och den upplästa tidningsartikeln avslutas något abrupt. I den upplästa tidningstexten är intervjuavsnitt bortredigerade, så det finns en del abrupta övergångar även där.

x-log	text/spontan	transkription
Monolog	Textbaserad	nyheter1.doc Upplästa nyheter
		nyheter2.doc Upplästa nyheter
		beraetta-text.doc Uppläst tidningsartikel
	spontan	beraetta-tal-a.doc Återberättad tidningsartikel
		beraetta-tal-b.doc Återberättad tidningsartikel
		beraetta-tal-c.doc Återberättad tidningsartikel
Dialog	textbaserad	radioteater.doc
	spontan	dialog1.doc Kart-dialog
		dialog2.doc Kart-dialog
		dialog3.doc Kart-dialog
		dialog4.doc Kart-dialog

Tabell 1 visar en översikt över de transkriberade filerna. (Filnamnen står högst upp i hörnet på de pappersko-pior du fått).Texterna som skall annoteras finns i 4 filer:

Du får dem på papper, men respektive namn kommer att finnas i filhuvudena!

4 Gör såhär:

Läs papprena med transkriptionerna och sätt in interpunktion så att transkriptionerna blir läsliga. Det enklaste är att arbeta med blyertspenna, så att du kan sudda om du ångrar dig. Använd dig av de skiljetec-ken/interpunktionshjälpmedel som du behöver.

Vad gäller dialogerna så är de redan strukturerade i talarturer, d.v.s. du ser om det blir ett talarbyte (t.ex. talare B talar istället för talare A). Detta är gjort eftersom jag bedömde att det skulle bli alltför svårtolkat att få hela dialogen i en klump, utan information om talare.

Använd // (dubbelt snedstreck) för att markera nytt stycke. I dialog blir "nytt stycke" litet malplacerat, markera därför istället "nytt ämne", men använd fortfarande //. Glöm inte att markera "nytt ämne" även vid talarbyte, om du tycker att det är en ämnesgräns där. Om du inte markerar "nytt ämne" vid talarbyte kommer det att tolkas som om båda båda talarnas bidrag ligger inom samma ämne – vilket naturligtvis ibland är fallet.

Försök att använda en större uppsättning av skiljetecken än bara punkt och komma. På omstående sida finns en Kom-ihåg-tabell över skiljetecken (jag tror inte att du kommer att behöva några andra tecken än de som finns i tabellen):
Dessa skiljetecken är i stort sett vad du kommer att behöva använda!

komma	,
punkt	.
frågetecken	?
utropstecken	!
semikolon	;
kolon	:
parentes	()
tankestreck	–
anföringstecken	“ “
tre punkter	...
nytt stycke	//

Nedan ser du exempel på hur interpunktionen kan föras in. I Exempel 1 ser du den ursprungliga transkriptionen, och i Exempel 2 ser du hur den ser ut efter det att skiljetecken satts in. Som du ser så har jag också fört in stor bokstav i meningsbörjan och vid namn, det kan du göra om du vill (bara skriv över den lilla bokstaven med en stor), men om du tycker att det är lättare att fokusera på interpunktionen om du struntar i stor bokstav så är det också ok!

Exempel 1. **FÖRE!**

```
<Talare A> maria olsson och det är den tjugoförsta mars nittonhundra nittio sju
<Talare B> ja ska jag säga det samma eller eller det vill säga motsvarand lars
andersson och det är väl fortfarande den tjugoförsta mars då i så fall
<Talare A> då ska vi se då har vi en en s karta här framför oss och jag har
landstigit på en plats på den här ön och det börjar vid en en in i en bukt en ganska
ovalt formad bukt inne på västra sydvästra sidan utav den här ön har du den också
<Talare B> jo då eee och tydligen så är ju formen på våra öar identiska så att så att
själva den bukten ska det inte vara några problem att hitta
```

Exempel 2. **EFTER!**

```
<Talare A> Maria Olsson. Och det är den tjugoförsta mars, nittonhundra nittio sju.
<Talare B> Ja, ska jag säga det samma? Eller, eller det vill säga motsvarande. Lars
Andersson, och det är väl fortfarande den tjugoförsta mars då i så fall.//
<Talare A> Då ska vi se, då har vi en en s karta här framför oss, och jag har
landstigit på en plats på den här ön. Och det börjar vid en en in i en bukt, en
ganska ovalt formad bukt, inne på västra sydvästra sidan utav den här ön. Har du den
också?
<Talare B> Jo då! Eee, och tydligen så är ju formen på våra öar identiska, så att så
att själva den bukten ska det inte vara några problem att hitta.
```

Gör texterna i den ordning som står i tabellen, d.v.s. börja med monologerna och sluta med dialogerna.

4.1 Vad gör jag när de pratar i munnen på varandra?

I de spontana dialogerna förekommer ibland överlappande tal. Det som sägs i dessa avsnitt är transkriberat på separata rader, men i verkligheten talar de båda talarna samtidigt. Ett exempel på hur överlappande tal är transkriberat ser du i exempel 3 (för tydlighetens skull är talarturerna, d.v.s. talarnas olika bidrag till dialogen, numrerade).

Exempel 3

```
Tur 1. <Talare B> jaha
Tur 2. <Talare A> det är rätt långt ner i norr
Tur 3. <Talare B> i sydvästra hörnet
```

```
Tur 4. <Talare A> no alltså nord nordväst
Tur 5. <Talare B> nordväst
Tur 6. <Talare A> sydväst <skratt>
```

Turerna 3 och 4 är överlappande, d.v.s. de sägs samtidigt. Det innebär att talare B egentligen säger tur 3 och tur 5 i ett svep. Det säger sig själv att det inte är det allra lättaste att sätta ut skiljetecken i detta, eftersom vanlig skrift är ett trubbigt verktyg för att representera överlappande tal. Jag vill ändå uppmuntra dig till att göra så bra du kan och skriv gärna en kommentar där du finner det helt omöjligt.

4.2 Jag kan inte bestämma mig för hur jag ska göra! Vad gör jag då?
Ibland kan man ha två eller fler alternativ, och inte veta riktigt vilket man ska välja; hur man än gör så blir man inte helt nöjd. Du kan markera såna passager med genom att sätta dem inom hakparenteser:
[när texten står inom hakparenteser, som här, så betyder det att jag är litet osäker]
Då ser jag att du tyckte att just detta stället var svårt, och du behöver inte sitta alltför länge och tveka om hur du ska göra.
Om det känns fullkomligt omöjligt är du alltid varmt välkommen att fråga mig om du undrar över något. Du kan skriva mail (sofia@ling.su.se), ringa (08/16 17 61) eller komma förbi mitt rum (C348) och fråga!!

När du är färdig lämnar du in dina papperskopior med markeringar i, antingen i mitt fack eller på mitt rum. Därefter får du dina biocheckar av mig, och den mer avkopplande delen av experimentet - för dig :) - börjar!

Du är alltid varmt välkommen att fråga mig om du undrar över något. Du kan skriva mail (sofia@ling.su.se), ringa (08/16 17 61) eller komma förbi mitt rum (C348) och fråga!!

Tack så hemskt mycket och lycka till!

Sofia

Sofia Gustafson-Capková
Institutionen för lingvistik
Stockholms universitet
106 91 Stockholm
tel: 08/16 17 61
e-post: sofia@ling.su.se
rum: C348
web: www.ling.su.se/staff/sofia

II. Instructions to the subjects, boundary annotation task, condition Listen

In this experiment condition we have data from 8 subjects. The task for the subjects is to insert punctuation in the transcripts of the monologues and dialogues with access to the speech signal.

Instruktioner

Sofias undersökning av interpunktion

1 Syfte

Hur visar en talare att vissa delar av det som sägs hänger mer ihop än andra delar? Detta experiment går ut på att undersöka hur lyssnare upplever meningsstrukturen i olika typer av tal. Hela experimentet beräknas ta c:a tre timmar.

2 Beskrivning av uppgiften

Du har fått inspelningar av olika slags tal samt nedteckningar (d.v.s. transkriptioner) av detta tal. I dessa transkriptioner finns dock inga skiljetecken, som punkt, komma etc, utsatta. Din uppgift är att genom att lyssna på ljudfilerna sätta in interpunktion, d.v.s. punkt, komma, frågetecken, nytt stycke etc. i transkriptionerna. (Ljudfilerna är tillsammans ungefär en timme, så den beräknade tiden, tre timmar, för experimentet inkluderar en del omlyssning!)

3 Beskrivning av materialet

Du har fått en CD med ljudfiler och transkriptioner av ljudfilerna på papper. CD:n är en data-CD, det innebär att du måste lyssna på den i dator, och inte i CD-spelare.

Materialet för denna undersökning består av flera olika typer av textbaserat eller spontant tal, där du har tillgång till transkriptioner av tal. De transkriptioner som är gjorda av det upplästa talet är mer lika traditionell text, medan det transkriberade spontantalet är mer olikt traditionell text.

De olika talstilarna är olika transkriberade, p.g.a. att förutsättningarna är olika, t.ex. så innehåller spontantal oavslutade meningar/meningsfragment och talspråkliga uttryck. Nedan följer en kort beskrivning av de olika taltyperna, samt en beskrivning av hur de är transkriberade.

3.1 Beskrivning av monologerna

Monologerna är uppdelade i en skriftbaserad och en spontan del. Den skriftbaserade delen består av upplästa nyheter (Dagens Eko) samt en tidningsartikel (Dagens Nyheter), och den spontana delen består av en återberättad nyhetsartikel (Dagens Nyheter). Båda transkriptionerna har hummanden och talspråkliga uttryck införda.

3.2 Beskrivning av dialogerna

Även dialogerna består av en skriftbaserad och en spontan del. Den skriftbaserade dialogen består av ett utsnitt ur en pjäs från Radioteatern ("Själen och Sankte Per" av Beate Grimsrud). Dialogen är strukturerad i turer.

Den spontana dialogen är strukturerad i turer, och hummanden och talspråkliga uttryck förts in. Spontandialogen är en s.k. Map Task-dialog, en kartuppgift. Kartuppgifter används ofta för att framkalla spontantal kring en speciell uppgift och går till på detta vis. Två talare sitter så att de inte kan se varandra och de har varsin karta. Den ena talaren är "instruktör" och den andra talaren är "följare". Instruktören har en väg utritad på sin karta, men följarens karta har det inte. Instruktörens uppgift är att ge följaren detaljerade instruktioner om vägen, så att följaren kan rita in vägen exakt på sin karta. För att göra dialogen mer dubbelriktad har

kartorna gjorts något olika, så att diskussion ska uppstå om den exakta placeringen och utformningen av olika landmärken på kartan.

3.3 Översikt över de filer som ska annoteras

Eftersom ljudfilerna är uppdelade i kortare avsnitt, för att det ska bli litet mer hanterligt, så består materialet av rätt många ljudfiler med tillhörande transkriptioner. På omstående sida ser du allt uppställt i en tabell. Där ser du vad de olika transkriptionerna heter. Detta namn återfinner du i filhuvudet på respektive transkription. I tabellen ser du också av vilken typ transkriptionen är.

Observera att du bara får ett utsnitt ur varje material. Det innebär att dialogerna och den upplästa tidningsartikeln avslutas något abrupt. I den upplästa tidningstexten är intervjuavsnitt bortredigerade, så det finns en del abrupta övergångar även där.

x-log	text/spontan	transkription	ljud
Monolog	Textbaserad	nyheter1.doc Upplästa nyheter	110-nyheter.wav 111-nyheter.wav 112-nyheter.wav
		nyheter2.doc Upplästa nyheter	113-nyheter.wav 114-nyheter.wav 115-nyheter.wav
		beraetta-text.doc Uppläst tidningsartikel	116-dn.wav 117-dn.wav 118-dn.wav
	spontan	beraetta-tal-a.doc Återberättad tidningsartikel	119-spontan.wav
		beraetta-tal-b.doc Återberättad tidningsartikel	120-spontan.wav
		beraetta-tal-c.doc Återberättad tidningsartikel	121-spontan.wav
Dialog	textbaserad	radioteater.doc	210-radioteater.wav
	spontan	dialog1.doc Kart-dialog	211-dialog.wav
		dialog2.doc Kart-dialog	212-dialog.wav
		dialog3.doc Kart-dialog	213-dialog.wav
		dialog4.doc Kart-dialog	214-dialog.wav

4 Gör såhär:

Lyssna till ljudfilerna och sätt in interpunktion i de transkriptioner du fått som papperskopior. Sträva efter att interpunktionen motsvarar det du hör, d.v.s. lita mer på dina öron, än på rigida interpunktionsregler. Arbeta gärna med blyerts, så kan du sudda om det skulle behövas. Använd dig av de skiljetecken som du behöver.

Vad gäller dialogerna så är de redan strukturerade i talarturer, d.v.s. du ser i transkriptionerna om det blir ett talarbyte (t.ex. talare B talar istället för talare A).

Använd // (dubbelt snedstreck) för att markera nytt stycke. I dialog blir "nytt stycke" litet malplacerat, markera därför istället "nytt ämne", men använd fortfarande //. Glöm inte att markera "nytt ämne" även vid talarbyte, om du tycker att det är en ämnesgräns där. Om du inte markerar "nytt ämne" vid talarbyte kommer det att tolkas som om båda båda talarnas bidrag ligger inom samma stycke – vilket naturligtvis ibland är fallet.

Försök att använda en större uppsättning av skiljetecken än bara punkt och komma. Nedan finns en Kom-ihåg-tabell över skiljetecken (jag tror inte att du kommer att behöva några andra tecken än de som finns i tabellen).

komma	,
punkt	.
frågetecken	?
utropstecken	!
semikolon	;
kolon	:
parentes	()
tankestreck	–
anföringstecken	“ “
tre punkter	...
nytt stycke	//

Nedan ser du exempel på hur interpunktionen kan föras in. I Exempel 1 ser du den ursprungliga transkriptionen, och i Exempel 2 ser du hur den ser ut efter det att skiljetecken satts in. Som de ser så har jag också fört in stor bokstav i meningsbörjan och vid namn, det kan du göra om du vill (bara skriv över den lilla bokstaven med en stor), men om du tycker att det är lättare att fokusera på interpunktionen om du struntar i stor bokstav så är det också ok! (Exemplet är från allra första början på en spontan dialog, så den inleds med att talarna presenterar sig, och därefter börjar de med kartuppgiften. Observera att den allra första snutten “ presentationen “ inte finns med i ljudfilerna, detta p.g.a. anonymisering av inspelningen).

Exempel 1. **FÖRE!**

```
<Talare A> maria olsson och det är den tjugoförsta mars nittonhundra nittio sju
<Talare B> ja ska jag säga det samma eller eller det vill säga motsvarand lars
andersson och det är väl fortfarande den tjugoförsta mars då i så fall
<Talare A> då ska vi se då har vi en en s karta här framför oss och jag har
landstigit på en plats på den här ön och det börjar vid en en in i en bukt en ganska
ovalt formad bukt inne på västra sydvästra sidan utav den här ön har du den också
<Talare B> jo då eee och tydligen så är ju formen på våra öar identiska så att så att
själva den bukten ska det inte vara några problem att hitta
```

Exempel 2. **EFTER!**

```
<Talare A> Maria Olsson. Och det är den tjugoförsta mars, nittonhundra nittio sju.
<Talare B> Ja, ska jag säga det samma..? Eller, eller... det vill säga motsvarande.
Lars Andersson, och det är väl fortfarande den tjugoförsta mars då i så fall.//
<Talare A> Då ska vi se, då har vi en... en s karta här framför oss, och jag har eee
landstigit på en plats på den här ön. Och det börjar vid en... en... in i en bukt, en
ganska ovalt formad bukt, inne på västra, sydvästra sidan utav den här ön. Har du den
också?
<Talare B> Jo då! Eee och tydligen så är ju formen på våra öar identiska, så att så
att själva den bukten ska det inte vara några problem att hitta.
```

Gör texterna i den ordning som står i tabellen, d.v.s. börja med monologerna och sluta med dialogerna. När du är färdig lämnar du transkriptionerna i mitt fack i postrummet – eller kom förbi och lämna det i mitt rum!

4.1 Vad gör jag när de pratar i munnen på varandra?

I de spontana dialogerna förekommer ibland överlappande tal. Det som sägs i dessa avsnitt är transkriberat

på separata rader, men i verkligheten talar de båda talarna samtidigt. Ett exempel på hur överlappande tal är transkriberat ser du i exempel 3 (för tydlighetens skull är talarturerna, d.v.s. talarnas olika bidrag till dialogen, numrerade).

Exempel 3
```
Tur 1. <Talare B> jaha
Tur 2. <Talare A> det är rätt långt ner i norr
Tur 3. <Talare B> i sydvästra hörnet
Tur 4. <Talare A> no alltså nord nordväst
Tur 5. <Talare B> nordväst
Tur 6. <Talare A> sydväst <skratt>
```

Turerna 3 och 4 är överlappande. Det innebär att talare B egentligen säger tur 3 och tur 5 i ett svep. Det säger sig själv att det inte är det allra lättaste att sätta ut skiljetecken i detta, eftersom vanlig skrift är ett trubbigt verktyg för att representera överlappande tal. Jag vill ändå uppmuntra dig till att göra så bra du kan – och skriv gärna en kommentar där du finner det helt omöjligt.

4.2 Jag kan inte bestämma mig för hur jag ska göra! Vad gör jag då?
Ibland kan man ha två eller fler alternativ, och inte veta riktigt vilket man ska välja; hur man än gör så blir man inte helt nöjd. Du kan markera såna passager med genom att sätta dem inom hakparenteser:

[när texten står inom hakparenteser, som här, så betyder det att jag är litet osäker]

Då ser jag att du tyckte att just detta stället var svårt, och du behöver inte sitta alltför länge och tveka om hur du ska göra. Om det känns fullkomligt omöjligt är du alltid varmt välkommen att fråga mig om du undrar över något. Du kan skriva mail (sofia@ling.su.se), ringa (08/16 17 61) eller komma förbi mitt rum (C348) och fråga!!
När du är färdig lämnar du in dina papperskopior med markeringar i, antingen i mitt fack eller på mitt rum. Därefter får du dina biocheckar av mig, och den mer avkopplande delen av experimentet - för dig :) - börjar!

Tack så hemskt mycket och lycka till!

Sofia

Sofia Gustafson-Capková
Institutionen för lingvistik
Stockholms universitet
106 91 Stockholm
tel: 08/16 17 61
e-post: sofia@ling.su.se
rum: C348
web: www.ling.su.se/staff/sofia

III. Instructions to the subjects, prominence annotation task, condition Read

In this experiment condition we have data from 10 subjects. The task for the subjects is to insert prominence markings in the transcripts of the monologues and dialogues.

Instruktioner

Sofias läsundersökning

1 Syfte

Vad är det egentligen som blir betonat? Syftet med detta experiment är att undersöka hur vi förväntar oss att intonationen realiseras i olika talstilar. Beräknad tid för experimentet är ca tre timmar.

2 Beskrivning av uppgiften

Du får ett antal nedteckningar av tal, s.k. transkriptioner. Din uppgift är, att på grundval av dessa transkriptioner, stryka under de ord eller fraser du tror skulle vara mest betonade i talet. Mer betonade innebär ord eller fraser som blir mer framhävda genom att de uttalas med högre volym, intensitet eller att de förlängs. Det handlar om att markera de ord eller fraser som är mest framträdande, det är alltså inte nödvändigt att det är ett enda ord som är mest framträdande, utan det kan kanske vara två eller fler. Exempel på denna skillnad visas i exempel 1 (fet stil indikerar starkare betoning):

Exempel 1
Den röda **bilen**. (Ett ord framhävt)
Den **röda bilen**. (En fras framhävd)

Naturligtvis skulle man också kunna tänka sig andra alternativ som Den **röda** bilen"eller "**Den röda** bilen". Det är inte helt lätt att avgöra i relation till vad som ett ord eller en fras är mer framträdande. Som riktlinje kan man välja en enhet stor som en mening ungefär, och sedan försöka markera vad som är mest framträdande i hela meningen. Man måste dock komma ihåg att detta bara är ett riktmärke eftersom det finns meningar som kan innehålla flera framträdande ord. En del av transkriptionerna består av spontantal, där korrekta meningar, såna som vi är vana att se dem i text, inte alltid går att hitta. Transkriptionen av spontantalet är dessutom gjord utan skiljetecken som punkt, komma och frågetecken etc. Detta gör att det kan vara svårt att komma åt enheten "en mening". Använd alltså omfånget meningsom ett riktmärke, och inte som en tvingande regel, och lita till din egen intuition vad gäller mer eller mindre framträdande ord och fraser.

3 Beskrivning av materialet

Du får allt material som transkriptioner på papper. Materialet för denna undersökning består av flera olika typer av tal. De olika talstilarna är olika transkriberade, p.g.a. att förutsättningarna är olika, t.ex. så innehåller spontantal oavslutade meningar/meningsfragment och talspråkliga uttryck. Detta gör att det transkriberade spontantalet inte ser ut som vanlig text. Nedan följer en kort beskrivning av de olika taltyperna, samt en beskrivning av hur de är transkriberade.

3.1 Monologerna

Monologerna är uppdelade i en skriftbaserad och en spontan del. Den skriftbaserade delen består av upplästa nyheter (Dagens Eko) samt en tidningsartikel (Dagens Nyheter), och den spontana delen består av en återberättad nyhetsartikel (Dagens Nyheter). Eftersom de upplästa nyheterna är baserade på text är även dina transkriptioner strukturerade med skiljetecken i traditionella meningar.
Den återberättade tidningsartikeln är strukturerad i styckeliknande avsnitt. Hummanden och tvekljud har först in i båda transkriptionerna.

3.2 Dialogerna
Även dialogerna består av en skriftbaserad och en spontan del. Den skriftbaserade dialogen består av ett utsnitt ur en pjäs från Radioteatern. Här har du tillgång till text som är ortografiskt strukturerad på vanligt sätt. Den spontana dialogen är strukturerad i turer, och även här har hummanden och talspråkliga uttryck förts in. Spontandialogen är en s.k. Map Task-dialog, en kartuppgift. Kartuppgifter används ofta för att framkalla spontantal kring en speciell uppgift och går till på detta vis. Två talare sitter så att de inte kan se varandra och de har varsin karta. Den ena talaren är "instruktör" och den andra talaren är "följare". Instruktören har en väg utritad på sin karta, men följarens karta har det inte. Instruktörens uppgift är att ge följaren detaljerade instruktioner om vägen, så att följaren kan rita in vägen exakt på sin karta. För att göra dialogen mer dubbelriktad har kartorna gjorts något olika, så att diskussion ska uppstå om den exakta placeringen och utformningen av olika landmärken på kartan.

3.2.1 Ett par ord om dialogstruktur
I dialog turas talarna om att tala. Varje sådan omgång kallas en "tur". Varje tur behöver inte innehålla något framhävt ord/fras. Ett enstaka kort "mmm" eller "ja" utgör en tur, men det behöver inte vara framhävt. Likaväl som en tur inte behöver innehålla något framhävt ord/fras, kan den innehålla ett eller flera framhävda ord/fraser (jämför med vad som sades om meningar).

4 Exempel på annotering (ur kartuppgift-dialogen)
Ett exempel på hur en annotering skulle kunna se ut ser du i exempel 2. Utsnittet är från början av Dialog1. Du ser också exempel på turer.

Exempel 2 `Tur 1 <Talare A> maria olsson och det är den tjugoförsta mars nittonhundra nittio sju`
`Tur 2 <Talare B> ja ska jag säga det samma eller eller det vill säga motsvarand lars andersson och det är väl fortfarande den tjugoförsta mars då i så fall`
`Tur 3 <Talare A> då ska vi se då har vi en en s karta här framför oss och jag har landstigit på en plats på den här ön och det börjar vid en en in i en bukt en ganska ovalt formad bukt inne på västra sydvästra sidan utav den här ön har du den också`
`Tur 4 <Talare B> jo då och tydligen så är ju formen på våra öar identiska`

När du stryker under kan du också stryka under halva ord, om träffar på gränsfall, d.v.s. om du anser att framhävningen börjar mitt i ett visst ord. Observera också att turerna inte är numrerade i det transkriberade materialet. Tänk också på att det finns överlappande tal i vissa delar av transkriptionen. Det som sägs i dessa avsnitt är transkriberat på separata rader, men i verkligheten talar de båda talarna samtidigt. Ett exempel på överlappande tal ser du i exempel 3 (för tydlighetens skull är turerna numrerade även här).

Exempel 3 Tur 1. <Talare B jaha
Tur 2. <Talare A det är rätt långt ner i norr
Tur 3. <Talare B i sydvästra hörnet
Tur 4. <Talare A no alltså nord nordväst
Tur 5. <Talare B nordväst
Tur 6. <Talare A sydväst <skratt>

Turerna 3 och 4 är överlappande. Det innebär att talare A egentligen säger tur 3 och tur 5 i ett svep. Det kan vara bra att påminna sig, eftersom det ibland inte tycks finns något framhävt ord/fras i vissa korta turer.

5 Översikt over materialet
I Tabell 1 ser du vad de olika transkriptionerna heter. Detta namn återfinner du i filhuvudet på respektive transkription. Du ser också av vilken typ transkriptionen är. Observera att du bara får ett utsnitt ur varje material. Det innebär att dialogerna och den upplästa tidningsartikeln avslutas något abrupt. I den upplästa tidningstexten är intervjuavsnitt bortredigerade, så det finns en del abrupta övergångar även där.

x-log	text/spontan	transkription
Monolog	Textbaserad	nyheter1.doc Upplästa nyheter
		nyheter2.doc Upplästa nyheter
		beraetta-text.doc Uppläst tidningsartikel
	spontan	beraetta-tal-a.doc Återberättad tidningsartikel
		beraetta-tal-b.doc Återberättad tidningsartikel
		beraetta-tal-c.doc Återberättad tidningsartikel
Dialog	textbaserad	radioteater.doc
	spontan	dialog1.doc Kart-dialog
		dialog2.doc Kart-dialog
		dialog3.doc Kart-dialog
		dialog4.doc Kart-dialog

6 Gör såhär:

Det enklaste sättet att göra sig en uppfattning om vad som är mer framhävt än något annat är att läsa texterna högt (eller halvhögt) för dig själv. Du stryker sedan löpande under de ord i transkriptionerna som du betonade mer än andra. Tänk bara på att det finns många sätt att betona olika texter, och i detta fallet är det är inget som är mer rätt än det andra, så ta bara det du tycker är bäst, och bekymra dig inte om att det finns alternativ – för det finns det! Använd gärna blyertspenna, så kan du sudda utan problem om du vill ändra något! x Gör texterna i den ordning som står i tabellen, d.v.s. börja med monologerna och sluta med dialogerna.

7 Efteråt...

När du är klar lämnar du in transkriptionerna med dina markeringar i mitt fack eller i mitt rum. Därefter får du dina biocheckar, och den mer avkopplande delen av experimentet - för dig :) - börjar!

Om du undrar över något kan du alltid fråga mig, gärna via mail till sofia@ling.su.se

Lycka till!

Sofia

Sofia Gustafson-Capková
Institutionen för lingvistik
Stockholms universitet
106 91 Stockholm
tel: 08/16 17 61
e-post: sofia@ling.su.se
rum: C348
web: www.ling.su.se/staff/sofia

IV. Instructions to the subjects, prominence annotation task, condition Listen

In this experiment condition we have data from 10 subjects. The task for the subjects is to insert prominence markings in the transcripts of the monologues and the dialogues.

Instruktioner

Sofias lyssningsexperiment

1 Syfte

Detta experiment går ut på att undersöka vilka ord eller fraser som lyssnare uppfattar som mer framträdande i olika talstilar.

2 Beskrivning av uppgiften

Din uppgift är att lyssna igenom ett antal ljudfiler och markera de ord eller fraser du tycker är mer framträdande än andra. Mer framträdande"innebär att du uppfattar dem som mer betonade, d.v.s. talaren lägger mer vikt och intensitet vid dessa ord eller fraser. Beräknad tid för experimentet är tre timmar. Ljudmaterialet omfattar c:a en timme, så en beräknad tid om tre timmar ger utrymme för en del omlyssning. Det handlar om att markera de ord eller fraser som är mest framträdande, det är alltså inte nödvändigt att det är ett enda ord som är mest framträdande, utan det kan kanske vara två eller fler. Exempel på denna skillnad visas i exempel 1 (fet stil indikerar starkare betoning):

Exempel 1 Den röda **bilen**. (Ett ord framhävt)
Den **röda bilen**. (En fras framhävd)

Naturligtvis skulle man också kunna tänka sig andra alternativ som Den **röda** bilen"eller "**Den röda bilen**".
Det är inte helt lätt att avgöra i relation till vad som ett ord eller en fras är mer framträdande. Som riktlinje kan man välja en enhet stor som en mening ungefär, och sedan försöka markera vad som är mest framträdande i hela meningen. Man måste dock komma ihåg att detta bara är ett riktmärke eftersom det finns meningar som kan innehålla flera framträdande ord. En del av ljudfilerna består av spontantal, där korrekta meningar, såna som vi är vana att se dem i text, inte alltid går att hitta. Transkriptionen av spontantalet är dessutom gjord utan skiljetecken som punkt, komma och frågetecken etc. Detta gör att det kan vara svårt att komma åt enheten "en mening". Använd alltså omfånget meningsom ett riktmärke, och inte som en tvingande regel, och lita till dina egna öron vad gäller mer eller mindre framträdande ord och fraser.

3 Beskrivning av materialet

Du får materialet som ljudfiler på en CD samt som nedteckningar av talet från ljudfilerna, d.v.s. transkriptioner - på papper. Materialet för denna undersökning består av flera olika typer av tal. De olika talstilarna är olika transkriberade, p.g.a. att förutsättningarna är olika, t.ex. så innehåller spontantal oavslutade meningar/meningsfragment och talspråkliga uttryck. Detta gör att det transkriberade spontantalet inte ser ut som vanlig text. Nedan följer en kort beskrivning av de olika taltyperna, samt en beskrivning av hur de är transkriberade.

3.1 Monologerna

Monologerna är uppdelade i en skriftbaserad och en spontan del. Den skriftbaserade delen består av upplästa nyheter (Dagens Eko) samt en tidningsartikel (Dagens Nyheter), och den spontana delen består av en återberättad nyhetsartikel (Dagens Nyheter). Eftersom de upplästa nyheterna är baserade på text är även dina transkriptioner strukturerade med skiljetecken i traditionella meningar. Den återberättade tidningsartikeln är strukturerad i styckeliknande avsnitt. Hummanden och tvekljud har först in i båda transkriptionerna.

3.2 Dialogerna

Även dialogerna består av en skriftbaserad och en spontan del. Den skriftbaserade dialogen består av ett utsnitt ur en pjäs från Radioteatern ("Själen och Sankte Per" av Beate Grimsrud). Här har du texten ortografiskt strukturerad på vanligt sätt. Den spontana dialogen är strukturerad i turer, och även här har hummanden och talspråkliga uttryck förts in. Spontandialogen är en s.k. Map Task-dialog, en kartuppgift. Kartuppgifter används ofta för att framkalla spontantal kring en speciell uppgift och går till på detta vis. Två talare sitter så att de inte kan se varandra och de har varsin karta. Den ena talaren är "instruktör" och den andra talaren är "följare". Instruktören har en väg utritad på sin karta, men följarens karta har det inte. Instruktörens uppgift är att ge följaren detaljerade instruktioner om vägen, så att följaren kan rita in vägen exakt på sin karta. För att göra dialogen mer dubbelriktad har kartorna gjorts något olika, så att diskussion ska uppstå om den exakta placeringen och utformningen av olika landmärken på kartan.

3.2.1 Ett par ord om dialogstruktur

I dialog turas talarna om att tala. Varje sådan omgång kallas en "tur". Varje tur behöver inte innehålla något framhävt ord/fras. Ett enstaka kort "mmm" eller "ja" utgör en tur, men det behöver inte vara framhävt. Likaväl som en tur inte behöver innehålla något framhävt ord/fras, kan den innehålla ett eller flera framhävda ord/fraser.

4 Exempel på annotering (ur kartuppgift-dialogen)

I exempel 2 ser du hur annoteringen skulle kunna se ut. Annoteringen är gjord på dialogmaterialet, början av fil dialog1.wav. Du ser också exempel på turer.

Exempel 2

```
Tur 1 <Talare A> maria olsson och det är den tjugoförsta mars nittonhundra nittio sju
Tur 2 <Talare B> ja ska jag säga detsamma eller eller det vill säga motsvarande lars
andersson och det är väl fortfarande den tjugoförsta mars då i så fall
Tur 3 <Talare A> då ska vi se då har vi en en s karta här framför oss och jag har
landstigit på en plats på den här ön och det börjar vid en en in i en bukt en ganska
ovalt formad bukt inne på västra sydvästra sidan utav den här ön har du den också
Tur 4 <Talare B> jodå och tydligen så är ju formen på våra öar identiska
```

När du stryker under kan du också stryka under halva ord, om träffar på gränsfall, d.v.s. om du anser att framhävningen börjar mitt i ett visst ord. Observera också att turerna inte är numrerade i det transkriberade materialet.

4.1 Vad gör jag när de pratar i munnen på varandra?

I de spontana dialogerna förekommer ibland överlappande tal. Det som sägs i dessa avsnitt är transkriberat på separata rader, men i verkligheten talar de båda talarna samtidigt. Ett exempel på hur överlappande tal är transkriberat ser du i exempel 3 (för tydlighetens skull är talarturerna, d.v.s. talarnas olika bidrag till dialogen, numrerade).

Exempel 3

```
Tur 1. <Talare B> jaha
Tur 2. <Talare A> det är rätt långt ner i norr
Tur 3. <Talare B> i sydvästra hörnet
Tur 4. <Talare A> no alltså nord nordväst
Tur 5. <Talare B> nordväst
Tur 6. <Talare A> sydväst <skratt>
```

Turerna 3 och 4 är överlappande. Det innebär att talare A egentligen säger tur 3 och tur 5 i ett svep. Det kan vara bra att påminna sig, eftersom det ibland inte tycks finns något framhävt ord/fras i vissa korta turer.

5 Översikt över materialet

För att ljudfilerna ska bli mer lätthanterliga har jag delat upp dem i kortare filer. Det innebär att det kan se mer ut än det egentligen är. För varje ljudfil finns en separat textfil med transkription. Du ser i tabell 1 vilka ljudfiler som hör ihop med vilka transkriptioner. Transkriptionerna har dessutom respektive filnamn i filhuvudet. Observera att du bara får ett utsnitt ur varje material. Det innebär att dialogerna och den upplästa

tidningsartikeln avslutas något abrupt. I den upplästa tidningstexten är intervjuavsnitt bortredigerade, så det finns en del abrupta övergångar även där.

x-log	text/spontan	transkription	ljud
Monolog	Textbaserad	nyheter1.doc Upplästa nyheter	110-nyheter.wav 111-nyheter.wav 112-nyheter.wav
		nyheter2.doc Upplästa nyheter	113-nyheter.wav 114-nyheter.wav 115-nyheter.wav
		beraetta-text.doc Uppläst tidningsartikel	116-dn.wav 117-dn.wav 118-dn.wav
	spontan	beraetta-tal-a.doc Återberättad tidningsartikel	119-spontan.wav
		beraetta-tal-b.doc Återberättad tidningsartikel	120-spontan.wav
		beraetta-tal-c.doc Återberättad tidningsartikel	121-spontan.wav
Dialog	textbaserad	radioteater.doc	210-radioteater.wav
	spontan	dialog1.doc Kart-dialog	211-dialog.wav
		dialog2.doc Kart-dialog	212-dialog.wav
		dialog3.doc Kart-dialog	213-dialog.wav
		dialog4.doc Kart-dialog	214-dialog.wav

Din uppgift är alltså att i transkriptionerna markera när du tycker att ett ord eller en fras är mer framhävd. Lyssnar till en ljudfil och markera i morsvarande transkription vilka ord/fraser du tycker är framhävda.

6 Gör såhär:

Lyssna på den första ljudfilen, och forsök att följa med i transkriptionen och stryka under de ord/ fraser som du tycker är framhävda. Gör det genom att stryka under de ord/fraser det rör sig om. Du får gärna spola tillbaka ljudfilen och lyssna om ifall du är osäker eller inte hinner med. Upprepa sedan detta med varje transkription. Använd gärna blyertspenna, så kan du sudda utan problem om du vill ändra något! Gör texterna i den ordning som står i tabellen, d.v.s. börja med monologerna och sluta med dialogerna.

7 Efteråt...

När du är klar lämnar du in transkriptionerna med dina markeringar i mitt fack eller i mitt rum. Därefter får du dina biocheckar, och den mer avkopplande delen av experimentet - för dig :) - börjar!

Om du undrar över något kan du alltid fråga mig, gärna via mail till sofia@ling.su.se

Lycka till!

Sofia

Sofia Gustafson-Capková
Institutionen för lingvistik
Stockholms universitet, 106 91 Stockholm
tel: 08/16 17 61, rum: C348
e-post: sofia@ling.su.se, web: www.ling.su.se/staff/sofia

Appendix 2

This appendix contains samples of the transcripts which were used in the boundary marking (section I.) task and the prominence marking task (section II.). Each speaking style is represented, but since the transcripts are identical within the tasks, we do not present separate samples for condition Read and condition Listen. Excerpts from the following transcripts are included:

I. The Boundary Annotation Task

1. Scripted monologue
2. Non-scripted monologue
3. Scripted dialogue
4. Non-scripted dialogue

II. The Prominence Annotation Task

1. Scripted monologue
2. Non-scripted monologue
3. Scripted dialogue
4. Non-scripted dialogue

To get access to the full transcripts which were used in the study, please contact the author.

I. The Boundary Annotation Task

1. Scripted monologue (News broadcast)

och dethär är dagens eko kvart i fem nya storbanken finland tog första steget giftskandalen på hallandsåsen båstad vill ha oberoende undersökningskommission och historiskt möte i belfast idag i ekostudion helena sjöholm och marianne hasslow ja idag blev det alltså klart med ytterligare en storaffär i bankvärlden det är nordbanken och finländska merita som går ihop och bildar nordens största bank samgåendet blir därmed en komplicerad teknisk affär men det var den enda form vi kunde välja det säger dom båda bankledningarna meritas vesa vaino vainio och nordbankens hans dalborg ja det är jeri juridiken som gör att den nya storaffären inte är glasklar så klar som när ett företag köper ett annat nu handlar det istället om att bilda ett nytt bolag som ska köpa aktier i dom gamla och dom gamla blir kvar som dotterbolag med organisationerna kvar i sverige och i finland och det statliga inflytandet blir kvar i nordbanken och så är det dom svenska reglerna för skatt aktiebeskattning som driver iväg huvudkontoret till helsingfors den svenska dubbelbeskattningen på utdelning är sämre än det finsk finländska sättet att ta ut skatt där får aktieägarna räkna av en del av den skatt som företaget redan betalt den nya banken blir en gigant med nordiska mått mätt var tredje invånare i norden är kund sex och en halv miljon privatkunder 300 000 företag 800 miljarder i omslutning och så vidare den här affären kan kanske bäst sammanfattas i en liknelse parterna ingår ett äktenskap men fortsätter att leva som singel på kort sikt i alla fall

2. Non-scripted monologue (Retold DN-article)

sverre sjölander har gett ut en bok som handlar om människors och djurs sett sätt att kommunicera med varandra och han drar paralleller eller liknelser till religionen för i rel i religionen har många komponenter som är väldigt grundläggande i m mänskliga beteendet menar han och dom det är män det som skiljer människan och djuren åt till exempel är detta att människan skaffar sig en religion ehmn och det är en en följd utav att att människan kan föreställa sig saker som inte är här och inte är nu att föreställa sig en annan tid både det som har varit och det som och kunna planera för framtiden och ur evolutionsmässig synpunkt så är människans förmåga att föreställa sig andra tider och platser av väldigt stor betydelse och att kunna kommunicera dom här föreställningarna med medmänniskorna betye gav en stor fördel ur ur evolutionistisk synpunkt menar sjölander det gör det gjorde att människan kunde ehm kommunicera med varandra och kunna samordna sina krafter till att till exempel jaga villebråd att komma överens om tider och strategier för att kunna nedlägga byte för att kunna klara sin energiförsörjning och så så att det blev en väldigt stark ehm press på evolutionen detta menar han språket är väldigt bra att komunicera med när det gäller viss typ utav information så att när till exempel när det gäller att beskriva platser som finns som inte är just här utan platser på andra ställen så passar språket väldigt bra f att att uttrycka såna saker att ge information

3. Scripted dialogue (Radio play)

<sp> ringde det ja hej och välkommen det gick ju smidigt ett ögonblick nu så ska jag ställa er en fråga jag sjöng förstår ni det har varit så tyst här nynnar nej jag borde inte sjunga inte prata nu det vet jag ja jag ska sluta alldeles strax jag slutar nu eee ja jag ska ställa er en fråga välkommen har jag sagt det ja men som jag pratar det är nästan litet skönt ojojojojoj det är rena tortyren som ni har fått vänta här har jag blivit gammal eee nu återstår bara formaliteter så är ni innanför min goda min gode eee ja det var just det vilket kön hade ni då ni levde hallå kom kom har jag tappat er
<s> mej nej
<sp> gott men ni hörde visst inte endast några formaliteter kvarstår en liten sådan enkel och rättvis sedan kan ni stiga in vilket kön hade ni hallå är ni där
<s> jaa
<sp> var ni man eller kvinna
<s> kön det var värre det nä det kommer jag faktiskt inte ihåg
<sp> va inte vad menar ni
<s> jag kommer inte ihåg vad jag hade för kön
<sp> ja det kommer som en chock alltså ni är bara en själ och jag har pratat för mycket mår ni inte bra det var bara en fråga jag skulle ställt den meddetsamma ah det har ingen betydelse jag tar om min fråga var ni kvinna eller man
<s> jag minns inte
<sp> minns inte jamen vi lugnar oss nu ni har fått frid eller strax ska ni få få vila i frid och jag ska återgå att vara som jag brukar såhär ser verkligheten ut: jag har bara den frågan sen är ni innanför vi har bara två kategorier här varsågod höger eller vänster det är min nästa replik jag kommer att stå litet snett hålla ut armen ungefär såhär peka mot en av dörrarna: varsågod det förstår ni
<s> ja
<sp> nå vad säger ni
<s> menar ni nyss så sa jag endast "japå den där andra frågan kan jag inte svara något säkert det får väl visa sig med tiden här är så
<sp> nä så är det inte här här är ingenting ingen tid att visa det ni ska svara mig nu meddetsamma blunda
<s> skulle påstå att jag blundar hela tiden inte blundar kanske men bilden är som tanken: trög
<sp> eeeee jag borde hjälpa er att minnas mhm stod ni eller satt ni när ni kissade
<s> <skratt> när jag kissade

4. Non-scripted dialogue (Map Task dialogue)

<talare a> då ska vi se då har vi en en sån karta här framför oss och jag har eee landstigit på en plats på den här ön och det börjar vid en en in i en bukt en ganska ovalt formad bukt inne på västra sydvästra sidan utav den här ön har du den också
<talare b> jo då ee och tydligen så är ju formen på våra öar identiska så att så att själva den bukten ska det inte vara några problem att hitta
<talare a> mmm

<talare b> jag har en säl tror jag att det är precis innanför landstigningspunkten och
<talare a> ja
<talare b> strax ovanför där så är det en hel hord med krabber
<talare a> ja just precis och då ska vi se då börjar vägen precis söder om den där sälen
<talare b> mmm
<talare a> där har jag en liten sån här ett ankare ute i havet också så här min båt ankrat och sen går jag in i land ee söder om den här sälen och gör en sväng norrut och följer den här vackra buktkanten
<talare b> okej
<talare a> i en mjuk kurva rakt eee norrut
<talare b> ska vi gå bakom ryggen på sälen då
<talare a> just precis och sen så fortsätter jag upp en bit och sen kommer det två stycken palmdungar och precis mitt
<talare b> ja just det du du går genom krabb-flocken då
<talare a> njej jag går jag går öster om krabbflocken
<talare b> okej
<talare a> så jag bara tittar på dom
<talare b> oke
<talare a> men jag gör ingenting me dom
<talare b> då snuddar du nästan vid en flod där då
<talare a> ja det är ett aningers när närmare floden än kust-kanten där men det är nästan mitt emellan
<talare b> okej
<talare a> och sen forstätter det i en mjuk eee liten buktning uppåt eee åt åt höger innan jag fortsätter mellan dom här palmdungarna
<talare b> du går emellan dom
<talare a> ja
<talare b> eee

II. The Prominence Annotation Task

1. Scripted monologue (News broadcast)

Och dethär är dagens Eko Kvart i fem. Nya storbanken, Finland tog första steget. Giftskandalen på Hallandsåsen. Båstad vill ha oberoende undersökningskommission. Och historiskt möte i Belfast idag. I Ekostudion Helena Sjöholm och Marianne Hasslow.

Ja, idag blev det alltså klart med ytterligare en storaffär i bankvärlden. Det är Nordbanken och Finländska Merita som går ihop och bildar Nordens största bank. Samgåendet blir därmed en komplicerad teknisk affär, men det var den enda form vi kunde välja, det säger dom båda bankledningarna: Meritas Vesa Vaino... Vainio och Nordbankens Hans Dalborg.

Ja det är jeri... juridiken som gör att den nya storaffären inte är glasklar, så klar som när ett företag köper ett annat. Nu handlar det istället om att bilda ett nytt bolag som ska köpa aktier i dom gamla och dom gamla blir kvar som dotterbolag med organisationerna kvar i Sverige och i Finland. Och det statliga inflytandet blir kvar i Nordbanken. Och så är det dom svenska reglerna för skatt... aktiebeskattning som driver iväg huvudkontoret till Helsingfors. Den svenska dubbelbeskattningen på utdelning är sämre än det finsk... finländska sättet att ta ut skatt. Där får aktieägarna räkna av en del av den skatt som företaget redan betalt.

Den nya banken blir en gigant med nordiska mått mätt. Var tredje invånare i Norden är kund. sex och en halv miljon privatkunder, 300 000 företag, 800 miljarder i omslutning och så vidare. Den här affären kan kanske bäst sammanfattas i en liknelse. Parterna ingår ett äktenskap men fortsätter att leva som singel. På kort sikt i alla fall.

2. Non-scripted monologue (Retold DN-article)

Sverre Sjölander har gett ut en bok som handlar om människors och djurs sett... sätt att kommunicera med varandra och han drar paralleller eller liknelser till religionen för i rel i religionen har många komponenter som är väldigt grundläggande i m mänskliga beteendet menar han//

och dom det är män... det som skiljer människan och djuren åt till exempel är detta att människan skaffar sig en religion ehmn och det är en en följd utav att att människan kan föreställa sig saker som inte är här och inte är nu att föreställa sig en annan tid både det som har varit och det som och kunna planera för framtiden//

och ur evolutionsmässig synpunkt så är människans förmåga att föreställa sig andra tider och platser av väldigt stor betydelse och att kunna kommunicera dom här föreställningarna med medmänniskorna betye... gav en stor fördel ur ur evolutionistisk synpunkt menar Sjölander//

det gör det gjorde att människan kunde ehm kommunicera med varandra och kunna samordna sina krafter till att till exempel jaga villebråd att komma överens om tider och strategier för att kunna nedlägga byte för att kunna klara sin energiförsörjning och så så att det blev en väldigt stark ehm press på evolutionen detta menar han //

språket är väldigt bra att komunicera med när det gäller viss typ utav information så att när till exempel när det gäller att beskriva platser som finns som inte är just här utan platser på andra ställen så passar språket väldigt bra f att att uttrycka såna saker att ge information//

3. Scripted dialogue (Radio play)

<SP> ringde det? Ja hej och välkommen, det gick ju smidigt. Ett ögonblick nu så ska jag ställa er en fråga. Jag sjöng förstår ni, det har varit så tyst här. Nynnar. Nej, jag borde inte sjunga! Inte prata nu, det vet jag. Ja jag ska sluta alldeles strax. Jag slutar nu. Ja, jag ska ställa er en fråga. Välkommen, har jag sagt det? Ja, men som jag pratar, det är nästan litet skönt. Ojojojojoj, det är rena tortyren som ni har fått vänta här! Har jag blivit gammal? Nu återstår bara formaliteter så är ni innanför, min goda... min gode... eee Ja, det var just det, vilket kön hade ni då ni levde? Hallå? Kom? Kom? Har jag tappat er?
<S> Nej... Nej...
<SP> Gott. Men ni hörde visst inte. Endast några formaliteter kvarstår. En liten sådan, enkel och rättvis sedan kan ni stiga in. Vilket kön hade ni? Hallå, är ni där?
<S> Jaa.
<SP> Var ni man eller kvinna?
<S> Kön? Det var värre det. Nä, det kommer jag faktiskt inte ihåg.
<SP> Va! Inte! Vad menar ni?
<S> Jag kommer inte ihåg vad jag hade för kön.
<SP> Ja. Det kommer som en chock. Alltså ni är bara en själ och jag har pratat för mycket. Mår ni inte bra? Det var bara en fråga, jag skulle ställt den meddetsamma. Ah, det har ingen betydelse, jag tar om min fråga! Var ni kvinna eller man?
<S> Jag minns inte.
<SP> Minns inte?! Jamen. Vi lugnar oss nu. Ni har fått frid. Eller strax ska ni få få vila i frid, och jag ska återgå att vara som jag brukar. Såhär ser verkligheten ut: Jag har bara den frågan. Sen, är ni innanför. Vi har bara två kategorier här. Varsågod höger eller vänster, det är min nästa replik. Jag kommer att stå litet snett, hålla ut armen ungefär såhär, peka mot en av dörrarna: Varsågod! Det förstår ni?
<S> Ja.
<SP> Nå, vad säger ni?
<S> Menar ni nyss? Så sa jag endast "ja". På den där andra frågan kan jag inte svara något säkert, det får väl visa sig med tiden. Här är så...
<SP> Nä! Så är det inte här! Här är ingenting! Ingen tid att visa det. Ni ska svara mig nu meddetsamma. Blunda!
<S> Skulle påstå att jag blundar hela tiden. Inte blundar kanske men... Bilden är som tanken: trög
<SP> Eeeee jag borde hjälpa er att minnas. Mhm stod ni eller satt ni när ni kissade?
<S> <skratt> när jag kissade?

4. Non-scripted dialogue (Map Task dialogue)

<Talare A> då ska vi se då har vi en en sån karta här framför oss och jag har eee landstigit på en plats på den här ön och det börjar vid en en in i en bukt en ganska ovalt formad bukt inne på västra sydvästra sidan utav den här ön har du den också
<Talare B> jo då ee och tydligen så är ju formen på våra öar identiska så att så att själva den bukten

ska det inte vara några problem att hitta
<Talare A> mmm
<Talare B> jag har en säl tror jag att det är precis innanför landstigningspunkten och
<Talare A> ja
<Talare B> strax ovanför där så är det en hel hord med krabber
<Talare A> ja just precis och då ska vi se då börjar vägen precis söder om den där sälen
<Talare B> mmm
<Talare A> där har jag en liten sån här ett ankare ute i havet också så här min båt ankrat och sen går jag in i land ee söder om den här sälen och gör en sväng norrut och följer den här vackra buktkanten
<Talare B> okej
<Talare A> i en mjuk kurva rakt eee norrut
<Talare B> ska vi gå bakom ryggen på sälen då
<Talare A> just precis och sen så fortsätter jag upp en bit och sen kommer det två stycken palmdungar och precis mitt
<Talare B> ja just det du du går genom krabb-flocken då
<Talare A> njej jag går jag går öster om krabbflocken
<Talare B> okej
<Talare A> så jag bara tittar på dom
<Talare B> oke
<Talare A> men jag gör ingenting me dom
<Talare B> då snuddar du nästan vid en flod där då
<Talare A> ja det är ett aningers när närmare floden än kust-kanten där men det är nästan mitt emellan
<Talare B> okej
<Talare A> och sen forstätter det i en mjuk eee liten buktning uppåt eee åt åt höger innan jag fortsätter mellan dom här palmdungarna
<Talare B> du går emellan dom
<Talare A> ja
<Talare B> eee