Weighted error minimization in assigning prosodic structure for synthetic speech

Weighted error minimization in assigning prosodic structure for synthetic speech

The work described in this thesis has been carried out under the auspices of
the J. F. Schouten School for User-System Interaction Research,
and was funded by SOBU, TU/e and KPN.

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Herwijnen, Olga M. van

Weighted error minimization in assigning prosodic structure for synthetic speech /
by Olga M. van Herwijnen. –
Eindhoven: Technische Universiteit Eindhoven, 2004. –
Proefschrift. –
ISBN 90-386-1908-1
NUR 616
Keywords: Speech synthesis / Prosody / Machine learning / Psycholinguistics

Printing: Universiteitsdrukkerij Technische Universiteit Eindhoven.

Weighted error minimization in assigning prosodic structure for synthetic speech

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van
de Rector Magnificus, prof.dr. R.A. van Santen, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op maandag 17 mei 2004 om 16.00 uur

door

Olga Marjolein van Herwijnen

geboren te Amsterdam

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. R.P.G. Collier
en
prof.dr. W.M.P. Daelemans

Copromotor:
dr. J.M.B. Terken

Acknowledgement

Writing a thesis is something one cannot accomplish without the support from supervisors, colleagues, friends and family. Therefore, I want to seize the opportunity to express some words of thankfulness to some people in particular.

First of all, I sincerely would like to thank Jacques Terken, René Collier, Walter Daelemans and Antal van den Bosch for their supervision, help and useful comments and for sharing their knowledge and enthusiasm. I learned a lot from you.

I thank Carlos Gussenhoven and Lou Boves for their constructive comments on the manuscript. Erwin Marsi and Marc Swerts are thanked for their valuable contributions as experts in the field of prosody research. Toni Rietveld is thanked for helping me with the statistical design for the experiments in Chapters 3 and 6. Esther Janse, Hanny den Ouden, Ielka van der Sluis and Piroska Lendvai are thanked for their pleasant company during trips to Aalborg, Trento and Budapest.

With respect to the experimental conditions, I am very grateful for the opportunity to use the Corpus Gesproken Nederlands. Furthermore, Martin Boschman and Kees Kuijpers are thanked for their help with the physical environment for the experiments.

The AiO's of the J.F. Schoutenschool are thanked for creating a joyous atmosphere, which made the working days (and the after-work gatherings) something to look forward to. Of course a special thanks goes out to all other (ex-)colleagues of the former IPO, the UCE-group and MTI-group in Eindhoven and the ILK-group in Tilburg.

I would like to express my gratitude to my family, especially my parents and sister, for supporting me. Finally, my dearest Mark is thanked for believing in me and for being there to share thoughts and experiences, in science and in daily life.

Olga van Herwijnen

Contents

General introduction 1

The quality of synthetic speech is still not fully acceptable, among other things due to the lack of proper prosodic structure. Phrase boundaries and accents are missing or allocated incorrectly. Correct prosodic structure reduces the time and effort that it takes listeners to process and understand artificially generated speech. In this thesis, existing Dutch Text-to-Speech systems have been evaluated to assess the major factors that cause errors in the allocation of accents and phrase boundaries. The aim of the research described here is to improve the assignment of prosodic structure on the basis of syntactic and lexical information. This project combines two lines of research: language engineering and psycholinguistics. We will apply language engineering techniques, taking into account the psycholinguistic information which we obtained through perception experiments. We will present a module for the assignment of prosodic structure (ECLIPSE), which will be evaluated by comparing it to a reference transcription by human experts.

1.1 Introduction

Nowadays, the use of synthetic speech is getting more common. For instance, when calling public transport information services, when making banking transfers via the telephone ('telebanking'), or when using car navigation systems, one will get the information in synthetic speech (in the vernacular also called 'computer speech'). When listening to such speech, one immediately notices that it often sounds unnatural. It takes more time and effort to understand the message than when the same utterance is spoken by a human. Scenario 1.1 illustrates the merit of improved intelligibility of synthetic speech.

(1.1) A businessman is on the road to one of his most important customers. Every day there are a number of new messages in his e-mail box, so when he's again in a traffic jam, he thinks he might as well go through these messages. Since he needs to focus on the surrounding traffic, he uses the e-mail 'narrator' on his notebook. Unfortunately, the intelligibility of the synthetic speech is quite poor, so he has to go through the message back and forth quite a few times, before he understands that today's appointment has been postponed by one hour. While playing the message a few times, he passes by a restaurant. If the intelligibility of the synthetic speech had been better, he could have gone for coffee, while seeing the traffic jam dissolve. Now it is too late.

There are a number of reasons why synthetic speech is more difficult to understand than natural speech: (i) in concatenative synthesis the speech output falters due to the improper concatenation of words or parts of words, (ii) the voice quality is often worse than that of a human, in particular in rule-based systems, and (iii) for both synthesis types the prosodic structure is often improper, which means that accents and phrase boundaries are allocated incorrectly. To improve the quality of Dutch synthetic speech much effort has already been invested to remedy the first two causes of lower intelligibility. Effort has mainly been put in the domain of segmental characteristics (e.g. Klabbers, 2000) and voice source parameters (e.g. van Dinther, 2003). The third reason for the unsatisfactory intelligibility of synthetic speech is prosody. We define prosody as a combination of melody and prosodic structure. Melody is defined as the particular shape of an intonation contour, which may vary independently of the prosodic structure. We define the prosodic structure of a sentence as the distribution of accents and phrase boundaries over the utterance. The prosodic structure reflects which words of an utterance are the most informative. These words should be emphasized by means of an accent (often realized as a pitch change). Accents make the words prominent to the listener, thus attracting his attention to these words. The prosodic structure also reflects which words belong together on the information level (Bolinger, 1989) and it indicates the information domains by means of phrase boundaries (often realized as pauses). The words within such a domain (i.e. between two boundaries) should be processed by the listener as one chunk of information.

Sentence 1.2 is an example of a sentence annotated with prosodic structure. Accents are indicated with asterisks and phrase boundaries with slashes.

(1.2) Hij zag het *meisje // met de *verrekijker ///

Figure 1.1 is an illustration of a possible pitch contour of the synthesized utterance of this sentence. This figure is an example of how the phonological (or prosodic) structure of a sentence can be realized by a speaker or Text-to-Speech system as a sequence of pitch rises and falls and slots of silence.

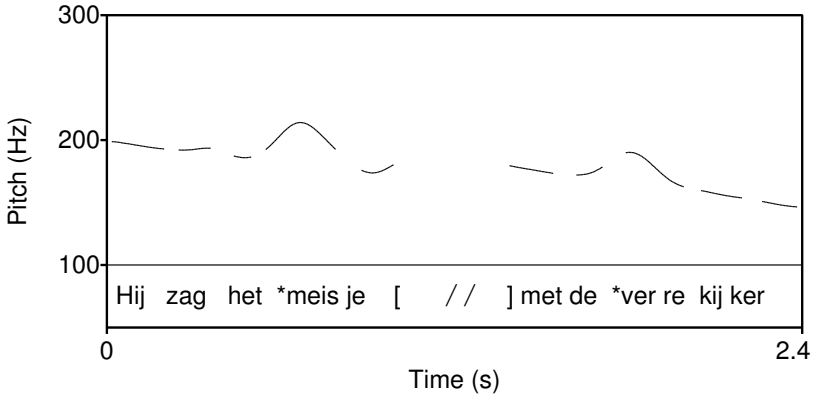


Figure 1.1: *Pitch contour of the synthesized utterance of sentence 1.2. The words 'meisje' and 'verrekijker' are accented, which is indicated by the asterisks in the text and the peaks in the pitch contour. There is a phrase boundary between the words 'meisje' and 'met', indicated by the slashes in the text and the pause in the speech signal.*

Correct prosodic phrasing and accentuation are important determinants of the acceptability of synthetic speech (Silverman et al., 1993; Koehn et al., 2000; Pan and Hirschberg, 2000). Incorrect or inadequate prosodic structure will distract or even mislead the listener. In today's Text-to-Speech systems the assignment of prosodic structure is still rather poor. This is one of the reasons why listeners often judge synthetic speech as only moderately acceptable. The acceptability of synthesized utterances furthermore depends on the physical implementation of accents and phrase boundaries. Although we are aware of the fact that these two aspects of synthetic speech cannot be fully appreciated when considered separately, the aim of the research described in this thesis is to improve the automatic assignment of prosodic structure, largely ignoring the aspect of melody.

Figure 1.2 shows the different steps in the process of text to speech conversion. Ideally, the input text is first submitted to semantic and pragmatic analysis before starting the syntactic analysis. The input text, together with semantic and pragmatic information, is then analyzed by a syntactic parser. The syntactic representation forms the input for the generation of prosodic structure. This prosodic structure and the input text are the

two sources of information that the Text-to-Speech system uses to generate a spoken version of the input text.

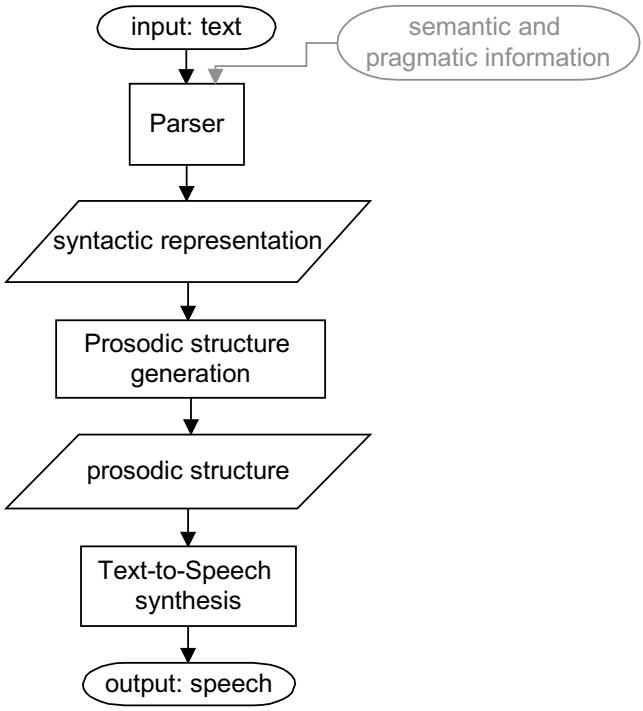


Figure 1.2: *Representation of the different steps in the process of text to speech conversion.*

The project started from the observation that the assignment of prosodic structure requires elaborate information about syntactic structure, and that one important cause of errors in the assignment of prosodic structure is the fact that syntactic ambiguities[1] in the input text cannot be resolved without access to other sources of information, most notably semantic information. In this project we explore a combination of several techniques and solutions to elaborate the syntactic information that we derive from a state-of-the-art syntactic parser, to resolve some syntactic ambiguities. This should result in improving the prediction of prosodic structure. Previous research (e.g. Prevost and Steedman, 1994) showed that several textual characteristics such as given/new status, contrastiveness and discourse structure are also determinants in the prediction of accents and phrase boundaries. Initially, the topic of our research was to improve prosodic structure prediction for free texts (like newspaper text and e-mail messages) taking into account all context effects. Later on, we have mainly focussed on the influence of syntax, which is more obvious in sentences in isolation, since pragmatic and semantic characteristics are less important to sentences in isolation.

[1]Note that syntactic ambiguity does not always lead to two semantically possible readings. This means that if one of the readings is semantically improper, the prosodic structure assigned according to that specific reading will be considered as incorrect.

Prosodic structure

Prosody is the ensemble of phonetic properties (like tempo, speech melody, loudness and prominence) that are characteristics of longer stretches of speech. A speaker can use prosody to indicate the relative prominence of words by providing them with an accent, and to indicate which sequences of words within a sentence form a coherent unit. The entirety of accents and phrase boundaries is referred to as prosodic structure.

In autosegmental phonology (Pierrehumbert, 1980; Beckman, 1996; Shattuck-Hufnagel and Turk, 1996) prosodic structure is described as a string of tonal events. The tonal events are either pitch accents or boundary tones, which are composed of one or more tone segments. These tone segments (H and L) are phonological abstractions of high and low pitch level. Which pitch patterns (i.e. sequences of H and L) count as a pitch accent must be specified by a phonological description. Accent is usually treated as a binary feature (+/- accent). Pitch accents are often defined as a local feature of pitch – usually realized as a sudden change in pitch level.

According to Beckman (1996) the intonational phrase is the phonological element for describing prosodic structure. The intonational phrase is defined by the distribution of H% and L% boundary tones. These boundary tones mark the end of the intonational phrase. Intonational phrases consist of one or more phonological phrases. The layered representation makes a binary approach for boundary strength inadequate. Sanderman (1996) treats boundary strength as a scalar feature distinguishing different strengths of boundaries. On the basis of Sanderman (1996) we will distinguish four strengths (i.e. no boundary, weak, medium and strong boundaries). Phrase boundaries will be realized by one or more of the following factors: sustained high pitch, pre-pausal lengthening, continuation rise and pause.

As mentioned before, in this thesis we focus on automatic allocation of accents and phrase boundaries. Proper allocation is important since variation in the accentuation pattern of an utterance may induce different meanings of the same text of a sentence. For example, sentences 1.3a and 1.3b consist of the same words, however the meaning is different. Sentence 1.3a means that the girl was playing next to the sandpit when she fell down into it, whereas sentence 1.3b means that the girl was already in the sandpit when she fell down.

(1.3) (a) *Het *meisje is in de *zandbak gevallen.*
The girl is into the sandpit fallen down.
"The girl fell into the sandpit."

(b) *Het *meisje is in de *zandbak *gevallen.*
The girl is in the sandpit fallen down.
"The girl fell down in the sandpit."

Note that sentence 1.3b can also be read as an answer to the question in sentence 1.4.

(1.4) *Is het *meisje in de *zandbak *gesprongen?*
Did the girl into the sandpit jump?
"Did the girl jump into the sandpit?

The accent on '*gevallen*' (fallen) in the answer then indicates the contrast with '*gesprongen*' (jumped) in the question. Here, the context plays an important role in assigning the proper prosodic structure, rising from proper contextual interpretation. Experience shows that if there exists a mismatch between context and the meaning induced by prosody, the context is the dominant factor in the interpretation. However, in most cases prosody is compatible with the contextually appropriate reading, so that this conflict doesn't arise.

Phrase boundaries indicate which words belong together syntactically and semantically. Comparably to accentuation, variation in the prosodic phrasing of an utterance may induce different meanings of the same sentence. For instance, sentences 1.5a and 1.5b consist of the same words, but the meaning is different. Sentence 1.5a means that he saw the girl who had the binoculars, whereas sentence 1.5b more readily lends itself to an interpretation that he saw the girl through the binoculars.

(1.5) (a) *Hij zag het meisje met de verrekijker ///*
He saw the girl with the binoculars.
"He saw the girl with the binoculars."

(b) *Hij zag het meisje / met de verrekijker ///*
He saw the girl through the binoculars.
"He saw the girl through the binoculars."

Relation syntax-prosody

A number of studies ('t Hart and Cohen, 1973; Pierrehumbert, 1980; Selkirk, 1984; Bachenko et al., 1986) have investigated the relationship between sentence prosody and syntactic structure. For instance, speakers tend to produce different prosodic structures depending on the presence of a sentence-internal major syntactic boundary. Listeners use these prosodic structures for the interpretation of syntactically ambiguous sentences (Beach, 1991). Researchers agree upon the influence of syntax on the prosodic structure, acknowledging the fact that there is no one-to-one mapping between the syntactic and the phonological representation. Selkirk (1984), for instance, states that "a sentence may correspond to one or more intonational phrases. An intonational phrase typically contains material belonging to a sequence of words and/or phrases, and it is not necessarily isomorphic to any constituent of syntactic structure". The rationale is

that the definition of what may constitute an intonational phrase is essentially semantic in character, meaning that syntactic information alone does not suffice for proper assignment of phrase boundaries and accents. For the purpose of prosodic structure prediction the syntactic information should be elaborated on the basis of information about semantic relations between words and phrases.

This idea is supported by several other studies. Keijsper (1985), for instance, argues that accentuation of a sentence is also decided on the basis of meaning. She states that a coherent text is one which only contains (contextually independent) meanings which can be interpreted in the given context. Accentuation can induce a different meaning of the text in that specific context. Appropriate accentuation is not only based on syntactic and lexical information, but also on word order and context.

With respect to prosodic phrasing Fach (1999), for instance, showed that for read speech only 65% of syntactic boundaries are coded in the prosodic boundaries, and Bachenko and Fitzpatrick (1990) claim that the syntax plays a necessary but *not* sufficient role in determining phrasing.

Selkirk (1986, 1995) proposed a theory of phrasing based on edge alignment of phonological phrases with syntactic XP's. The constraints on this for Dutch are formulated as:

> "For each XP there is a P such
> that the right edge of XP coincides with the right edge of P."
>
> where P stands for phonological phrase.

This means that prosodic phrase boundaries align with the right edges of the constituent major projections. The left boundary of the major projection is not necessarily marked by a prosodic boundary, although it could be due to the marking of the right edge of another major projection.

The studies mentioned above are based on full syntactic parsing of the sentences. However, it has also been discussed that for determining prosodic structure shallow syntactic information will suffice (e.g. Abney, 1991; Willemse and Boves, 1991). What is needed is information about the phonological phrases in a sentence. Gee and Grosjean (1983) already showed that phonological phrases are a good predictor of intonation. They see the sentence as made up of a number of "prosodic units". These units match the phonological and syntactic units of the sentence, however not completely. Abney (1991) argues that when we read a sentence we read it one chunk at a time. He states that these chunks correspond to prosodic patterns, where each chunk takes at least one accent and where phrase boundaries are most likely to fall between these chunks. According to Abney (1991) a chunk corresponds to a phonological phrase.

Following this reasoning we anticipate that for our purpose of assigning proper prosodic structure for synthetic speech we do not need a full syntactic parse. What

we minimally need to reach our goal is a shallow syntactic analysis. This analysis should provide us with information about what the immediate constituents are in the sentence in combination with information about sense units, that is, domains that consist of immediate constituents belonging together (major projections on main lexical categories). This is what we will add to the shallow syntactic information.

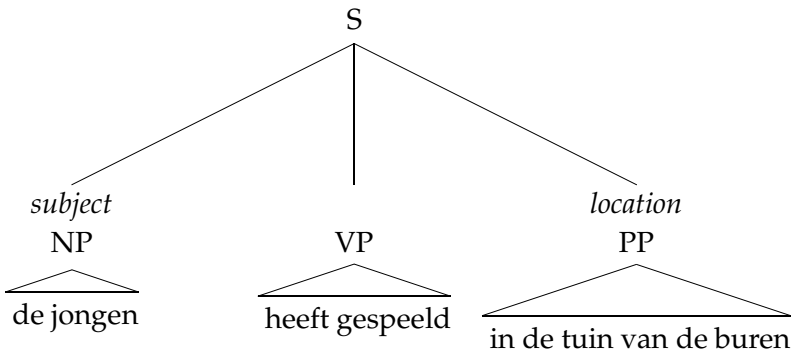
While two sentences are the same at the level of immediate constituents (as in example 1.6), there may be a difference in the co-occurrence of the constituents. Semantic information finally decides the attachment. This is indicated on the basis of representations 1.7 and 1.8.

(1.6) (a) *De *jongen heeft in de *tuin van de *buren gespeeld ///*
The boy has in the garden of the neighbors played.
"The boy has played in the neighbors' garden."

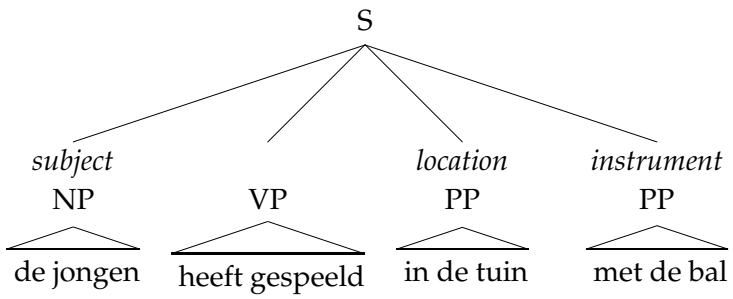
(b) *De *jongen heeft in de *tuin / met de *bal gespeeld ///*
The boy has in the garden with the ball played.
"The boy has played in the garden with the ball."

In example 1.7 the prepositional phrase (PP) '*van de buren*' is part of the PP '*in de tuin*', whereas in example 1.8 the PP '*met de bal*' is a separate PP. In the latter case, a phrase boundary preceding the second PP is acceptable, whereas in the first it is not. If this information is not available, both sentences will be realized with the same prosodic structure, which means that there are no clues for the listener that indicate which relation there exist between the PP and the noun or verb in the sentence.

(1.7)



(1.8)



In this thesis, we will focus on sentences in isolation. In such sentences pragmatic information will not have a large effect on the allocation of accents and phrase boundaries. Semantic information will mainly affect the prosodic structure with respect to ambiguities. To resolve some categories of ambiguities we will extend the syntactic information on the basis of lexical and co-occurrence information of the constituents. In section 1.2, we will describe our approach for elaborating syntactic information. In section 1.3, we will give an overview of the thesis outline. In section 1.4, we describe the material we used for the experiments and evaluations.

1.2 Approach

The project described in this thesis starts with the evaluation of existing TTS systems and an error analysis to assess the major accentuation and phrasing problems. Next, we explore the effect of using proper syntactic analysis of the sentence. Since existing TTS systems for Dutch do not adequately allocate phrase boundaries, we draw up a revised phrasing algorithm based on sentence length and punctuation. Next, the perceptual costs of the different kinds of accentuation and phrasing errors are determined. This knowledge is used for decisions about optimization strategies in machine learning experiments. These experiments are carried out to enhance the syntactic information. Finally, the resulting prosody module is evaluated.

The examples in section 1.1 show that the prosodic structure is informative to the listener for the comprehension of an utterance. This implies that for Text-to-Speech synthesis it is important to assign proper prosodic structure to a sentence. The question is what a proper prosodic structure is for a given sentence. When evaluating the prosodic structure of a sentence, as assigned by TTS systems, we need a reference transcription which is the norm for the proper prosodic structure for this specific sentence. By means of comparison with such a reference transcription we will be able to decide whether the automatically assigned prosodic structure is correct. Deviations from the reference transcription will be considered as errors. Correct accentuation of a word means that the word is accented in the Text-to-Speech version in accordance with the reference transcription. The same applies to the allocation of phrase boundaries.

A reference transcription can be obtained in several ways. For instance, a spoken version of text by one speaker can be used as a reference. A consensus transcription can be derived of two or more experts, by having them actually sit together and agree upon one transcription. Another possibility is statistical computation of an "average" transcription of the annotations of a number of experts. We chose to use the latter approach, since this method is not as time consuming as a consensus transcription, while still we use the transcriptions of more than one speaker. A disadvantage of this approach is that the "average" transcription may be a prosodic structure which has not been assigned by any of the experts. When part of the experts assigned one pattern

and part of the experts assigned another pattern, the consequence might be that in the reference transcription an uncommon pattern is derived. Before using this reference transcription, we therefore validated the transcription through computation of expert agreement and comparison with a spoken reference. These studies showed that our straightforward approach turns out to be unproblematic in most cases.

To measure the progress we make in assigning prosodic structure, we also use two metrics: (i) quantitative evaluation metrics from the domain of language engineering, and (ii) perceptual evaluation experiments. Evaluation of prosodic structure on the basis of a reference transcription by human experts provides us with objective information about the performance of the module that assigns prosodic structure. For this we use evaluation metrics from the domain of language engineering. We perform a subjective evaluation by means of perception experiments so that we can also assess the acceptability of the assigned prosodic structure, next to the exact agreement between the prosodic structure as assigned by the prosody module and the reference transcription.

In this thesis the improvement of physical realization of prosodic structure is set aside. The Text-to-Speech system that we will use for several comparison studies is Calipso[2], that has been evaluated earlier (Terken, 1993). This evaluation showed that for isolated utterances the naturalness of the physical implementation of the intonation was as good as the human intonation.

Psycholinguistics

Inappropriateness of prosodic structure may occur in two ways: (i) words may be accented that should remain unaccented, and phrase boundaries may occur at locations where there should be no phrase break, and (ii) words may remain unaccented when in fact they should be accented, and phrase boundaries may be omitted where there should be a phrase break. The different kinds of errors may not all be equally serious, which means that when we want to evaluate and improve algorithms for accentuation and phrase boundary allocation, we need to take into account the perceptual costs of different types of errors.

In this thesis the field of psycholinguistics is restricted to the domain of speech perception and language comprehension. Cutler et al. (1997) discuss the effect of prosody on the comprehension of speech. There is some freedom in the assignment of prosodic structure, in the sense that a given sequence of words with a particular meaning in a particular context may be associated with several equally acceptable prosodic structures, but this does not mean that all prosodic alternatives are acceptable to the listener. In certain cases listeners have clear intuitions that the prosodic structure is inappropri-

[2]Calipso is a Text-to-Speech system developed by J.R. de Pijper at the former IPO, Center for User-System Interaction, Eindhoven University of Technology.

ate. In this sense we may conclude that there is a linguistic basis for the intuitions of listeners.

Research on the role of prosody in the comprehension of spoken language has made clear that phrasing and accentuation may help the listener to impose structure onto the incoming speech signal, that is to group words and phrases that belong together, and to focus attention on the information that needs most processing effort. As a consequence, inappropriate prosodic structures may slow down the listener's comprehension of the incoming speech signal, as for instance is shown in experiments on accentuation (Nooteboom and Kruyt, 1987; Terken and Nooteboom, 1987) and on prosodic phrasing (Sanderman and Collier, 1997).

Language engineering

Research in the domain of language technology has given rise to data-oriented approaches to the analysis of linguistic structures (parsing), complementing or even replacing earlier rule-based approaches. Previous research has explored possibilities to apply methods from the domain of language engineering to the assignment of prosodic structure (e.g. Ostendorf and Veilleux, 1994; Pan and McKeown, 1999). A more specific technique which is currently explored for predicting prosodic structure is machine learning (e.g. Hirschberg and Prieto, 1996; Koehn et al., 2000; Marsi et al., 2002, 2003).

Machine learning algorithms extrapolate from the example to new input cases, either by extracting regularities from the examples for instance in the form of rules or decision trees, or by a more direct use of analogy in lazy learning algorithms such as memory-based learning.

Machine learning techniques are often explored for directly predicting the prosodic structure as a whole. In this thesis, we perform machine learning experiments to elaborate the syntactic information through resolving two syntactic ambiguities, since state-of-the-art parsers for Dutch do not provide us with sufficient syntactic information. On the basis of this elaborate syntactic structure we then assign prosodic phrase boundaries and accents.

Data for these machine learning experiments is derived from the Corpus Gesproken Nederlands (CGN, Spoken Dutch Corpus)[3] and the World Wide Web. The electronic (on-line) availability of such large amounts of data on the internet has made it possible to compute distributional patterns of linguistic structures and co-occurrence relations of words. We use the WWW to compute the co-occurrence relation of words, which provides us with probabilistic information about the extent to which these words belong together semantically. Data from the Spoken Dutch Corpus provides us with

[3]The *Spoken Dutch Corpus* is a database of contemporary Dutch as spoken by adults in the Netherlands and Flanders. The project is funded by the Flemish and Dutch governments and the Netherlands Organization for Scientific Research NWO. Its homepage is `http://lands.let.kun.nl/cgn/ehome.htm`.

lexical information, which we use as features in the machine learning experiments. We apply machine learning techniques using the perceptual costs of errors in accentuation and phrasing. This information is used to decide whether there should be a bias for fewer or more accents and boundaries.

ECLIPSE

The results of the experiments described above will eventually lead to a module for the prediction and assignment of prosodic structure. This module will be henceforth referred to as ECLIPSE (*Extensive Computation of Linguistic Information for Prosodic Structure Estimation*). Figure 1.3 shows the different steps in which ECLIPSE will assign prosodic structure to text.

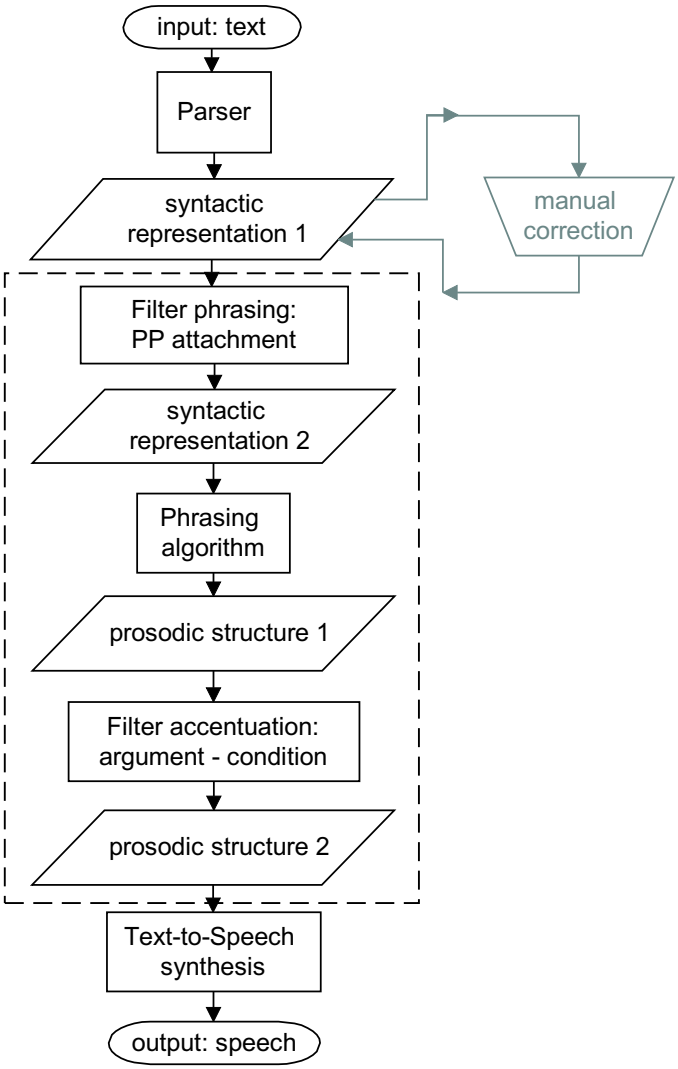


Figure 1.3: *Representation of the different steps in the process of predicting prosodic structure. The dashed line indicates the domain of the prosody module ECLIPSE.*

ECLIPSE starts with the syntactic representation provided by a parser. This syntactic representation only contains immediate constituents. Information about PP attachment, derived from machine learning experiments is added in the second syntactic representation. This contains a clustering of immediate constituents according to the PP attachment. From the syntactic representation a first prosodic structure is computed, which is the text annotated with phrase boundaries. Next, accents are assigned to specific words, taking information about the distinction argument–condition into account. The resulting prosodic structure is the text annotated with phrase boundaries and accents. This prosodic structure for the input text is than realized by a Text-to-Speech system.

The performance of ECLIPSE will be evaluated in comparison with the earlier mentioned reference transcription by human experts and the most extended system for Dutch prosody assignment PROS-3 (Dirksen, 1994). In PROS-3 a deterministic parser assigns a hierarchical syntactic representation to the input text. Next, a metrical analysis of this text is performed and focus is assigned. The metrical tree is used to derive a prosodic structure of the sentence, specifying locations for phrase boundaries and accents (applying the Rhythm rule for deaccenting in left-recursive metrical trees).

1.3 Thesis outline

In Chapter 2 we describe the evaluation of three existing TTS systems for Dutch. This evaluation is carried out to investigate the main problems that occur when assigning prosodic structure. For this evaluation we establish a reference transcription from prosodic annotations of 10 phonetic experts. We also evaluate an existing module for prosody assignment, under various conditions, to assess some first attempts to improve assignment of prosodic structure.

In Chapter 3 we describe two perception experiments we carried out to determine the perceptual costs of prosodic errors that are due to two major error causing factors found in Chapter 2. The first experiment is on the allocation of prosodic phrase boundaries at junctures preceding a noun- or verb-attached prepositional phrase (PP). The second experiment is on the accentuation of sentence final verbs when they are preceded by an argument or condition.

In Chapter 4 we describe machine learning experiments using a memory based and a rule-based algorithm, to predict whether a PP is noun or verb attached. The attachment decides whether or not the PP may be preceded by a prosodic phrase boundary. Furthermore, we investigate the benefit of PP attachment information for phrase boundary allocation.

In Chapter 5 we describe machine learning experiments using the same memory-based learner and rule-based learner, to predict whether the sentence final verb is preceded by an argument or a condition. The status of the nominal constituent that precedes the

verb decides whether or not the verb should be accented. Furthermore, we investigate the added value of information about the status of the nominal constituent for accent assignment.

In Chapter 6 we describe the evaluation of the resulting prosody module ECLIPSE, which has been constructed on the basis of our findings described in the preceding chapters. We evaluate the prosody module by means of a comparison with the reference transcription and PROS-3 (i) through computation of the performance measures and (ii) by means of a perception experiment.

Finally, in Chapter 7 we describe the main findings, and we discuss which problems remain unsolved.

1.4 Material

In this section we introduce the material that we used for the different experiments and evaluations. Most of the material that we use is part of the reference transcription of experts. The entire reference transcription consists of 2 newspaper articles and 15 e-mail messages. For many experiments, we used only part of the reference transcription, either because the evaluation could be performed on a smaller test set, or because we were only interested in some specific instances. Next to the reference transcription we performed experiments with instances from the Spoken Dutch Corpus (CGN). These instances were selected because they either contained an attached PP or a sentence final verb preceded by an argument or a condition. Table 1.1 gives an overview of the data that we used for the different experiments and evaluations.

Table 1.1: *Overview of material used for the different experiments and evaluations.*

experiment / evaluation	**chapter**	**data set**
reference transcription	2	newspaper + e-mail
expert agreement	2	newspaper (20 sentences)
comparison with spoken reference	2	reference (20 sentences)
evaluation TTS systems	2	reference
error analysis	2	reference (newspaper)
evaluation PROS-3	2	reference (20 sentences)
perception experiment phrasing	3	reference (20 sent. PP attachment)
perception experiment accentuation	3	reference (20 sent. ARG/COND)
ML experiment PP attachment	4	CGN (1004 sent.) + reference (157 sent.)
ML experiment ARG/COND	5	CGN (1613 sent.) + reference (61 sent.)
objective evaluation ECLIPSE	6	reference (24 sentences)
subjective evaluation ECLIPSE	6	reference (24 sentences)
expert judgement ECLIPSE	6	reference (24 sentences)

Evaluation of Dutch TTS systems

2

In this chapter we describe an evaluation study which was carried out to investigate what problems exist in automatic prosody assignment by state-of-the-art Text-to-Speech systems. For this evaluation we compared prosodic structure assigned by TTS systems to a reference transcription from annotations of text by 10 human experts. The results of the evaluation showed that the main problem for both phrasing and accentuation is incorrect or insufficient syntactic information. A second evaluation study and a perception experiment proved that proper syntactic information together with a revised phrasing algorithm improve the assignment of prosodic structure.

[0]This chapter is based on van Herwijnen and Terken (2000, 2001a,b)

2.1 Introduction

In this thesis, we are concerned with improving the prosodic structure in synthetic speech. Whenever one wants to improve a system, one needs to have profound knowledge about the current status of the system's performance. For our goal, this implies that we have to assess the state-of-the-art for the assignment of prosodic structure by Dutch Text-to-Speech systems. To assess the state-of-the-art we will evaluate TTS-systems. This will provide us with information about the problems that exist (i.e. what kind of error arise) when predicting prosodic structure automatically. By means of an error analysis of the prosodic structures as generated by the TTS-systems, we will identify actions for future improvements in automatic prosody assignment.

For the evaluation of the TTS-systems we need a reference prosodic structure. This reference will be considered to be a proper prosodic structure. We will refer to deviations from this reference structure as errors. In section 2.2, we describe the method we used to obtain a reference structure from 10 experts (henceforth referred to as reference transcription). This reference transcription is validated through computing the expert agreement and comparing the expert annotations of 2 newspaper articles and 15 e-mail messages to spoken versions of the same text. We show that the reference transcription is a proper tool for evaluation of systems that automatically assign prosodic structure.

In section 2.3, we describe the evaluation of three Text-to-Speech systems for Dutch. Comparison with the reference transcription shows that there is a considerable discrepancy between the prosodic structure (both for phrase boundaries and accents) as assigned by the systems and the reference transcription. Moreover, we discuss the error analysis to assess the major error-inducing factors. Two major factors turn out to be the incorrect insertion of prosodic phrase boundaries inside syntactic phrases and incorrect accentuation of sentence final verb phrases.

Since, the major error-inducing factors are dependent on proper syntactic information, we will make a first attempt to improve prosodic structure prediction through evaluation of PROS-3 (Dirksen, 1994) under various conditions. In section 2.4, we describe the evaluation of PROS-3 as such, a system that assigns prosodic structure to text on the basis of the output of a robust syntactic parser. Secondly, we investigate the merit of using improved syntactic information as input for PROS-3. And thirdly, we evaluate PROS-3 in combination with improved syntactic information and a revised phrasing algorithm. For evaluation we compare the output of PROS-3 under these three conditions with the reference transcription. Besides, to investigate whether the results of the evaluation study are a proper reflection of listeners' preferences for the reference transcription over PROS-3 under the three conditions, a perception experiment will be conducted.

Finally, in section 2.5, we discuss what we learned from the evaluation studies and what error-inducing factors we will investigate more thoroughly to achieve more acceptable prosody assignment.

2.2 Reference transcription

A reference transcription can be derived from spoken text or annotations of written text. Usually, the prosodic structure of a spoken version of text derived from one single speaker is taken as a reference. Since this single version is not the only proper prosodic structure for that specific text, we will compute a reference transcription from annotations of 10 experts. To come to a reference transcription there are two options. The first option is to have the annotators sit together and discuss their annotations (or realizations) to come to one 'consensus' transcription. The second option is statistical computation of an 'average' transcription. We chose to use the latter, because this is a more efficient use of resources (i.e. less time consuming for the annotators).

To be able to compute a sensible reference transcription there should be a reasonable extent of agreement between the annotators. We will assess the inter-annotator agreement in section 2.2.1. A possible argument against our approach is that the annotations may result in a different prosodic structure than the spoken versions. To exclude this discrepancy, we will validate the reference transcription through comparison of the annotated and spoken versions of three experts in section 2.2.2.

Ten experts (linguists or phoneticians familiar with the assignment of prosodic structure) were asked to read two newspaper articles (Volkskrant, 2000) and 15 e-mail messages. They should decide which words they would accentuate and where in the text they would allocate a phrase boundary when they would read the text aloud. The experts were asked to assign a prosodic structure to the text through annotation with markers for accents and phrase boundaries (four levels of boundaries were distinguished). The text consisted of 1919 words in 127 sentences. The average sentence length was 15 words.

From the gathered results of the ten experts a single prosodic structure was derived. For each juncture (i.e. a potential location for a phrase boundary) and each word, we decided whether there should be a phrase boundary or an accent respectively, according to the following procedures.

Distribution of phrase boundaries

Four phrase boundary strengths were distinguished: no boundary, weak, medium and strong boundary. For every juncture in the text, the mean score of the experts was computed by summing the scores of the ten experts ($\Sigma_{max} = 30$, with score ranging from 0 for no boundary to 3 for a strong boundary). The summed scores were mapped onto boundary values for each juncture in the following way.

0 - 6	no boundary
7 - 14	weak boundary
15 - 23	medium boundary
24 - 30	strong boundary

According to this distribution the reference transcription assigns 96 weak, 76 medium and 135 strong boundaries. The average phrase length is about 6 words.

Distribution of accents
For every word in the text the annotations of the ten experts were summed. The score for no accents is 0 and 1 for an accent ($\Sigma_{max} = 10$). Only when the score for a word was 7 or higher, the word was marked as accented in the reference transcription, otherwise the word would remain unaccented. According to this distribution the reference transcription assigns 595 accents. This means that almost one out of three words is accented.

A consequence of the criterium that an accent is only assigned at a score of 7 or higher, is that when the score for a word is 5 or 6, there will not be an accent in the reference transcription. If a system would assign an accent to such a word, we treat it as an error, although it is less problematic than when the score was 1 or 2. Figure 2.1 shows the distribution of the summed scores.

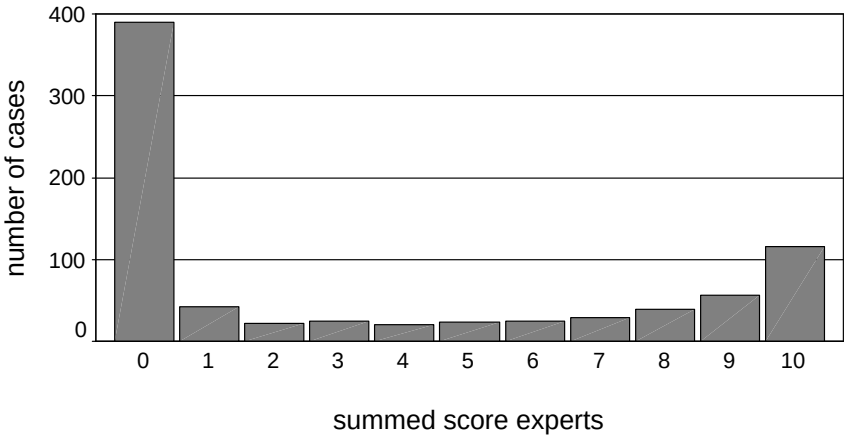


Figure 2.1: *Number of cases per summed score of the experts, for accents.*

2.2.1 Expert agreement

In order to compute a sensible mean text representation from the results of the experts, there should be a reasonable level of agreement between experts. To assess the agreement on classification tasks the kappa coefficient (Carletta, 1995) is used. The kappa coefficient κ measures pairwise agreement among a set of coders making category judgements, correcting for expected chance agreement. κ is computed according to equation 2.1.

$$\kappa = \frac{(P(A) - P(E))}{(1 - P(E))} \tag{2.1}$$

In this equation $P(A)$ is the proportion of times the coders agree and $P(E)$ is the proportion of times that the coders are expected to agree by chance.

To assess the expert agreement we selected 20 sentences from the newspaper articles. For allocation of accents κ is 0.72 (averaged over 45 sets of coders or experts), for allocation of phrase boundaries (boundary or no boundary) κ is 0.66. Content analysis researchers generally think of $\kappa > 0.80$ as good reliability, with $0.67 > \kappa > 0.80$ allowing tentative conclusions to be drawn (Carletta, 1995). Others think of $0.60 > \kappa > 0.80$ as substantial agreement (Landis and Koch, 1977; Rietveld and van Hout, 1993). Anyway, the κ-values are far from impressive.

Closer inspection of the annotations showed that there were clear differences in the numbers of accents and phrase boundaries allocated by the individual experts (see Table 2.1). We assume that the relatively modest kappa's are mainly the consequence of this variation in the amount of accents and boundaries assigned by the experts. Support for this interpretation comes from the observation that for any two experts who have approximately the same amount of accents and boundaries, pairwise kappa's are around 0.78.

Table 2.1: *Number of phrase boundaries and accents for each expert and the reference transcription.*

	boundaries				
expert	**weak**	**medium**	**strong**	**total**	**accents**
01	5	5	33	43	75
09	12	12	20	44	105
07	26	21	29	76	114
06	18	22	21	61	122
05	11	14	21	46	129
10	22	20	20	62	130
02	55	15	21	91	136
04	21	10	20	51	136
08	28	17	20	65	138
03	27	19	20	66	151
reference	16	17	20	53	112

Even when there is a clear difference in number of assigned accents and boundaries, there is a large overlap in which words are marked for accent (or which junctures are marked with a boundary). To assess the amount of overlap, we computed the kappa's for maximal and partial (half) overlap for all two combinations of experts. An illustration of this is given in Figure 2.2, which shows examples of maximal overlap and partial overlap for accent assignment by expert 01 compared with expert 03 (the extremes in number of assigned accents: respectively 75 and 151 accents) and for expert 02 and expert 08 (both in the middle of the range: respectively 136 and 138 accents).

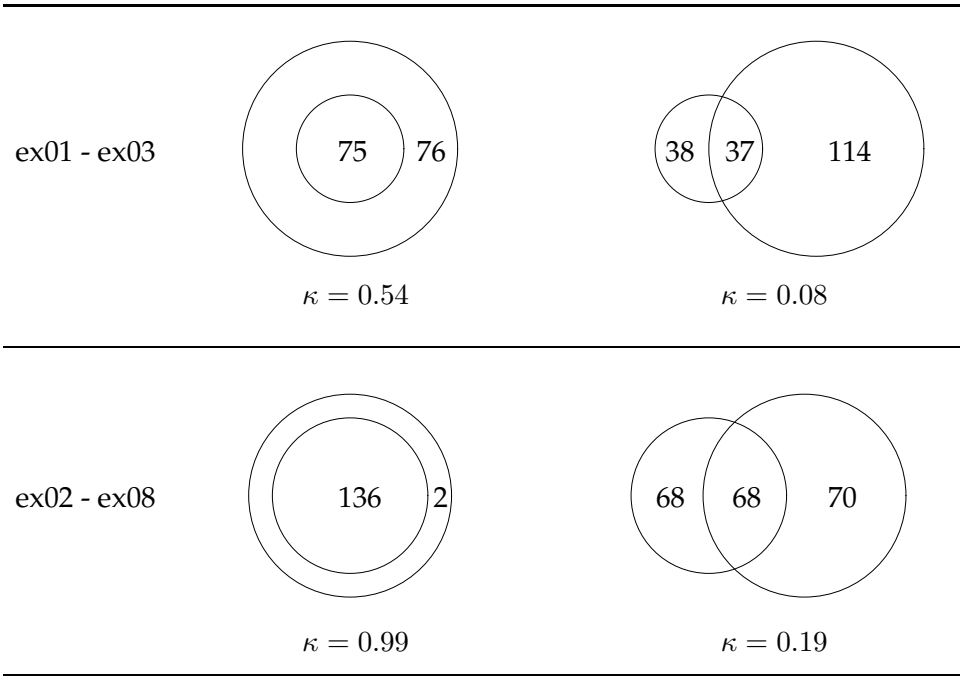


Figure 2.2: *Comparison of kappa's with maximal overlap (left side) and partial overlap (right side) of allocation of accents. The upper distributions represent the comparison of expert 01 and 03. The lower distributions represent the comparison of expert 02 and 08.*

From this we conclude that with a maximal overlap the maximal kappa's for accent assignment are within the range of 0.54 through 1.00, with an average of 0.85, which means that a mean κ of 0.72 indicates a reasonable amount of overlap. The same is true for the assignment of phrase boundaries, where the range for maximal overlap is $0.40 - 0.97$, with an average of 0.76 and the observed mean κ is 0.66, indicating a reasonable amount of overlap.

This analysis shows that the somewhat low kappa's are mainly the consequence of the different number of accents assigned by the experts. Experts who assign a large number of accents, mainly assign the same accents as experts who assign a small number of accents, and they also assign some extra accents. They do not particularly assign different accents. The same is true for boundaries.

Considering these pairwise kappa's and the reliability indicated by the kappa's according to the literature, we consider it valid to compute a single prosodic structure on the basis of the expert annotations. This 'average' prosodic structure (henceforth referred to as reference transcription) is given in Appendix A.

For comparison, the agreement between the mean κ for the three Text-to-Speech systems and the reference transcription is somewhat lower, especially for accents. Here $\kappa = 0.65$ (versus $\kappa = 0.72$ for the experts) for allocation of accents and $\kappa = 0.53$ (versus $\kappa = 0.66$) for allocation of phrase boundaries.

2.2.2 Validity of the reference transcription

As we mention in section 2.1, the annotations by the 10 experts may result in a different prosodic structure than when they actually read the text aloud. We will validate the reference transcription through comparison of the annotated and spoken versions of three experts to rule out that there exists a considerable discrepancy between the annotated and spoken versions.

Spoken reference

A production task was performed to obtain the spoken versions from the text. Three of the ten experts mentioned above were asked to read the two newspaper articles aloud. From these spoken versions of the texts 20 sentences were selected (the same sentences as mentioned in section 2.2.1). These sentences were analyzed to obtain the prosodic structures that the speakers realized. The analysis consisted of two parts: (i) pitch contour analysis together with auditory analysis to indicate which words were accented, and (ii) analysis in the time domain together with auditory analysis to indicate at which junctures phrase boundaries were realized. The strength of the phrase boundary was determined on the basis of pause duration, melodic aspects (such as continuation rise) and segmental factors (such as phrase final lengthening). An accent is assigned if there is an accent lending rise or fall in the pitch contour. A weak boundary is assigned if there is no pause or a short pause, associated with a continued high pitch and/or phrase final lengthening. A medium boundary is assigned if there is a 200–500ms pause, associated with a continuation rise and/or phrase final lengthening. A strong boundary is assigned if there is a pause with a duration longer than 500ms, associated with a continuation rise and/or phrase final lengthening.

As mentioned in section 2.2.1, there should be a reasonable level of agreement between experts to be allowed to compute a mean representation. No conspicuous differences were found in the numbers of accents and phrase boundaries allocated by the individual speakers (see Table 2.2). To assess the agreement, we computed the kappa coefficient again. For allocation of phrase boundaries $K = 0.73$ and for allocation of accents $K = 0.77$. This means that the agreement between the three experts was satisfactory (see section 2.2.1). Therefore we consider it valid to compute a mean spoken reference.

Table 2.2: *Number of accents and phrase boundaries for each expert and the mean spoken reference.*

	boundaries				
expert	**weak**	**medium**	**strong**	**total**	**accents**
04	16	16	21	53	130
05	17	14	21	52	129
08	23	11	21	55	133
reference	16	13	21	50	131

The distribution of boundaries corresponds with that described in section 2.2.1. However, as the spoken versions are produced by only three experts, we had to adjust the criteria for phrase boundary strength and accent. In this mean spoken reference a word is marked for accent when the score for that word was 2 or 3. The criteria for distribution of phrase boundaries are given below.

0 - 1	no boundary
2 - 4	weak boundary
5 - 7	medium boundary
8 - 9	strong boundary

Annotated versus spoken versions

For the three speakers the prosodic structures of their spoken versions of the 20 sentences from the newspaper articles were compared to the prosodic structures of their own annotations of the same 20 sentences. Comparing the numbers of phrase boundaries and accents for the spoken versions (Table 2.2) to those for the annotations (Table 2.1), we see that there is no large discrepancy in the number of phrase boundaries and accents between the reference transcription and the spoken reference. These results give a first impression of the capability of speakers to predict on paper which prosodic structure they would assign when they actually read the text aloud.

To obtain a more revealing view on the performance, more fine-grained measures were applied. We computed the accuracy, precision, recall and F_β-value (van Rijsbergen, 1979), which are measures typically used in the Information Retrieval domain.

Accuracy is the fraction of predictions that are correct. **Precision** is a measure of the ratio between hits and incorrect insertions (or false alarms). **Recall** is a measure of the ratio between hits and incorrect omissions (or misses). The **F_β-value** is a measure combining precision and recall.

Table 2.3: *Computation of accuracy, precision and recall.*

		reference	
		accent	no accent
prediction	accent	A	B
	no accent	C	D

Table 2.3 and equations 2.2–2.4 show how the measures accuracy, precision and recall are computed for accents. For phrase boundaries the performance measures are computed in a similar way.

$$accuracy = \frac{(A + D)}{(A + B + C + D)} \tag{2.2}$$

$$precision = \frac{A}{(A+B)} \tag{2.3}$$

$$recall = \frac{A}{(A+C)} \tag{2.4}$$

In these equations B denotes insertions and C denotes omissions. The precision becomes higher as the number of insertions decreases. The recall becomes higher as the number of omissions decreases.

The F_β-value is computed with equation 2.5.

$$F_\beta = \frac{((\beta^2+1) * prec * rec)}{(\beta^2 * prec + rec)} \tag{2.5}$$

If $\beta = 1$ the precision and recall have the same weights. If β is chosen zero, then the F_β-value equals the precision. Since we assume that precision and recall are of equal importance here, the assumption $\beta = 1$ was made.

Phrase boundaries

Precision and recall are measures for bimodal values (zero or one; present or absent). Because there are several boundary strengths, the computation of these performance measures for phrase boundaries is somewhat less straightforward. We had to find a way to derive a bimodal value from the existing four-modal value for boundaries (no boundary, weak, medium or strong boundary).

Confusion matrices were computed per expert (see Table 2.4). From these we derived a bimodal value for phrase boundaries, insertions and omissions according to two methods. One method does not take into account the boundary strength (it only makes a distinction between boundary and no boundary). However, there is a large difference between strong boundaries and weak boundaries. For phrase boundaries a standard consistency criterion is agreement within +/- 1 level (Pitrelli et al., 1994). Therefore, we used a second (rather stringent) method that does take into account boundary strength (when the system assigns a lower boundary than the experts did, we call it a quasi omission, when the system assigns a higher boundary than the experts did, we call it a quasi insertion). Quasi omissions and quasi insertions are added up to the real omissions and insertions.

We computed the performance measures according to both methods. The first method shows to what extent speakers are able to predict where they would produce a phrase boundary. The second method is even more exact, it shows to what extent speakers are able to predict where they would produce a phrase boundary and what the boundary strength would be.

Table 2.4: *Confusion matrices per expert for allocation of phrase boundaries, comparing the annotations on paper with their spoken versions.*

expert04		**annotation**			
		no	weak	medium	strong
spoken	no	301	6		
	weak	6	7	3	
	medium	2	8	6	
	strong			1	20

expert05		**annotation**			
		no	weak	medium	strong
spoken	no	301	6	1	
	weak	11	2	3	
	medium	3	3	8	
	strong			1	20

expert08		**annotation**			
		no	weak	medium	strong
spoken	no	290	15		
	weak	5	13	5	
	medium			11	
	strong			1	20

Table 2.5 gives the performance measures for allocation of phrase boundaries for the three speakers. For computation of these measures the annotations were taken as reference and the spoken versions as test case (as in Table 2.3).

Table 2.5: *Performance measures per expert for allocation of phrase boundaries, comparing annotations on paper with their spoken versions.*

	method 1				**method 2**			
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**accuracy**	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
E04	96	81	85	83	93	66	79	72
E05	94	73	84	78	92	73	63	67
E08	94	91	77	83	93	88	69	77

When we consider method 1 with respect to phrasing, the results show that the spoken versions of the sentences correspond rather well with the speakers' annotations of the sentences. The performance measures for expert 05 are somewhat lower than those for expert 04 and expert 08, but are still reasonably good. This means that speakers are capable of predicting where they would allocate phrase boundaries when reading text aloud.

When we consider method 2, the performance measures are somewhat less promising. The measures for expert 08 are still reasonably good, but the measures for expert 04 and expert 05 are worse. This means that though speakers are capable of predicting at which junctures they would allocate phrase boundaries, there is less agreement in predicting the boundary strength.

Accents

Accent is a bimodal value (accent or no accent), thus the computation of the performance measures is straightforward. We again computed confusion matrices for the three speakers (see Table 2.6).

Table 2.6: *Confusion matrices per expert for allocation of accents, comparing annotations on paper with their spoken versions.*

expert04		**annotation**	
		accent	no accent
spoken	accent	116	14
	no accent	20	210

expert05		**annotation**	
		accent	no accent
spoken	accent	101	28
	no accent	28	203

expert08		**annotation**	
		accent	no accent
spoken	accent	119	14
	no accent	19	208

Table 2.7 gives the performance measures for the allocation of accents for the three speakers. Again, the annotations were taken as reference and the spoken versions as test case. With respect to accentuation, the results show that the spoken versions of the sentences correspond rather well with the speakers' annotations of the sentences. As for allocation of phrase boundaries, the performance measures for expert 05 are somewhat lower than those for expert 04 and expert 08, but are still reasonably good. This means that speakers are capable of predicting to which words they would assign accents when reading text aloud.

Table 2.7: *Performance measures per expert for allocation of accents, comparing annotations on paper with their spoken versions.*

	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
E04	91	89	85	87
E05	84	78	78	78
E08	91	99	86	88

Reference transcription versus spoken reference

First, the number of accents and phrase boundaries allocated by the reference transcription (of all 10 experts) and the spoken reference (of 3 experts) were compared. There is no large discrepancy in the number of phrase boundaries in the two references, although the strength of the allocated boundary is not always the same. The number of accents allocated by the spoken reference is somewhat higher than the number of accents allocated by the reference transcription. Table 2.8 shows the confusion matrices for the comparison.

To obtain a more revealing view on the performance, again the accuracy, precision, recall and F_β-value were computed. Table 2.9 gives these performance measures for phrase boundaries (for both methods described above) and accents. Again, the reference transcription was taken as reference and the spoken reference was taken as test case.

Table 2.8: *Confusion matrices for allocation of phrase boundaries and accents for reference transcription versus spoken reference.*

		consensus			
		no	weak	medium	strong
spoken	no	303	7		
	weak	4	7	5	
	medium		2	11	
	strong			1	20

		consensus	
		accent	no accent
spoken	accent	106	25
	no accent	6	223

Table 2.9: *Performance measures for comparison between reference transcription and spoken reference.*

	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
bound (method 1)	97	92	87	89
bound (method 2)	95	84	76	80
accent	91	81	95	87

These results show that with respect to phrase boundary allocation, the spoken reference corresponds rather well with the reference transcription when we consider method 1. When we consider method 2, the correspondence is slightly lower, but still rather good. With respect to accent allocation the spoken reference corresponds rather well with the reference transcription.

Conclusion

Although we found reasonably high performance measures when comparing the reference transcription and the spoken reference, there still there is some variation between the annotation and spoken version for each single expert. This variation can be partly explained by the fact that there was a time span (i.e. two months) between the annotation and the production of the spoken version. We expect that the agreement would be even higher when the recordings had been made right after the annotation task. Another explanation is that there is some variation possible in the assignment of prosodic structure (even within the same context). When speakers pronounce the same sentence a number of times, they will not always realize the same prosodic structure. This means that when we compare two spoken versions of the same sentence, produced by the same speaker, we would also find some variation in prosodic structure.

From the results of the comparison between the annotations and the spoken versions for all three experts and the comparison between the reference transcription and the spoken reference we conclude that speakers are rather well capable of predicting what prosodic structure they will realize when reading text aloud. This means that annotating text is a good strategy to obtain the prosodic structure which would be realized when reading the text aloud. This strategy can be used instead of the more time consuming strategy where the prosodic structure is obtained by analysis of spoken text. This implies that we can freely use the reference transcription for our evaluation studies.

2.3 Evaluation of three TTS systems

In order to identify the major error-inducing factors for state-of-the-art automatic prosody assignment, we evaluate three Text-to-Speech systems for Dutch: Fluent Dutch, KIK and RealSpeak[1]. Fluent Dutch and KIK make use of diphone synthesis, whereas RealSpeak makes use of unit selection (Hunt and Black, 1996). In order to evaluate the acceptability of the prosodic structure generated by these systems, we compare it to the reference transcription (described above) of two newspaper articles and 15 e-mail messages.

For this comparison we had to obtain the prosodic structure that the systems assign. For FD and KIK we directly acquired the structure from the files containing the phoneme transcriptions of the sentences, in which the prosodic structure was appended. For RS this information was not directly available. The prosodic structure had to be determined on the basis of the speech output, through listening and analysis of pitch and spectral information.

[1]Fluent Dutch (FD) is a commercial product of Fluency, Van Dale Lexicography; KIK is a former joint research system by Eindhoven University of Technology, Nijmegen University and KPN (a Dutch telephone company); RealSpeak (RS) is a commercial product by the former L&H.

In van Herwijnen and Terken (2000) we showed that there is no considerable difference between the two text genres (newspaper and e-mail). The numbers of allocated phrase boundaries and accents are comparable, and so are the numbers of incorrect insertions and omissions. We also showed that for the two genres there is no appreciable difference on the textual level (i.e. sentence length and occurrence of the different syntactic categories). Considering these results, we collapsed the data for the newspaper articles and e-mail messages per system for further comparison with the reference transcription. We will perform separate analyses for the allocation of phrase boundaries and accents.

2.3.1 Phrasing

The evaluation focuses on the location of phrase boundaries and their strength. First, we compared the number of allocated phrase boundaries by the TTS systems with that allocated by the reference transcription (see Figure 2.3).

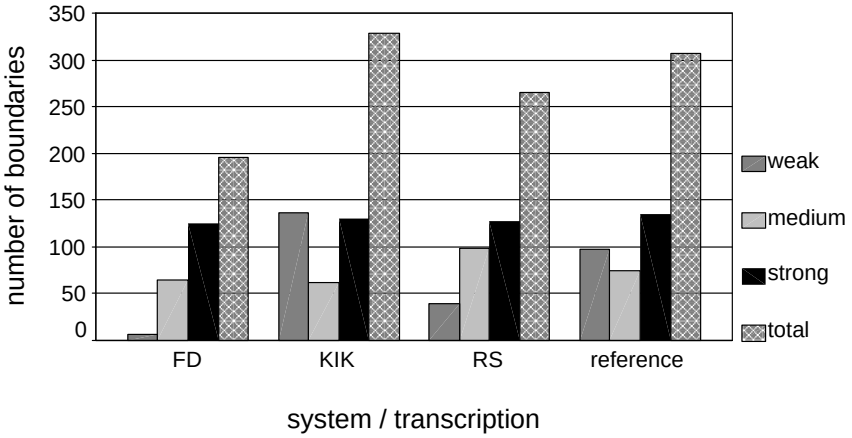


Figure 2.3: *Number of (weak, medium and strong) boundaries allocated by the TTS systems and the reference transcription.*

When we investigate the number of assigned phrase boundaries, it turns out that FD realized relatively few boundaries. Most marked boundaries here are strong boundaries. As the allocation of boundaries in this system was based on punctuation, these strong boundaries are almost completely attributable to sentence end. Sentence end was indicated by a period or carriage return. The number of strong boundaries is quite equal to the number allocated by the reference transcription. Medium boundaries could probably be ascribed to the commas. As there were very few other punctuation marks, weak boundaries are very rare.

KIK realized more boundaries than the reference transcription. The number of medium boundaries is almost equal to the number indicated by the reference transcription. The same is true for strong boundaries. The number of weak boundaries on the other hand is rather high. Further investigation showed that the realization of these boundaries

seems to be due to rhythmic aspects (i.e. after every 6 or 7 words a weak boundary was realized). Many of these are on a juncture where a boundary is undesirable (e.g. within a syntactic constituent).

RS realized more boundaries than FD and somewhat less than the reference transcription. The number of boundaries of the strong level is almost equal to that assigned by the reference transcription. The number of medium boundaries is a little higher than that indicated by the transcription, and the number of weak boundaries is considerably lower than that indicated by the transcription.

The next step is to inspect the mismatch between the systems and the reference transcription by computing the performance measures, not only for the number of assigned boundaries, but also for their strength. For this purpose we treat all mismatches as errors, meaning that both incorrect insertions (i.e. in the transcription there is no boundary allocated, whereas there is by the system) and incorrect omissions (i.e. a boundary has been included in the transcription while no boundary was generated by the system) were considered to be phrasing errors. Again, we computed the performance measures when abstracting from boundary strength (method 1), and when taking boundary strength into consideration (method 2). Table 2.10 gives the performance measures for the three systems for prosodic phrasing. The baseline indicates the performance measures for assignment of phrase boundaries on the basis of punctuation only.

Table 2.10: *Performance measures per TTS system compared to the reference transcription for allocation of phrase boundaries. Method 1 only considers the location, Method 2 also considers boundary strength.*

	method 1				**method 2**			
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**accuracy**	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
FD	94	100	64	78	93	96	58	72
KIK	92	73	78	75	91	70	73	71
RS	93	84	73	78	92	78	67	72
baseline	94	100	63	77	93	95	60	74

We found that there is a considerable discrepancy between automatically allocated phrase boundaries and boundary allocation by human experts. As we mentioned before, we consider deviations from the reference transcription to be incorrect (even though there is a certain amount of freedom of the speaker resulting in more than one proper prosodic structure per sentence). The accuracy and F_β-value are rather similar for the three systems. However, the three systems perform differently in terms of precision and recall. FD has almost no incorrect boundary insertions, although this is at the expense of allocating very few phrase boundaries. As a result FD generates rather long phrases. KIK, on the other hand, has many incorrect boundary insertions, which are mainly weak boundaries. To prevent these incorrect insertions proper syntactic analysis is required. RS allocates slightly fewer boundaries than the experts do, and the

distribution over boundary level resembles that of the experts reasonably well. However, this does not necessarily mean that the prosodic structure assigned by this system is acceptable. We observed that many of the boundaries annotated for RS may have been spurious due to the pitch discontinuities that are inherent in the unit selection approach, but qualify nevertheless as boundaries on the basis of melodic criteria.

2.3.2 Accentuation

First, we compared the number of allocated accents by the TTS systems with that allocated by the reference transcription (see Figure 2.4).

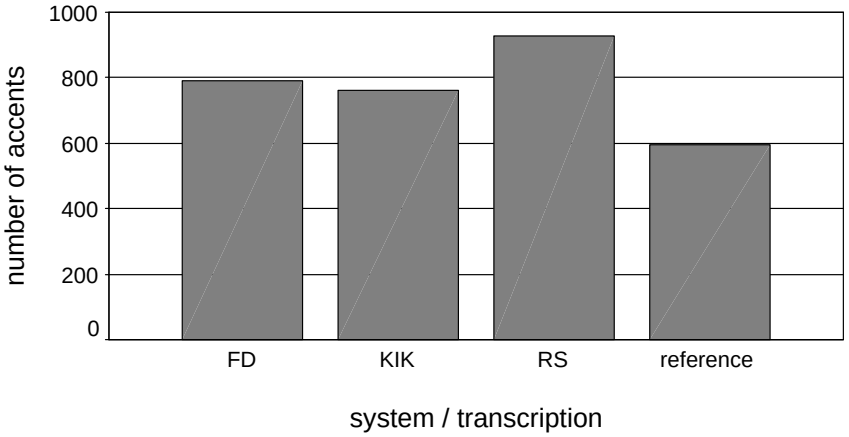


Figure 2.4: *Number of accents allocated by the TTS systems and the reference transcription.*

All three systems realized considerably more accents than the reference transcription. For FD and KIK this is due to the fact that they assign accents on the basis of Part-of-Speech information only and have a strong tendency to accent all content words. For RS, the fact that so many words are accented is dependent on the accidental properties of the concatenation units in the database.

We found that there is a considerable discrepancy between automatically allocated accents and allocation by human experts. However FD has the highest accuracy and F_β-value and KIK has the lowest values, the performance values (see Table 2.11 where the baseline indicates the performance measures for assignment of accent on all content words and no accents on function words.) do not differ very much for the three systems. RS has a reasonably high recall value. This is due to the number of assigned accents. Logically, if many accents are assigned, there will be only few accent omissions (which results in a high recall). However, the precision is quite low, due to the large number of incorrect accent insertions. Overall, automatically generated prosodic structure contains too many accents and they are not distributed correctly over the prosodic domains.

Table 2.11: *Performance measures per TTS system compared to the reference transcription for allocation of accents.*

	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
FD	79	62	82	71
KIK	76	59	76	66
RS	75	56	87	68
baseline	60	44	98	61

2.3.3 Error analysis

In order to get an indication of the causes for discrepancies between the prosodic structures as assigned by the TTS systems and the reference transcription, we conducted a more detailed analysis of the mismatches. We investigated several factors that may be held responsible for these mismatches, such as lexical and syntactic factors and textual influences. We will investigate these factors for both prosodic phrasing and accentuation.

Phrasing

We performed a manual analysis of the phrasing errors in the newspaper articles for the three TTS systems compared to the reference transcription. We distinguished five factors: first major constituent, punctuation, conjunctions, enumerations and boundaries within syntactic phrases.

In sentences with a long first major constituent (as in example 2.6a) speakers often realize a weak boundary (indicated by /) after this constituent. This makes the sentence easier to comprehend. On the other hand, insertion of a boundary within the first constituent disturbs the listener.

As punctuation is helpful to readers, allocating boundaries on account of punctuation helps the listener. In example 2.6b, a strong boundary should be allocated at the end of the sentence, which is indicated by a period. A strong boundary is also allocated if there is a semicolon in the text. A medium boundaries is allocated if there is a comma, colon or parenthesis.

Conjunctions, such as '*doordat*' (because), indicate a contrast or reason. Speakers use the strategy of allocating a medium boundary to emphasize this contrast (as in example 2.6c). To make clear that a sequence of nouns is an enumeration, speakers use the strategy to insert a weak boundary after each part of the enumeration and close the enumeration with a medium boundary (as in example 2.6d). With respect to boundaries within syntactic phrases (example 2.6e), insertion of such boundaries are often disturbing to the listener, because it effects the meaning of the sentence. Thus, in this example, the weak boundary is incorrect.

(2.6) (a) *Pakweg drie jaar geleden / opperde de liberale*
About three years ago / proposed the liberal
jongerenorganisatie / om van het koningschap een louter ceremoniële
youth organization / of the monarchy a purely ceremonial
functie te maken.
function to make.
"About three years ago, the liberal youth organization proposed that the monarchy should have a purely ceremonial function."

(b) *Het andere deel der natie omhelst hem. ///*
The other part of the nation embraces him. ///
"He is embraced by the other part of the nation."

(c) *Veel toeschouwers merkten alleen iets van de eclips //*
Many observers perceived only somewhat of the eclipse //
doordat het enkele graden kouder werd.
because it by several degrees colder got.
"Many observers only perceived somewhat of the eclipse because it got colder by several degrees."

(d) *Energiebedrijf NUON (Gelderland / Flevoland / en Friesland)*
Energy company NUON (Gelderland / Flevoland / and Friesland)
// besloot ...
// decided ...
"Energy company NUON (Gelderland, Flevoland and Friesland) decided ..."

(e) *De ernst / van haar letsel is nog onbekend.*
The seriousness / of her injury is still unknown.
"The seriousness of her injury is still inknown."

Table 2.12 shows the percentages of incorrect boundary insertions and omissions per factor and per system. These data show that the factors punctuation, first major constituent and boundary insertion inside a syntactic phrase are the major factors that cause phrasing errors. The last two factors are syntactic factors, which means that we would need to obtain proper syntactic analyses to prevent such errors.

As we saw before, FD has almost no incorrect boundary insertions, because phrasing is done on the basis of punctuation. The numbers of incorrect insertions and omissions are more equal for KIK and RS. These two systems have many incorrect boundary insertions which are due to the factor syntactic phrase. This means that there often is a boundary inserted between two words in the same syntactic phrase (as in example 2.6e). Across systems the major problem is the allocation of weak boundaries at the right positions. In order to solve this problem, phrasing algorithms need correct syntactic analyses, where we should focus on preventing allocation of boundaries within syntactic phrases. Since we found that many occurrences of an incorrect inserted boundary within a syntactic constituent are between an NP and a PP, the main

phrasing problem we will deal with in this thesis is the allocation of boundaries at junctures preceding (attached) PP's.

Table 2.12: *Percentage of incorrect phrase boundary insertions and omissions per factor, for the three TTS systems.*

	FD		KIK		RS	
factor	ins.	om.	ins.	om.	ins.	om.
first major constituent	0	38	12	8	23	0
punctuation	2	0	1	25	0	2
conjunction	0	20	0	5	0	11
enumeration	0	6	0	1	3	2
insertion inside syntactic phrase	0	0	23	0	27	0
other	0	8	2	1	0	7
wrong level	2	24	1	21	3	22
total	4	96	39	61	56	44

Accentuation

A Part-of-Speech tagger for Dutch[2] was used to determine the syntactic categories of all words in the two newspaper texts. The number of words per Part-of-Speech (POS) was counted and the percentages of incorrect accent insertions and omissions were computed for all categories per TTS system. Table 2.13 shows the number of words in each category together with the percentage of errors.

Table 2.13: *Percentages of incorrect accent insertions and omissions per POS, for the three TTS systems.*

		FD		KIK		RS	
POS	**# of words**	ins.	om.	ins.	om.	ins.	om.
Noun	122	39	0	38	4	39	2
Verb	119	39	0	23	7	31	3
Preposition	115	0	2	0	2	0	2
Article	111	0	0	1	0	0	0
Adverb	56	7	23	30	11	19	17
Adjunct	51	24	10	20	20	22	15
Pronoun	51	0	20	6	14	3	17
Conjunct	40	0	5	0	3	0	4
Numeral	20	5	15	10	25	8	20

The error percentages of the three systems are quite the same. Overall the percentage of errors in the noun and verb category seems much higher than the percentage of errors in the other categories. When looking at the percentage of incorrect accent insertions

[2]The tagger used here is the MBT tagger (Daelemans et al., 1996), developed at ILK, Tilburg University. This tagger is a fast and accurate POS tagger that is automatically generated from a tagged example corpus by Memory-Based Learning techniques.

and omissions we observe that by and large the categories of content words (noun, verb, adjective, adverb and numeral) show mainly incorrect insertions, whereas the categories of function words (preposition, article, pronoun and conjunct) show mainly incorrect omissions. As such this overview gives no information about the underlying cause of the mismatches. Therefore, by manual analysis we distinguished several factors that caused phrasing errors. These factors are discourse context (e.g. given vs. new information), rhythmic considerations, lexical considerations, syntax and other.

Since the constituents '*De Graaf*' and '*lid van D66*' in example 2.7a had already been mentioned in the previous sentence, these words are not accented due to the fact that it is given information in the discourse context. The experts annotated the entire texts at once, so they could use the context information. As mentioned before, we do not deal with contextual errors.

The word '*plaatsje*' in example 2.7b should not be accented for rhythmical considerations, because in that case three successive words would be accented. This would not sound natural, as in such cases, speakers will deaccent the middle word.

With respect to lexical considerations (example 2.7c) the word '*anders*' always indicates a contrast. The lexicon should provide information that this words always should be accented. With respect to syntax (example 2.7d) the verb '*geopereerd*' should be accented because it is preceded by the condition '*met spoed*', which is a different focus domain.

(2.7) (a) *De *Graaf is niet bij *toeval *lid van *D66 ...*
De Graaf is not by accident a member of D66 ...
"De Graaf is not an accidental member of D66 ..."

(b) *Bij het *Belgische *plaatsje *Virton ...*
At the Belgian town Virton ...
"At the Belgian town Virton ..."

(c) *Anders dan in het *verleden ...*
Differently than in the past ...
"Differently than in the past ..."

(d) *Het *kind is met *spoed *geopereerd.*
The child has urgently had an operation.
"The child has had an urgent operation."

We computed the percentage of incorrect insertions and omissions per factor per TTS system. These percentages are shown in Table 2.14. The total percentage of accent insertions is remarkably higher than the percentage of omissions. This is due to the fact that all systems assigned many more accents than the reference transcription. Syntax is the major factor which causes errors. Moreover, almost all accent omissions are due to a lack of syntactic information (or incorrect information). Therefore, syntactic errors

need to be dealt with by better syntactic analysis of the text. Lexical errors can be dealt with to some extent by adjusting their accentability status in the lexicon. Rhythmical factors can be partly dealt with provided that the syntactic analyses delivers correct syntactic groupings. In order to deal with contextual errors we would need discourse modelling. Since syntax is the main error-inducing factor we will try to reduce this type of errors, starting with an investigation of the effect of proper syntactic analysis. We found that many occurrences of an incorrectly inserted or omitted accent are on sentence final verbs, therefore the main problem of accentuation we will deal with in this thesis is accenting the sentence final verb when it is preceded by an argument or a condition.

Table 2.14: *Percentage of incorrect accent insertions and omissions per factor, for the three TTS systems.*

	FD		**KIK**		**RS**	
factor	ins.	om.	ins.	om.	ins.	om.
context	25	2	24	3	22	1
rhythmic	16	0	13	1	12	1
lexical	2	4	6	3	10	0
syntax	36	12	34	15	34	14
other	2	1	0	1	4	2
total	81	19	77	23	82	18

As for phrasing, context effects are outside the scope of this thesis, therefore we will not investigate methods to prevent contextual errors (e.g. accentuation errors against given/new information).

2.4 Evaluation of PROS-3

In this section we describe the evaluation of PROS-3 (Dirksen, 1994). PROS-3 implements a theory about the assignment of prosodic structure on the basis of syntactic and pragmatic information. At the same time, it constitutes a module of a system that generates prosody on the basis of syntactic information produced by a state-of-the-art syntactic parser. In the latter context, it was observed that the assignment of phrase boundaries and accents is often inadequate. However, it is unclear whether this is due to the inadequate syntactic information or to inadequacies in the theory underlying PROS-3 or both.

The evaluation described in this section is divided into two stages: (i) the performance of PROS-3 when compared to the prosodic structure as assigned by the reference transcription, and (ii) the performance of PROS-3 compared to the prosodic structure as assigned by the reference transcription as judged in a perception experiment.

PROS-3

PROS-3 is a system that assigns prosodic structure to text on the basis of a syntactic representation of the input text. This syntactic representation describes the word category of each word together with the relations between the words.

PROS-3 is applied in two steps. First, a binary branching tree is computed on the basis of the syntactic representation. This metrical tree specifies the weak-strong relations between the sister nodes in the syntactic representation and which syntactic categories are eligible for Focus. Next, the metrical tree is turned into a prosodic structure specifying the location of phrase boundaries and accents within sentences. Strong boundaries are assigned on the basis of punctuation: period and semicolon are indicators for a strong boundary. Within sentences, PROS-3 determines the location of boundaries of Intonational (or I-) phrases and boundaries of Phonological (or Phi-) phrases. I-phrases are application domains for rules that specify intonation and are often separated by a speech pause and marked by a pitch movement; I-phrase boundaries are realized as medium boundaries. Phi-phrase boundaries are application domains for supra-segmental phonological rules (e.g. they block coarticulation); Phi-phrase boundaries are realized as weak boundaries.

The allocation of accents within phrases is based on the Focus-Accent Theory (Baart, 1987). One or more constituents of a sentence are marked as +F(ocus). The relation between +F and its realization as an accent located on a word is mediated by the metrical tree: one daughter of a branching node is characterized as strong and the other as weak, depending on the functor-argument relation between the two daughter nodes. The grammar specifies which phrasal categories are eligible for focus. In PROS-3 accent is assigned to individual words (Dirksen, 1994). The lexicon, used by PROS-3, specifies that certain words (e.g. pronouns) are typically deaccented. This may block the Focus rule from applying. In these cases, deaccentuation of words affects the strong-weak labelling. Finally, accents allocated to words within a sequence of accented words within a prosodic domain may be deleted for rhythmical reasons.

Since the metrical tree is constructed on the basis of the output of the syntactic parser, it is obvious that both phrasing and accentuation are strongly influenced by the performance of the parser. Therefore, we investigate the improvement in performance when PROS-3 is applied on the basis of correct syntactic information.

2.4.1 Comparison with reference transcription

We evaluate the assignment of prosodic structure using PROS-3 according to three protocols: (A) evaluation of PROS-3 on the basis of automatically derived syntactic structure by a robust parser, (B) evaluation of PROS-3 on the basis of correct syntactic structure and (C) evaluation of PROS-3 on the basis of correct syntactic structure, in combination with a revised algorithm for prosodic phrasing. The prosodic structure as assigned according to these three protocols will be compared to the prosodic structure as assigned by the reference transcription.

Sentence 2.8 shows one of the sentences that was presented to the experts and PROS-3. Sentence 2.9 shows an example of the reference transcription of the sentence. Accents are indicated by *, phrase boundaries by /, where the number of slashes indicates the boundary level.

(2.8) *Hoezeer er ook een verbod geldt op het lekken*
However much there also a ban is on leaking information
uit de ministerraad, toch is het beraad niet supergeheim.
from the council of ministers, still is the meeting not topsecret.
"However much there is a ban on leaking information from the council of ministers, the meeting is not topsecret."

(2.9) *Hoezeer er ook een *verbod geldt op het *lekken uit de *ministerraad // *toch is het beraad *niet *supergeheim ///

Evaluating the performance of PROS-3 in combination with a robust parser

Next to its status of being a theory about prosodic structure, PROS-3 has also been implemented as a module of a system for Text-to-Speech conversion in the POLYGLOT project. In order to provide this implementation of the PROS-3 algorithm with the desired syntactic information, a robust parser, STP, was developed as part of the POLYGLOT project (Dirksen, 1992b).

This parser provides a syntactic representation for every input text. Parts of the sentence that cannot be integrated into the syntactic representation are left unanalyzed and connected to the root node. The grammar rules contain information about functor-argument relations between adjacent syntactic categories that is needed by PROS-3 to convert the syntactic representation into a metrical representation. In addition, the grammar rules contain information about phrasing: boundaries between major syntactic constituents are hard-coded as prosodic phrase boundaries. Elements of the sentence for which no analysis can be provided are assigned the category of major constituent and realized as separate prosodic domains and accented as such.

Since STP provides the information that is needed to drive PROS-3, the combination of STP and PROS-3 was used to evaluate the performance of PROS-3 in combination with a state-of-the-art robust parser. This performance constitutes a baseline. The procedure described here was followed for twenty sentences from the newspaper articles. Sentence 2.10 shows an example of the output of PROS-3. Compared to the reference transcription PROS-3 assigns more phrase boundaries and these are often allocated at junctures that are not eligible for a boundary. PROS-3 also assigns more accents.

(2.10) Hoezeer er ook een *verbod / geldt *op het lekken uit de *ministerraad //
*toch / is het *beraad / *niet / *supergeheim ///

Evaluation of PROS-3 based on improved syntactic structure

As allocation of accents and phrase boundaries is strongly dependent on syntactic structure, we expect that correct syntactic structure will lead to more adequate allocation of accents and phrase boundaries by PROS-3. The robust parser providing the input for the PROS-3 module does not always yield a proper syntactic tree. In order to test the appropriateness of the prosody assignment by PROS-3 with correct syntactic input, the syntactic tree is manually edited to obtain a proper syntactic structure. Nevertheless, there remain some phrasing and accentuation errors as can be seen in Sentence 2.11, which shows an example of the output of PROS-3 with improved syntactic input. This means that only improving syntactic information is not a sufficient solution.

(2.11) Hoezeer / er ook een *verbod geldt // op het *lekken / uit de *ministerraad // *toch / is het *beraad / *niet / *supergeheim ///

Evaluation of PROS-3 on the basis of improved syntactic structure, in combination with a revised algorithm for prosodic phrasing

When providing PROS-3 with adequate syntactic information, it became clear that the phrasing algorithm implemented in PROS-3 gave rather poor results. For that reason, an alternative phrasing algorithm was defined. This algorithm consists of three steps:

- *Step 0*: Assignment of strong and medium boundaries, based on punctuation.
- *Step 1*: Assignment of medium boundaries, based on length of prosodic phrases and syntactic structure.
- *Step 2*: Assignment of weak boundaries, based on length of prosodic phrases and syntactic structure.

With regard to *Step 0* a strong boundary (Type 3) is assigned at the location of a period or semicolon in the sentence. A medium boundary (Type 2) is assigned at the location of a comma, colon or parenthesis in the sentence. For *Step 1* the length of the prosodic phrase should be determined (by counting words). If the phrase contains more than 16 words, a medium boundary is allocated in the middle of the phrase if this is not in conflict with the syntactic domains. If the middle of the phrase is within a syntactic constituent, the medium phrase boundary is allocated at the nearest (left or right) syntactic domain boundary. The procedure of *Step 2* resembles that of the previous step, but here the maximum phrase length is 8 words and weak boundaries (Type 1) are allocated instead of medium.

The information concerning the location of phrase boundaries as determined by the algorithm is then incorporated in the syntactic representation of the sentence. This syntactic and phrasal information constitutes the input for the procedure that assigns

accents to words in the sentence. The accentuation algorithm is the same as in the other protocols. Again, these procedures are performed on proper syntactic input, which prevents phrase boundaries within syntactic constituents. Sentence 2.12 shows an example of the output PROS-3 in combination with proper syntactic structure and the revised phrasing algorithm. This example shows that the allocation of phrase boundaries leads to an appreciable improvement. The allocation of accents has not changed.

(2.12) Hoezeer er ook een *verbod geldt / op het *lekken uit de *ministerraad //
*toch is het *beraad *niet *supergeheim ///

We have applied the maximum phrase length for *Step 1* and *Step 2* at 16 and 8 words respectively. These numbers are determined by analysis of the transcriptions by the ten phonetic experts. For faster or slower speech, one can easily alter these maximum phrase lengths, resulting in an acceptable number of boundaries for that specific speech rate.

Results phrasing

With respect to phrasing there are two main questions. The first is whether a better syntactic information leads to a better performance of the procedure for assignment of phrase boundaries. The second is whether the revised algorithm for allocation of phrase boundaries performs better than the old algorithm.

For the three protocols we counted the number of correct phrase boundaries, incorrect boundary insertions and omissions assigned to the 20 sentences from the two newspaper articles. These numbers give a first impression of the performance of PROS-3 (see Figure 2.5).

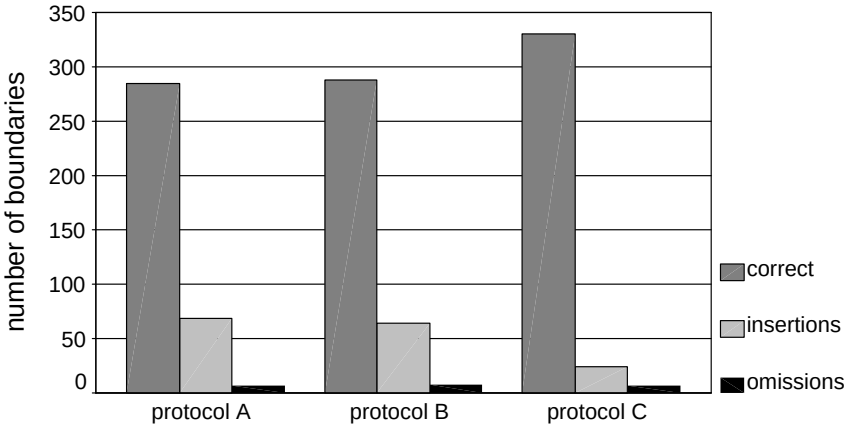


Figure 2.5: *Number of correct boundaries and incorrect boundary insertions and omissions per protocol compared to the reference transcription. Protocol A is PROS-3 in combination with a robust parser, protocol B is PROS-3 on the basis of correct syntactic structure and protocol C is PROS-3 on the basis of correct syntactic structure and the revised phrasing algorithm.*

When we compare the numbers of correct boundaries and incorrect boundary insertions and omissions for the three protocols, there is no noticeable difference between protocols A and B, but protocol C performs differently. We see that the revised algorithm induces fewer incorrect insertions. The performance measures accuracy, precision, recall and F_β-value were computed for more exact results.

To obtain a bimodal value (which is necessary for the computation of the performance measures) for phrase boundaries, insertions and omissions can be computed by two methods (as was done in section 2.2.2). Method 1 abstracts away from boundary strength (it only makes a distinction between boundary and no boundary). Method 2 also considers boundary strength.

Table 2.15: *Performance measures for allocation of phrase boundaries, per protocol, compared to the reference transcription.*

	method 1				**method 2**			
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**accuracy**	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
protocol A	79	40	88	55	76	31	74	44
protocol B	80	41	85	55	76	31	66	42
protocol C	92	66	88	75	88	53	70	60

Table 2.15 shows that for allocation of phrase boundaries the performance of protocol C (the protocol with the revised algorithm for assignment of phrase boundaries) is substantially better than that for protocols A and B. This improvement is true for all four measures and for both methods.

Results accentuation

With respect to accentuation our main question is whether a better syntactic input leads to a better performance of the procedure for assignment of accents. For the three protocols we counted the number of correct accents, incorrect accent insertions and omissions for the 20 sentences from the two newspaper articles. These numbers give a first impression of the performance of PROS-3 (see Figure 2.6).

When we compare the numbers of correct accents and incorrect accent insertions and omissions for the three protocols, we see that correct syntactic structure leads to improved accent assignment, as protocol B and C have more correct accents and fewer incorrect accent insertions. The revised phrasing algorithm does not induce an accentuation improvement. The performance measures accuracy, precision, recall and F_β-value were computed for more exact results.

Table 2.16 shows that for allocation of accents protocol B and protocol C perform better than protocol A. This means that PROS-3 performs better with correct syntactic input than with inaccurate syntactic structure. The performance of protocol B and C is comparable, indicating that the revised phrasing algorithm had no effect for improv-

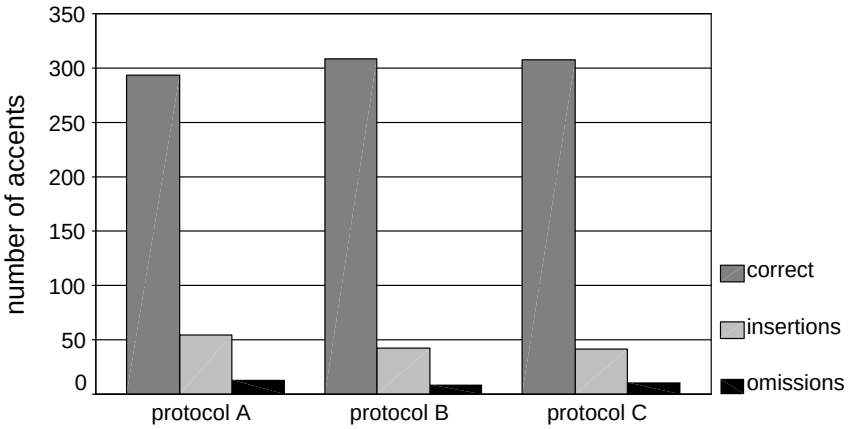


Figure 2.6: *Number of correct accents and incorrect accent insertions and omissions per protocol compared to the reference transcription. Protocol A is PROS-3 in combination with a robust parser, protocol B is PROS-3 on the basis of correct syntactic structure and protocol C is PROS-3 on the basis of correct syntactic structure and the revised phrasing algorithm.*

Table 2.16: *Performance measures for allocation of accents, per protocol, compared to the reference transcription.*

	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
protocol A	82	65	88	75
protocol B	86	71	92	80
protocol C	86	71	90	79

ing accentuation, unlike our expectations that improved prosodic phrasing would also result in improved accentuation, because the algorithm assigns at least one accent in every prosodic phrase.

When assigning prosodic structure with PROS-3 we obtain the best result if we apply PROS-3 in combination with the revised phrasing algorithm on correct syntactic information. For all protocols there is a certain amount of discrepancy between the reference transcription and automatically generated prosody. However, impressionistically, prosodic phrasing provided by protocol A and B is often inadequate, while that for protocol C is usually appropriate. With respect to accentuation, providing PROS-3 with correct syntactic information already gives a considerable improvement, but this can only be appreciated in combination with adequate phrasing. A perception experiment was performed to put this impression to the test.

2.4.2 Perception experiment

Ostendorf and Veilleux (1994) already suggested that "the best test of a phrase break algorithm is in perceptual judgements of synthesized speech", in our opinion this holds

not only for prosodic phrasing but also for accentuation. Therefore, the perception experiment presented here puts the results of the comparison between the assignment of prosodic structure by the three protocols and the reference transcription to the test. In this experiment listeners judgements were collected about acceptability of the prosodic structure.

Method

The 20 sentences mentioned in section 2.4.1 were processed by Calipso Text-to-Speech synthesis. Grapheme input was processed by this system, resulting in a phoneme representation, which was corrected manually. Furthermore, the prosodic structures resulting from the three protocols and the reference transcription were assigned. For each sentence four versions were generated (versions with the prosodic structure perceived by protocol A, B and C and the reference transcription (referred to as protocol H)).

The sentences were presented pair-wise to 20 listeners. These pairs consisted of two versions of one sentence. Pairs were presented in all possible sequences ($20*4*3 = 240$ pairs). The sentences were presented over headphones. Because of the duration of the experiment we presented only half of the stimuli (120) to each listener (partitioning by Latin-square). Listeners were asked to indicate which of the two sentences was the most acceptable by clicking with the mouse on a button on the screen. All listeners were native speakers of Dutch and none of them reported hearing problems. They were all students in the age of 18 through 29 and they were not familiar with the research which this perception experiment was part of.

Results

Figure 2.7 shows to what extent each protocol is preferred by the listeners. If the listeners had no preference for one of the protocols, all protocols would be rated for 25% (indicated by the reference line). The values in Figure 2.7 indicate that protocol B is not preferred over protocol A, that protocol C and H are highly preferred over protocol A and B, and that the listeners slightly prefer protocol H over protocol C.

A more detailed comparison is made in Table 2.17. This table shows that version H and C are significantly (according to the binomial test, with $p < 0.01$) preferred over version A and B. There is also a significant preference for version H when comparing it with version C.

The results of the perception experiment show that listeners have no preference for sentences generated on the basis of PROS-3 with automatically derived syntactic structure or sentences generated on the basis PROS-3 with correct syntactic structure. From this we conclude that improved syntactic structure alone does not improve the acceptability of the prosodic structure, according to the listeners' judgements. However, the difference between these two protocols and the protocol of the sentences based

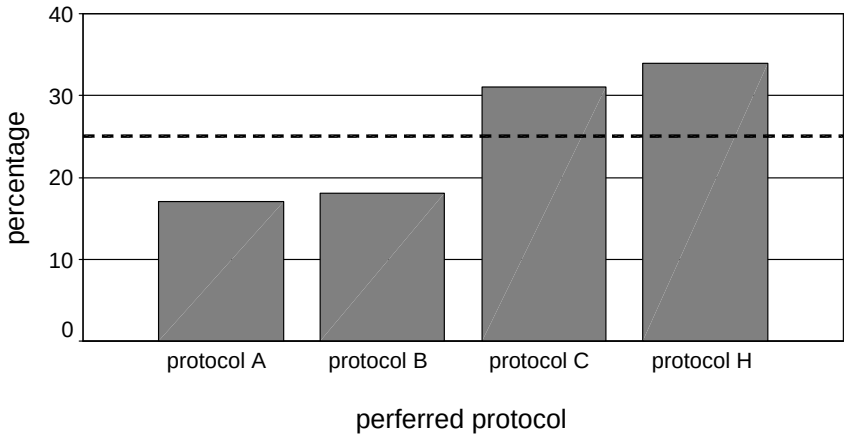


Figure 2.7: *Overall extent of listeners' preference per protocol.*

Table 2.17: *Listeners' preferences (in percentages) for the second version, when comparing two versions of the same sentences.*

version 1	**version 2**	**preferred 2**
A	B	51
A	C	70*
A	H	74*
B	C	69*
B	H	74*
C	H	58*

*significant with $p < 0.01$

on correct syntactic structure in combination with the revised algorithm for prosodic phrasing, leads to a major preference for the latter protocol over the other two. Still, listeners have a slight (but not significant) preference for the reference transcription (protocol H) above the most acceptable algorithm-based protocol C.

From the results of the objective evaluation of the performance of PROS-3 when compared to the reference transcription we conclude that we obtain the best result when applying PROS-3 in combination with the revised algorithm for prosodic phrasing, on correct syntactic input. Providing PROS-3 with correct syntactic information already gives a considerable improvement with respect to accentuation, but we notice that this can only be fully appreciated in combination with adequate phrasing.

2.5 Discussion and conclusion

In this chapter we showed that computing a reference transcription from the annotations of text by experts is a good alternative for the derivation of a human reference through analysis of spoken versions of text. The reference transcription we obtained with this approach was used for the evaluation of three state-of-the-art Text-to-Speech systems for Dutch and PROS-3 under three conditions.

The evaluation of the TTS systems showed that there are substantial differences between automatically generated prosodic structure and the prosodic structure as assigned by human experts. The analysis shows that a considerable number of incorrect insertions and omissions is made by the systems. These are merely due to incorrect or insufficient syntactic information about the input text.

Next, we evaluated PROS-3 as such, since it is the algorithm which we will use as a starting point for our attempts on creating an improved module for assignment of prosodic structure. As mentioned before we learned that incorrect or insufficient syntactic information was the major error-inducing factor for allocation of phrase boundaries and accents. Therefore, we made a first attempt to investigate the effect of proper syntactic information by evaluating PROS-3 in combination with such information (derived through manual correction of the output of the robust parser). Since PROS-3 assigns too many phrase boundaries, we defined a revised phrasing algorithm. We also evaluated PROS-3 in combination with proper syntactic information and the revised phrasing algorithm. These evaluation studies showed that proper syntactic information alone did not improve the allocation of phrase boundaries, but did do so in combination with the revised phrasing algorithm. Proper syntactic information did improve the allocation of accents and the revised phrasing algorithm did not make a difference here.

A perception experiment in which listeners had to indicate which version of a sentence was the most acceptable, showed that they highly prefer the version based on correct syntactic information in combination with the revised algorithm for prosodic phrasing. Still, there is a slight preference for the version based on the reference transcription. These results support our expectation that improved accentuation can only be appreciated in combination with adequate phrasing.

From the error analysis we learned that the major problem for phrasing turns out to be allocation of boundaries at junctures preceding an attached prepositional phrase. The major problem for accentuation is allocation of accents at sentence final verbs preceded by a nominal constituent. Therefore, in following chapters we will focus on finding a method for predicting the status of the PP (i.e. noun or verb attachment) and the status of the nominal constituent that precedes the sentence final verb (i.e. argument or condition). This will be done after a study on computing the perceptual costs of these types of errors.

3 Tolerance for errors

In this chapter we describe two experiments we perform to investigate the effect of prosodic phrasing and accentuation on the acceptability of synthetic speech. In the first experiment we show that listeners are more tolerant towards an incorrect omitted phrase boundary than towards an incorrect inserted boundary at the juncture preceding an attached PP. Thus, we rather allocate too few boundaries than too many. This implies that in machine learning experiments we should have a bias for predicting noun attachment, because for noun attached PP's correct phrasing means that there is no boundary allocated preceding the PP, whereas for verb attached PP's there is. Furthermore, we show that we had better not allocate a medium boundary preceding an attached PP. In the second experiment we show that incorrect accent insertions on a sentence final verb are as bad as accent omissions. Thus, we should find an optimum in accent allocation, such that there are as few accent insertions and omissions as possible. This implies that in machine learning experiments we should be as good in predicting arguments as in predicting conditions. Finally, when comparing two different pitch contours there is no clear evidence of an effect of the shape of pitch contour.

3.1 Introduction

From the previous chapters we learned that in state-of-the-art Text-to-Speech synthesis systems assignment of prosodic structure (accents and phrase boundaries) is not yet a solved problem. Accents and phrase boundaries are often omitted or allocated in the wrong places. Previous research (e.g. Nooteboom and Kruyt, 1987; Sanderman and Collier, 1997) showed that correct prosodic information helps the listener when processing text, whereas incorrect prosodic structure may impede the listener's comprehension. This means that it will take more time and effort from the listener when processing speech with incorrect prosodic structure, and in the worst case the listener might not understand the conveyed information correctly.

We introduce the contrast *correct* versus *incorrect* prosodic structure. In this context, the notion 'correct' means that a boundary or accent is allocated (or not) according to the syntactic structure, whereas 'incorrect' means that a boundary or accent is allocated (or not) in contradiction to the syntactic structure. Other contrasts that we refer to in this chapter are *insertion* versus *omission* and *acceptable* versus *unacceptable*. The notion 'insertion' means that a boundary or accent is allocated where there should not be one according to the syntactic structure. The notion 'omission' means that a boundary or accent is not allocated where there should have been one according to the syntactic structure. The notion 'acceptable' means that listeners approve with the phrasing structure or accentuation structure, whereas 'unacceptable' means that the listeners disapprove of it.

For the assignment of prosodic structure we focus on allocation of prosodic phrase boundaries and accents. Prosodic phrasing indicates which parts of an utterance belong together, syntactically and semantically (Bolinger, 1989; Sanderman, 1996). This information is used by the listener to deduce the relations between the words in the sentence. For instance, adjectives provide information about the status of a noun. When there is a phrase boundary inserted (i.e. incorrectly allocated) between the adjective and the noun, the listener might wrongly conclude that there is no semantic or syntactic relation between the two words.

A similar problem occurs when the accentuation of a sentence is incorrect. Accents provide information about which words the listener should pay attention to. Chafe (1974) states that accents highlight the information that should be at the center of attention of the listener. However, when a word is accented while it should not have been, the attention of the listener is erroneously attracted by that word. In this case the listener might have less attention for the more important words, which means that he should deduce the information that was meant to be provided, by back-tracking the sentence.

From the evaluation study (described in Chapter 2) we learned that many phrasing errors occur at junctures preceding the PP in [NP PP] or [PP PP] sequences. Therefore, we focus on the allocation of prosodic phrase boundaries at junctures preceding a

prepositional phrase in Dutch. We only consider sentences in which the PP is preceded by a nominal phrase or another prepositional phrase. In certain sentences a prosodic phrase boundary can be realized at such a juncture (indicated with [ ] in example 3.1), whereas in other sentences a phrase boundary preceding the PP is incorrect. The appropriateness of a phrase boundary depends on the status of the PP. If the PP is noun attached (as in example 3.1a), a phrase boundary preceding the PP is undesirable. If the PP is verb attached (as in example 3.1b), a phrase boundary preceding the PP is possible, although not mandatory.

(3.1) (a) *He accused the president [ ] of the National Bank.*

(b) *He accused the president [ ] of the bank robbery.*

Since TTS-systems will always make some errors when performing phrase boundary allocation, we want to find out which type of error is the least problematic for the listener (i.e. the least unacceptable; the type towards which the listeners is the most tolerant), so that we can try to shift the number of phrasing errors in the direction of the one that is less problematic. We perform a perception experiment in which subjects are asked to indicate their preference[1] for the sentence with or without a phrase boundary preceding the PP. We assume that the tolerance for errors can be interpreted as an indicator for the perceptual costs of errors: if the tolerance for an error is low (meaning that the acceptability of an utterance is low, due to the error), this error induces high perceptual costs. The results of the perception experiment will be used for the allocation of prosodic phrase boundaries in synthetic speech. Besides, the results will also be used for another part of our research, which is a machine learning experiment to predict whether the PP is noun attached or verb attached (described in Chapter 4).

From the evaluation study for Dutch we also learned that many accentuation errors are made on the sentence final verb phrase. These errors are due to the lack of information about the status of the nominal constituent preceding the sentence final verb. This nominal constituent can be either a condition or an argument to the verb. Linguistic investigations (Gussenhoven, 1984; Baart, 1987; Marsi, 2001) showed that conditions can be left out and are not subcategorized for by the verb. All other constituents are arguments. We argue that the sentence final verb should be accented in sentences with a condition preceding it, whereas the verb should not be accented in sentences with an argument preceding the verb. This means that the appropriateness of accentuation of the verb depends on the status of the nominal constituent preceding the verb.

As for allocation of phrase boundaries, TTS-systems will always make some errors when allocating accents. Therefore, we also want to find out whether an inserted accent or an omitted accent is less problematic for the listener. This way we can try to

[1]Note that the notion 'preferred sentence' is used here for the version that is most appreciated by the listeners. When they have to choose between two incorrect versions, the one that is most appreciated will be referred to as the preferred version.

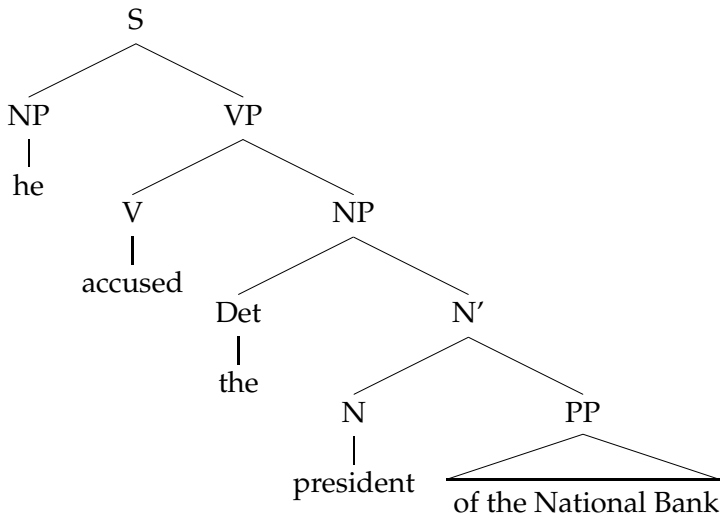
shift the number of accentuation errors in the direction of the one that is less problematic. We perform a perception experiment in which subjects are asked to indicate their preference for either the sentence with or without an accented sentence final verb. The results of this study will also be used for another part of our research, which is a machine learning experiment to predict the status of the nominal constituent preceding the sentence final verb (described in Chapter 5).

In section 3.2, we describe the perception experiment on the allocation of prosodic phrase boundaries at junctures preceding the PP. In section 3.3, we describe the perception experiment on the accentuation of sentence final verbs. In section 3.4, we discuss what the results of these experiments imply for further experiments and for the allocation of prosodic structure in synthetic speech.

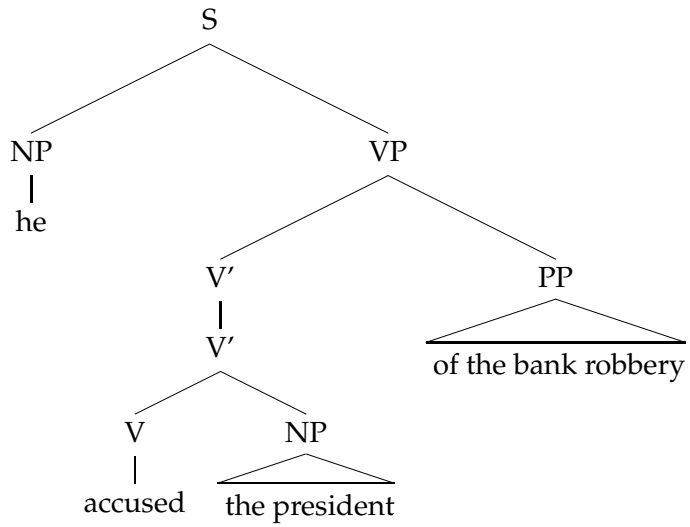
3.2 Prosodic phrasing in case of PP attachment

Although there is no one-to-one mapping between syntax and prosody (e.g. Selkirk, 1984), the syntactic structure of a sentence forms a restriction for phrase boundary allocation. We focus on the allocation of prosodic phrase boundaries at junctures preceding a noun attached or verb attached PP. If the PP is NOUN attached, it is syntactically related to the preceding PP or NP. This means that the attached PP and the preceding PP or NP comprise one constituent. If the PP is VERB attached, it is not syntactically related to the preceding PP or NP, but to the verb in the sentence. Then, the attached PP forms a single constituent. The syntactic tree in example 3.2 is a representation of a phrase with a noun attached PP. Here, a phrase boundary preceding the PP '*of the National Bank*' would be inappropriate. The tree in example 3.3 is a representation of a phrase with a verb attached PP. Here, a phrase boundary preceding the PP '*of the bank robbery*' would be appropriate, although not mandatory.

(3.2)



(3.3)



Previous studies showed that there are further restrictions on phrase boundary allocation. Scharpff and van Heuven (1988) found that listeners either prefer sentences with phrase boundaries at junctures marking syntactic constituents or sentences without boundaries. They also found that listeners are disturbed by phrase boundaries automatically allocated after a fixed number of words and by boundaries at junctures within syntactic constituents. Sanderman (1996) found that the freedom of the speaker may lead to a complete absence of prosodic phrase boundaries, whereas it may not lead to unlimited insertion of boundaries. There is an upper limit of the number of phrase boundaries which is acceptable to the listener. The effects are larger for synthetic speech than for natural speech. We argue that in synthetic speech listeners are helped by correct allocation of some extra (weak) boundaries, whereas in natural speech these boundaries would be thought redundant by the listener.

As we mentioned in section 3.1, we investigate whether listeners are more tolerant towards phrase boundary insertions or omissions. Our first hypothesis is that **listeners prefer the utterance with correct prosodic phrasing over the utterance with incorrect prosodic phrasing**. If a boundary is incorrectly inserted in a sentence with a NOUN attached PP (henceforth referred to as NOUN sentence), we expect that listeners have a preference for the utterance with correct phrasing (i.e. the utterance in which the PP is not preceded by a phrase boundary). If a boundary is incorrectly omitted in a sentence with a VERB attached PP (henceforth referred to as VERB sentence), we expect that listeners also have a preference for the utterance with correct phrasing (i.e. the utterance in which the PP is preceded by a phrase boundary). Considering the findings reported by Sanderman (1996) we expect that an incorrectly inserted phrase boundary is more problematic than an omitted boundary. Our second hypothesis then is that **the preference for correct prosodic phrasing is stronger for NOUN sentences than for VERB sentences**.

Various cues for perceiving a prosodic phrase boundary are known. One or more of the following cues can be used for the realization of a phrase boundary: pause, dec-

lination reset, pre-boundary lengthening and pitch movement. The cues which are used determine the perceived boundary strength. We are interested in the allocation of phrase boundaries within a sentence. Strong phrase boundaries usually mark the end of a sentence, allocation of a strong phrase boundary within a sentence is not desirable. Therefore, we focus on weak and medium phrase boundaries.

When reading a text aloud, speakers realize different boundary strengths, we therefore hypothesize that there is an effect of boundary strength. Since a NOUN attached PP should not be preceded by a phrase boundary, we expect that the preference of the listener for the utterance without a boundary is larger when compared with a medium boundary than with a weak boundary. Hypothesis 3a then is **for NOUN sentences an incorrectly inserted weak boundary is less problematic than an incorrectly inserted medium boundary**. Since a VERB attached PP can be preceded by a phrase boundary (but this is not mandatory), we expect that listeners have a preference for the utterance with a weak boundary over an utterance without a boundary or with a medium boundary. Hypothesis 3b then is **for VERB sentences a weak boundary is preferred over no boundary or a medium boundary**.

3.2.1 Method

Material

As we already mentioned in section 3.1, the correctness of phrase boundary allocation depends on the status of the succeeding prepositional phrase. Therefore, we selected two types of sentences: 10 sentences with a NOUN attached PP and 10 with a VERB attached PP. Sentence 3.4(a) is an example of a NOUN sentence: the PP '*van zijn fiets*' is noun attached. This means that a phrase boundary allocated at the juncture preceding the PP would be inappropriate. Sentence 3.4(b) is an example of a VERB sentence: the PP '*tot grote rust*' is verb attached. Thus, a phrase boundary allocated at the juncture preceding the PP is appropriate.

(3.4) (a) *De buurman beweerde dat zijn zoontje de remmen [ ] van zijn*
The neighbor claimed that his little son the breaks [ ] of his
fiets had gemaakt.
bike had repaired.
"The neighbor claimed that his little son had repaired the breaks of his bike."

(b) *In het nieuwsblad staat dat de aangekondigde zonsverduistering [ ]*
In the newspaper it says that the announced solar eclipse [ ]
tot grote rust heeft geleid.
to great quietude has led.
"It says in the newspaper that the announced solar eclipse has led to great quietude."

Figure 3.1 shows the three different realizations of prosodic phrase boundaries we used in the perception experiment. Type 0 is the schematic representation of no boundary preceding the attached PP (represented by the final peak). Type 1 represents a weak boundary realized by a continued high pitch and pre-pausal lengthening. Type 2 represents a medium boundary realized by a continuation rise followed by a 350ms pause.

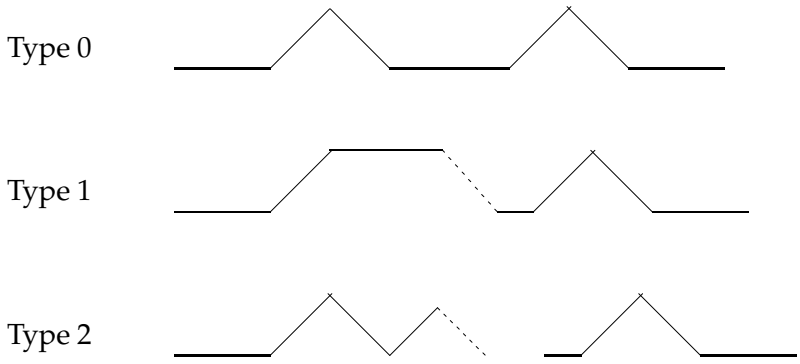


Figure 3.1: *Schematic representation of the three realizations of prosodic phrase boundaries. Type 0 represents no boundary, Type 1 represents a weak boundary and Type 2 represents a medium boundary. The dotted lines mark the non-audible (imaginary) continuation of the pitch contour.*

All 20 sentences (given in Appendix B) were realized with all three phrase boundary types at the juncture preceding the PP. This resulted in $20 * 3 = 60$ utterances. For the synthesis we used the female voice of Calipso Text-to-Speech synthesis.

Experimental Design

The utterances were presented pairwise to 20 subjects. Each pair consisted of two realizations of the same sentence, with two of the three phrase boundary types (Type 0, Type 1 or Type 2). All three possible combinations were presented in both orders. This resulted in $20 * 3 * 2 = 120$ pairs of utterances.

We distributed the utterance pairs over two subject groups according to a Latin square design. This means that group I was presented with the 20 sentence pairs, with all three possible combinations of two boundary types, but in only one order (utterance X - utterance Y), while group II was presented with the same 20 sentence pairs, with all three possible combinations of two boundary types, in the opposite order (utterance Y - utterance X). This resulted in 60 utterance pairs per subject. The utterance pairs were presented in random order.

The experiment was conducted in a sound treated room. The stimuli were presented over head phones, while at the same time the text of the sentence was displayed on a screen. The silence between the two utterances in a pair was 400ms. Each utterance pair was presented twice. After the second presentation, subjects were asked to indicate on a 7-point scale which utterance of the pair they preferred and to what extent.

Subjects had to indicate their judgement by clicking with a mouse on a button on the screen. These buttons indicated the preferences shown in Table 3.1.

Table 3.1: *The 7-point scale for judging utterance pairs.*

-3	strong preference for utterance 1
-2	medium preference for utterance 1
-1	slight preference for utterance 1
0	no preference
+1	slight preference for utterance 2
+2	medium preference for utterance 2
+3	strong preference for utterance 2

Prior to the actual experiment there was a training phase where the subjects could get acquainted with the procedure. In this training phase subjects were presented with 12 utterance pairs that were not part of the actual experiment. The total duration of the experiment was about 30 minutes.

All subjects were native speakers of Dutch and none of them reported hearing problems. They were all (PhD) students in the age of 21 through 31 and they were not familiar with the research which this perception experiment is part of.

Statistical Design

We submitted our data to an analysis of variance for paired comparisons (Scheffé, 1952). This method is developed for experiments in which preferences are expressed on a scale of 7 points or more. In the analysis the hypothesis of subtractivity is statistically tested. This hypothesis states that there exist parameters α_1 and α_2 characterizing both versions of a sentence in an utterance pair, such that the average preference for item 1 over item 2 is defined as $\alpha_1 - \alpha_2$. The α values, computed from the preference judgements, can be considered as the one-dimensional perceptual scaling for both realizations of the sentence. From the variances of the scores we compute a yardstick. Differences between α values are only significant ($p < 0.05$) if they are larger than the yardstick. The yardstick Y is computed according to equation 3.5:

$$Y = q_{0.95}\sqrt{\hat{\sigma}^2/(2rm)} \tag{3.5}$$

where $\hat{\sigma}^2$ is the estimate of the variance, $2rm$ are the degrees of freedom (the number of subjects times the number of variants, which is 3 boundary types) and $q_{0.95}$ is the critical value of the Studentized Range. The value of the Studentized Range is taken from Ferguson and Takane (1989, Table L, page 570).

3.2.2 Results and discussion

The preference scores for prosodic phrasing (averaged over all subjects and both utterance orders) are given in Table 3.2. These results show that the preference scores for sentences with a NOUN attached PP are larger than for sentences with a VERB attached PP. When we look at the preference scores for the comparison Type 0 - Type 1 and Type 0 - Type 2, we see that the preference for Type 0 is larger when compared with Type 2 than with Type 1, for both NOUN and VERB. The scores indicate that for both sentence types, there is a preference for the utterance without a phrase boundary preceding the PP. This preference is smaller for VERB than for NOUN. The scores for the comparison Type 1 - Type 2 show that for both NOUN and VERB there is a preference for a Type 1 boundary.

Table 3.2: *Mean preference scores for prosodic phrasing resulting from the comparison of utterances realized with the three boundary types. The minus sign indicates that there is a preference for the first mentioned boundary type in the comparison. Type 0: no boundary, Type 1: weak boundary and Type 2: medium boundary.*

	sentence type	**comparison**	**preference score**
P	NOUN	Type 0 - Type 1	-0.46
Q	NOUN	Type 0 - Type 2	-1.43
R	NOUN	Type 1 - Type 2	-1.13
S	VERB	Type 0 - Type 1	-0.12
T	VERB	Type 0 - Type 2	-0.40
U	VERB	Type 1 - Type 2	-0.29

Figure 3.2 shows that for NOUN attachment there is a clear preference for the weaker boundary (Type 0 or Type 1), indicated by the negative preference scores, whereas for VERB attachment there is no clear preference for either the weaker or the stronger boundary: the preference scores are almost equal to 0 (which means no preference).

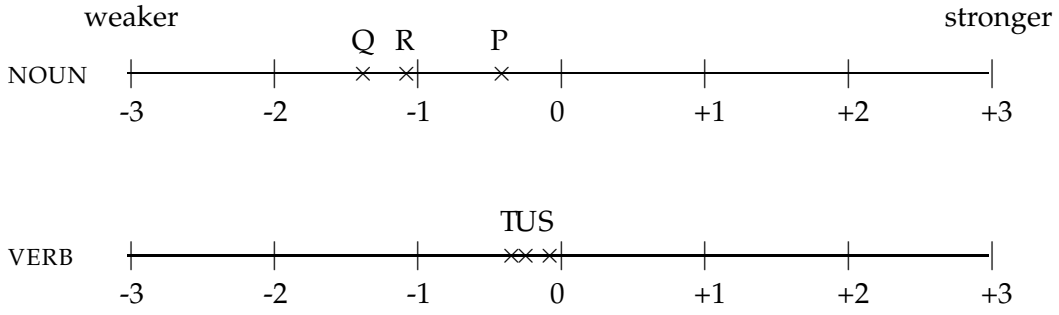


Figure 3.2: *Schematic representation of the preference scores for the comparisons given in Table 3.2, for both* NOUN *and* VERB *attachment. The score -3 represents a maximal preference for the weaker boundary, whereas the score +3 represents a maximal preference for the stronger boundary.*

Hypothesis 1: Listeners have a preference for the utterance with correct prosodic phrasing.

Correct phrasing means for NOUN sentences that there is no phrase boundary allocated at the juncture preceding the attached PP, whereas for VERB sentences it means that there is a boundary preceding the PP, however this boundary is not mandatory.

For testing this first hypothesis we used the judgements of all utterance pairs consisting of an utterance with a boundary and an utterance without a boundary. We abstracted from boundary strength. We performed separate analyses for the 10 NOUN sentences (comparison P and Q in Table 3.2) and the 10 VERB sentences (comparison S and T in Table 3.2). Analysis of variance for paired comparisons (as introduced in section 3.2.1) showed that for NOUN sentences the difference between the two versions ($\alpha_1 - \alpha_2$) is 0.942, with a yardstick (Y) of 0.206. So $\alpha_1 - \alpha_2 > Y$, which means that the difference between α_1 and α_2 is significant ($p < 0.05$) for NOUN sentences. For VERB sentences $\alpha_1 - \alpha_2$ is 0.257, with a yardstick of 0.205. This means that the difference between the two versions is also significant for VERB sentences. However, as opposed to our expectations for VERB sentences, listeners do not prefer the sentence with a prosodic phrase boundary, but the sentence without a phrase boundary.

The results for NOUN sentences prove that our hypothesis was correct, however for VERB sentences this is not true. For sentences with a verb attached PP listeners preferred the utterance without a boundary, while we expected a preference for the one with a boundary preceding the verb attached PP. This preference might be induced by the relatively strong character of the medium boundary (Type 2), which is allocated after the first part of the utterance (indicated by // in example 3.6).

(3.6) In het nieuwsblad staat // dat de aangekondigde zonsverduistering [ ] tot grote rust heeft geleid.

The remainder of the utterance is not that long (only 9 words), so listeners do not need another boundary. However, they do not really disapprove of a weak boundary (see comparison S in Figure 3.2). If we only compared utterances for VERB attachment without a boundary with ones with a weak boundary, we might find no preference at all for one of the two versions. In that case we have no evidence that we should not allocate some extra weak boundaries in synthetic speech, which would be helpful to the listener.

Hypothesis 2: The preference for correct prosodic phrasing is stronger for NOUN sentences than for VERB sentences.

For testing this second hypothesis we again used the judgements of all utterance pairs consisting of an utterance with a boundary and one without a boundary. As for hypothesis 1, we abstracted from boundary strength. We compared the difference be-

tween the utterances with correct and incorrect phrasing for NOUN sentences (comparison P and Q in Table 3.2), with this difference for VERB sentences (comparison S and T in Table 3.2). To investigate whether the difference is larger for NOUN sentences than for VERB sentences we computed $\alpha_1 - \alpha_2$ per sentence. For NOUN sentences $\alpha_1 - \alpha_2 > Y$ was true for 6 out of 10 sentences, whereas for VERB sentences it was true for only 1 out of 10 sentences. In the latter case listeners preferred the utterance without a prosodic phrase boundary preceding the PP.

Although the difference between α_1 and α_2 is not significant for all sentences, we performed a t-test for independent samples, which showed that this difference is significantly larger for NOUN sentences than for VERB sentences ($t = 2.96, p < 0.05, df = 18$).

The fact that for only 1 VERB sentence a significant preference was found for the utterance without a boundary preceding the PP, proves that listeners are not really disturbed by allocation of a phrase boundary preceding the PP. Furthermore, the results show that the costs of incorrect prosodic phrasing are larger for NOUN than for VERB sentences. For NOUN sentences incorrect phrasing means incorrect allocation of boundaries. This implies that it is better to allocate too few boundaries than too many.

Hypothesis 3a: For NOUN sentences an incorrectly inserted weak boundary is less problematic than an incorrectly inserted medium boundary.

For testing this hypothesis we used the judgements for NOUN only. We compared utterances without a boundary (Type 0) with utterances with a weak boundary (Type 1) or with a medium boundary (Type 2). We computed the differences between correct and incorrect phrasing for comparison P and Q in Table 3.2 per sentence summed over all subjects. For comparison P, $\alpha_1 - \alpha_2 > Y$ was true for 3 sentences, and for comparison Q this was true for 7 sentences.

Although again the difference between α_1 and α_2 is not significant for all sentences, we performed a t-test for independent samples, which showed that the preference for correct phrasing is significantly larger ($t = -4.39, p < 0.05, df = 18$) when the utterance without a boundary is compared with the utterance with a medium boundary, than when it is compared with the utterance with a weak phrase boundary.

These results imply that for NOUN sentences listeners are more tolerant towards incorrect insertion of a weak boundary than of a medium boundary. This means that our phrasing algorithm should be rather restrictive with the allocation of medium boundaries, whereas it can more freely allocate weak boundaries at junctures preceding an attached PP.

Hypothesis 3b: For VERB sentences a weak boundary is preferred over no boundary or a medium boundary.

For this hypothesis we used the judgements for VERB only. We computed the differences between Type 0 and Type 1 and between Type 1 and Type 2 boundaries. We computed $\alpha_1 - \alpha_2$ for comparison S and U in Table 3.2 per sentence summed over all subjects. For comparison S, $\alpha_1 - \alpha_2 > Y$ was true for only 1 sentence out of 10, and for comparison U this was true for none of the 10 sentences. A t-test for independent samples showed that the preference for a weak boundary (Type 1) was not significantly larger ($t = -0.387, p = 0.70, df = 18$) when the utterance with a weak boundary was compared with either one without a boundary or one with a medium boundary.

These results imply that for VERB sentences there is no preference for one of the boundary types. Listeners are not disturbed by the allocation of a boundary nor by no boundary. This means that our phrasing algorithm could freely allocate a boundary at the juncture preceding a verb attached PP, although a boundary is not mandatory.

The overview in Table 3.3 shows for which comparisons we found significant differences.

Table 3.3: *Overview of results of the comparisons for prosodic phrasing per hypothesis.*

hypothesis	comparison			significant
1	NOUN correct	vs.	NOUN incorrect	Y
1	VERB correct	vs.	VERB incorrect	Y
2	NOUN cor. - NOUN incor.	vs.	NOUN cor. - NOUN incor.	Y
3a	NOUN Type 0 - Type 1	vs.	NOUN Type 0 - Type 2	Y
3b	VERB Type 1 - Type 0	vs.	VERB Type 1 - Type 2	N

Summarizing, from the results we learned that for NOUN sentences the perceptual costs of incorrect prosodic phrasing are larger than for VERB sentences. For VERB sentences there even is a slight preference for the utterance with incorrect phrasing (i.e the utterance without a boundary). For NOUN sentences the difference between correct and incorrect phrasing is larger in case of incorrect insertion of a medium boundary than in case of incorrect insertion of a weak boundary. For VERB sentences there is no preference for the utterance with a weak boundary when compared with utterances without a boundary or with a medium boundary.

Since we argue that in synthetic speech more phrase boundaries should be allocated than in natural speech, we still think that allocation of weak boundaries preceding VERB attached PP's may help the listener. However, we should be rather stringent regarding incorrect boundary insertions in sentences with a NOUN attached PP. Our phrasing algorithm should only allocate weak boundaries at junctures preceding attached PP's, so that the least acceptable error (i.e. incorrect insertion of a medium boundary) would never occur.

As mentioned in section 3.1, the outcome of this perception experiment will be used for machine learning experiments for predicting the PP attachment. Considering the above mentioned reasons we should optimize on the recall for NOUN attachment. A high recall for NOUN attachment means that we only miss few cases of NOUN attachment (this could also be described as 'bias' towards NOUN attachment), which means that we have only few incorrect boundary insertions.

3.3 Accentuation of sentence final verbs

Previous research has shown that accentuation expresses the focus structure of a sentence (Baart, 1987; Gussenhoven, 1992; Birch and Clifton, 1995): [-focus] domains should remain unaccented, whereas [+focus] domains should get an accent. The problem we try to solve here is accentuation of the sentence final verb phrase. For accentuation of the sentence final verb we apply the Sentence Accent Assignment Rule (SAAR) (Gussenhoven, 1982, 1984). SAAR distinguishes three constituents: argument, condition and predicate and it is applied in two steps. First, focus domains are allocated and second, the exact location of the accent in that specific domain is decided.

We will concentrate on those cases where the predicate is preceded by either an argument ARG (as in example 3.7a) or a condition COND (as in example 3.7b).

(3.7) (a) *Hij heeft het hele boek gelezen.*
He has the entire book read.
"He has read the entire book."

(b) *Hij heeft de hele nacht gelezen.*
He has the entire night read.
"He has been reading the entire night."

If the predicate is preceded by an argument, they comprise one focus domain. This is shown in example 3.8a, where [ ] indicates the focus domain. To every focus domain at least one accent will be assigned. If a focus domain contains an argument and a predicate, the accent (indicated by *) will be on the argument, because it is strong in relation to the functor predicate. The scope of the accent concerns the whole predicative expression. If the nominal constituent is a condition (as in 3.8b) it is a separate focus domain. In this case the scope of the accent does not include the predicate, which means that the predicate will also receive an accent.

(3.8) (a) Hij heeft [het hele *boek gelezen].

(b) Hij heeft [de hele *nacht] [*gelezen].

Accentuation of a certain word is a cue for the listener that this specific word is informative. If a word is incorrectly accented or incorrectly not accented, the listener will

need more time and effort to understand the utterance. Therefore, our first hypothesis is that **listeners prefer the sentence with correct accentuation of the sentence final verb over the sentence with incorrect accentuation**. For ARG sentences correct accentuation means an unaccented sentence final verb, and for COND it means an accented sentence final verb.

In Dutch, the strongest cue to accent perception is (proper) movement of pitch. For Dutch two generally accepted pitch movements which are accent lending are the accent lending rise and accent lending fall. These pitch movements often succeed each other, resulting in the so called pointed hat and the flat hat ('t Hart and Cohen, 1973; 't Hart and Collier, 1975). Both the pointed hat and the flat hat consist of an accent lending rise followed by an accent lending fall. For the pointed hat both rise and fall are realized on the same syllable, whereas for the flat hat the accent lending rise is realized on the first accent receiving syllable and the fall is realized on the second accent receiving syllable. In case of a flat hat contour, the pitch will remain high between the rise and fall.

A flat hat contour is often used when the two accented words are semantically (Kruyt, 1985; Mallant, 1992) or syntactically related (Baart, 1989; Birch and Clifton, 1995), for instance when the verb subcategorizes for a certain argument. If a semantic or syntactic connection exists between two words (as in example 3.9a), but the two accents are both realized by two pointed hats, the listener gets no prosodic cue about the relation of the words. If no such relation exists between two words (as in example 3.9b), but the two accents are realized by a flat hat contour, the listener gets an incorrect prosodic cue that the two words are semantically or syntactically related.

(3.9) (a) *Zijn zoontje heeft de remmen van zijn fiets gemaakt.*
His son has the breaks of his bike repaired.
"His son has repaired the breaks of his bike."

(b) *Haar dochter heeft de pop in de wieg gelegd.*
Her daughter has the doll in the crib put.
"Her daughter has put the doll in the crib."

Listeners have a different expectation when being presented with a pointed hat than with a flat hat. According to the rules for Dutch intonational structure the fall of a flat hat can only occur on the final accent (i.e. not succeeded by other accents). Therefore, after perceiving the fall of a flat hat contour, listeners know that the remainder of the sentence will contain no accents. The fall of the pointed hat does not lead to such expectations as it can be succeeded by another accent lending pitch movement. Besides, the high pitch plateau, between the rise and fall of a flat hat contour, is an indication for the listener that a final accent (realized as an accent lending fall) will follow. Pointed hats do not provide the listener with such cues about the accentual structure of the remainder of the sentence. Because of these different expectations of the listener when perceiving a pointed hat or a flat hat, our second hypothesis is that **there is an effect of**

pitch contour. Preference scores for the sentence with or without an accented sentence final verb will differ when the accents are realized as a flat hat than when realized as pointed hats.

If the sentence final verb is incorrectly unaccented in a COND sentence through using the pitch contour with two pointed hats, there is no bias for an ARG sentence until the realization of the verb. This means that the listener did not get correct information about the syntactic structure of the utterance, but he has neither been exposed to incorrect information up to the final part of the utterance. Therefore, the listener will only be marginally disturbed by this incorrectly deaccenting of the verb. The same is true for incorrect accentuation of the sentence final verb in ARG sentences through using the pitch contour with two pointed hats. Therefore, we hypothesize that **for pitch contours consisting of all pointed hats, an inserted accent is as bad as an omitted accent** (hypothesis 3a).

If the sentence final verb is incorrectly unaccented in a COND sentence where the accents are realized as a flat hat, this flat hat is allocated on the two words which precede the sentence final verb and that are marked for accent. These two words are not semantically related while the pitch contour would suggest such a relation. Thus, this would incorrectly induce a bias for an ARG sentence. This bias will cause a more difficult understanding of the sentence, so that the listener will need more time and effort to recover the correct meaning of the sentence by back-tracking the sentence. If the sentence final verb is incorrectly accented in an ARG sentence where the last two accents are realized as a flat hat, the flat hat is allocated on the argument and the verb. This means that there is no information available about the status of the nominal constituent (whether it is an argument or a condition) that precedes the verb, until this constituent is realized. The listener will not need much back-tracking for a correct understanding of the utterance (because he has no information about the status of the nominal constituent, as opposed to incorrect information). Since accent omissions provide the listener with incorrect information whereas accent insertions deprive the listener of information, we hypothesize that **for flat hat contours an omitted accent is worse than an inserted accent** (hypothesis 3b).

3.3.1 Method

Material

As we already mentioned in section 3.1, the accentuation of the sentence final verb is influenced by the status of the preceding constituent. We selected two types of sentences: 10 sentences with an argument preceding the sentence final verb (ARG) and 10 sentences with a condition preceding the verb (COND). All sentences contain two accent marked words preceding the sentence final verb. Sentence 3.10(a) is an example of an ARG sentence: the constituent '*een toespraak*' is an argument to the verb. The

sentence final verb '*gehouden*' should not be accented. Sentence 3.10(b) is an example of a COND sentence: the constituent '*met spoed*' is a condition. The sentence final verb '*geopereerd*' should be accented.

(3.10) (a) *De directeur heeft een toespraak gehouden.*
The director has a speech given.
"The director has given a speech."

(b) *Het kind is met spoed geopereerd.*
The child has urgently had an operation.
"The child has had an urgent operation."

Figure 3.3 shows schematic representations of the four pitch contours that we used for the perception experiment on accentuation of the sentence final verb. Type A and C are patterns with so called pointed hats. Type B and D are patterns with a so called flat hat. Type A and B represent sentences with an accent on the sentence final verb. Type C and D represent sentences without an accent on the sentence final verb.

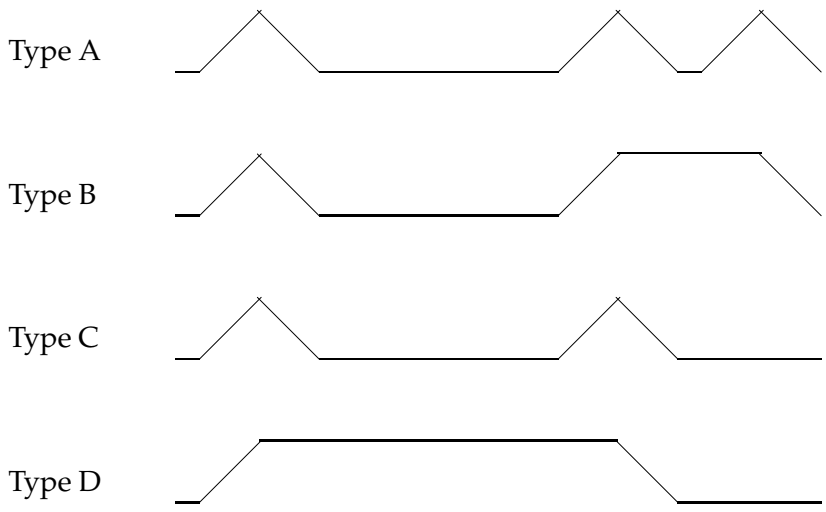


Figure 3.3: *The four different pitch contours used for accentuation. Type A is the schematic representation of a phrase with 2 accented words and an accented sentence final verb, with all accents realized as pointed hats. Type B represents a phrase with 2 accented words and an accented sentence final verb, with the first accent realized as a pointed hat and the last 2 accents realized as a flat hat. Type C represents a phrase with 2 accents and an unaccented sentence final verb, with both accents realized as pointed hats. Type D represents a phrase with 2 accents and an unaccented sentence final verb, with both accents realized as a flat hat.*

All 20 sentences (given in Appendix B) are realized with and without an accent on the sentence final verb (i.e. with correct and incorrect accentuation). The sentence final verb is the third word in the sentence that can be marked for accent. In addition, all sentences are realized with two different pitch contours: (i) a pitch contour with all accents realized as pointed hats (Type A and C), and (ii) a pitch contour with the

last two accents realized as a flat hat (Type B and D). This resulted in $20 * 2 * 2 = 80$ utterances. For the synthesis we again used the female voice of Calipso Text-to-Speech synthesis.

Experimental Design

The utterances were presented pairwise to 20 subjects. Each pair consisted of a sentence realized with an accent on the sentence final verb and the same sentence realized without an accent on the verb, or in the opposite order. The paired two utterances were both realized with either all pointed hats or a flat hat. This resulted in 80 pairs of utterances (20 sentences $*$ 2 orders $*$ 2 accent realizations).

The experimental design of this perception experiment resembles the design of the perception experiment on prosodic phrasing (as described in section 3.2.1). The total duration of this experiment on accentuation differs from that of the experiment on prosodic phrasing, lasting 15 minutes instead of 30 minutes and in the training phase subjects were presented with 8 (instead of 12) utterance pairs which were not part of the actual experiment.

Statistical Design

The results of this perception experiment on accentuation will be analyzed with the same statistical approach as we described for the experiment on prosodic phrasing (see section 3.2.1).

3.3.2 Results and discussion

The preference scores (averaged over all subjects and both utterance orders) for accentuation of the sentence final verb are given in Table 3.4. These scores indicate that there is a preference for the utterance with correct accentuation. For ARG sentences subjects prefer utterances realized with Type C and Type D pitch contours. For COND sentences subjects prefer utterances with Type A and Type B pitch contours. If we compare the scores for 'pointed hats' with the scores for 'flat hat', we see that the scores for pointed hats are higher than for flat hat. Moreover, the absolute preference scores for ARG and COND sentences for 'pointed hats' are almost equal (1.44 vs. 1.65), whereas the absolute score for ARG for 'flat hat' is lower than the absolute score for COND for 'flat hat' (1.02 vs. 1.59).

Figure 3.4 shows that for ARG sentences there is a clear preference for a pitch contour in which the sentence final verb is unaccented (Type C or Type D), whereas for COND sentences there is a preference for a pitch contour in which the sentence final verb is accented (Type A or Type B).

Table 3.4: *Mean preference scores for accentuation resulting from the comparison of utterances realized with the four pitch contours for accentuation. The minus sign indicates a preference for the first mentioned accentuation type in the comparison, whereas the plus sign indicates a preference for the second mentioned accentuation type in the comparison.*

	sentence type	comparison	pitch movement	preference score
J	ARG	Type A - Type C	pointed hats	+1.44
K	ARG	Type B - Type D	flat hat	+1.02
L	COND	Type A - Type C	pointed hats	-1.65
M	COND	Type B - Type D	flat hat	-1.59

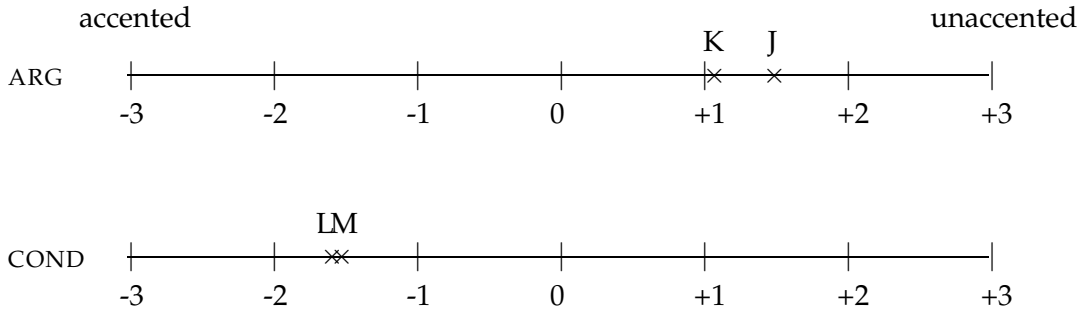


Figure 3.4: *Schematic representation of the preference scores for the comparisons given in Table 3.4, for both* ARG *and* COND *sentences. The score -3 represents a maximum preference for an accented sentence final verb, whereas the score +3 represents a maximum preference for an unaccented verb.*

Hypothesis 1: Listeners have a preference for the utterance with correct accentuation.

Correct accentuation means for ARG sentences that the sentence final verb is unaccented, whereas for COND sentences it means that the verb is accented.

For testing this first hypothesis we used the judgements of all utterance pairs. This means that we abstracted from the division in 'pointed hats' and 'flat hat'. We performed separate analyses for the 10 sentences in which the sentence final verb should be unaccented (ARG) and for the 10 sentences in which the verb should be accented (COND). Judgements were summed over all subjects. Analysis of variance for paired comparisons showed that for ARG sentences the difference between the two versions ($\alpha_1 - \alpha_2$) is 1.123, with a yardstick (Y) of 0.193. So, $\alpha_1 - \alpha_2 > Y$ is true, indicating that the difference between α_1 and α_2 is significant ($p < 0.05$) for ARG sentences. For COND sentences $\alpha_1 - \alpha_2$ is 1.615, with $Y = 0.178$, indicating that that the difference between the two versions is also significant for COND sentences.

The results prove that for both ARG and COND sentences our hypothesis was correct: listeners prefer the utterance with correct accentuation.

Hypothesis 2: There is an effect of pitch contour.

We hypothesize that preference scores for the sentence with correct accentuation are different when the accents are realized as a flat hat or realized as pointed hats.

For testing this second hypothesis, we again used all judgements. However, we performed separate analyses for 'pointed hats' and 'flat hat' and for ARG and COND sentences. We want to compare the difference between the utterances with correct and incorrect accentuation ($\alpha_1 - \alpha_2$) for pointed hats, with this difference for flat hats. To investigate whether the difference is larger for pointed hats than for flat hats we computed $\alpha_1 - \alpha_2$ per sentence, summed over all subjects. For ARG sentences realized with all pointed hats, $\alpha_1 - \alpha_2 > Y$ was true for 7 sentences out of 10, and for ARG sentences realized with a flat hat it was true for 6 sentences. For COND sentences realized with all pointed hats, $\alpha_1 - \alpha_2 > Y$ was true for 9 sentences out of 10 and for COND sentences realized with a flat hat it was also true for 9 sentences.

Although the difference between α_1 and α_2 is not significant for all sentences, we performed t-tests for independent samples. These tests showed that the difference between 'pointed hats' and 'flat hat' for ARG sentences is almost significant ($t = 2.01, p = 0.06, df = 18$), whereas this difference for COND sentences is clearly not significant ($t = -0.46, p = 0.66, df = 18$). These results show that the effect of pitch contour is not significant.

We expected a difference in preference for the correct sentence when we compared utterances realized with pointed hats and flat hats. However, the results showed that these differences are not significant for both ARG and COND sentences. For ARG there is still some tendency that our hypothesis is true, because the difference between 'pointed hats' and 'flat hats' is almost significant.

Hypothesis 3a: For pointed hat contours the perceptual costs of the two types of accentuation errors are equal.

For testing this hypothesis we used the listeners' judgements for 'pointed hats' only. We computed the difference between correct and incorrect accentuation for ARG and COND for each sentence separately, summed over all subjects.

Although the difference between the two versions is not significant for all sentences, we again performed a t-test for independent samples. This test showed that the difference between ARG and COND sentences is not significant for 'pointed hats' ($t = -1.14, p = 0.27, df = 18$).

These results indicate that for utterances realized with a 'pointed hat' contour an incorrect accent insertion is as bad as an accent omission. Thus, there is no difference in the perceptual costs of both types of errors. For our accentuation algorithm this means that it should find an optimum for allocating an accent or not.

Hypothesis 3b: For flat hat contours the perceptual costs of omitted accents are larger than the costs of inserted accents.

For testing this last hypothesis we used the listeners' judgements for 'flat hat' only. We computed the difference between correct and incorrect accentuation for ARG and COND for each sentence separately, summed over all subjects.

We again performed a t-test for independent samples. This test showed that the difference between ARG and COND sentences is significant for 'flat hat' ($t = -2.63, p < 0.05, df = 18$).

These results imply that for utterances realized with a 'flat hat' contour an incorrect accent omission is worse than an incorrect accent insertion. Thus, the perceptual costs of an omission are larger than those of an insertion. For our accentuation algorithm this means that is should rather allocate too many accents than too few.

The overview in Table 3.5 shows for which comparisons we found significant differences.

Table 3.5: *Overview of results of the comparisons for accentuation per hypothesis.*

hypothesis	**comparison**			**significant**
1	ARG correct	vs.	ARG incorrect	Y
1	COND correct	vs.	COND incorrect	Y
2	ARG pointed	vs.	ARG flat	N
2	COND pointed	vs.	COND flat	N
3a	ARG pointed	vs.	COND pointed	N
3b	ARG flat	vs.	COND flat	Y

Summarizing, from the results we learned that for both ARG and COND sentences there is a preference for the utterance with correct accentuation of the sentence final verb. The differences between preference scores for 'pointed hats' and 'flat hats' are not significant for either ARG or COND sentences, although for ARG there is a tendency that the difference between correct and incorrect accentuation is larger when accents are realized as pointed hats than when realized as a flat hat. If accents are realized as a flat hat, the difference between correct and incorrect accentuation in case of accent omission is significantly larger than in case of accent insertion. If accents are realized as pointed hats, the difference between correct and incorrect in case of accent insertion is not significantly different from this in case of accent omission.

From the outcome of this perception experiment we learn that proper accentuation is helpful to the listener when perceiving synthetic speech. We will use this outcome for machine learning experiments for predicting the status (argument or condition) of the nominal constituent preceding the sentence final verb. There is no univocal result when considering the acceptability of insertions and omissions of accents. Although

for flat hats there is an indication that listeners are more tolerant towards accent insertions than towards accent omissions. This is not true if all accents are realized as pointed hats.

Our speech synthesis system first decides which words will be marked for accent and only at the second stage, the accentuation pattern will be assigned. If our speech synthesis system were to work in the opposite order, that is if it first decided what type of accentuation pattern was to be allocated and next decided which words should be marked for accent, then we should apply different strategies for pointed hats and for flat hats with regard to optimization in machine learning experiments. For pointed hats we still should optimize on the accuracy (the overall score on ARG and COND). For flat hats we should optimize on the recall for COND sentences, because for this pitch contour accent omissions are less acceptable than accent insertions. A high recall for COND sentences means that we only miss few cases of COND (this could also be described as a 'bias' towards COND), which means that we have only few incorrect accent omissions. An easier solution would be that the accentuation algorithm only assigns pitch contours with all pointed hats, but for the sake of variation (which makes synthetic speech more vivid) we would prefer assignment of both pointed hats and flat hats.

We conclude that correct prediction of ARG and COND is equally important. For machine learning experiments this means that we should obtain a high accuracy, when comparing the predictions to the reference transcription (described in Chapter 2); we should try to have as few incorrect ARG and COND predictions as possible. This will result in as few accent insertions and omissions as possible.

3.4 Conclusion

From the perception experiment on prosodic phrasing we learned that listeners often prefer the utterance without a phrase boundary preceding an attached PP. This is not only true for sentences with a noun attached PP, but also for sentences with a verb attached PP. For VERB sentences we hypothesized that there would be a preference for the utterance with a boundary. However, the results prove that there was no significant preference for a weak boundary over no boundary or a medium boundary. For NOUN sentences incorrect insertion of a weak boundary turned out to be less problematic for the listener than insertion of a medium boundary. We argued that in synthetic speech some more weak boundaries should be allocated than in natural speech. This way we give the listener more time to process the speech signal and it makes the signal more composed.

Considering the results of the experiment, we will only allocate weak boundaries at junctures preceding attached PP's, so that if this allocation is incorrect, it is not that problematic for the listener. Besides, we will optimize on noun attachment in the ma-

chine learning experiments for predicting the attachment, so that we make as few incorrect NOUN predictions as possible, which results in as few incorrect weak boundary insertions as possible.

From the perception experiment on accentuation of sentence final verbs we learned that listeners prefer the utterance with correct accentuation. This means that the verb should not be accented if it is preceded by an argument, whereas it should be accented if it is preceded by a condition. In general there is no difference between the preference for correct accentuation of the verb when realized with a 'pointed hat' contour and the preference for correct accentuation of the verb when realized with a 'flat hat' contour. The results also show that if a sentence is realized with a pointed hat contour, incorrect accent insertions are as bad as incorrect omissions. If a sentence is realized with a flat hat contour, incorrect accent omissions are worse than incorrect insertions.

For the machine learning experiments on accentuation this means that we should make as few prediction errors as possible for both ARG and COND sentences. So, we should obtain a high accuracy, which means that there are only few incorrect ARG and COND predictions. Eventually, this results in few incorrect accent insertions and omissions.

Predicting PP-attachment 4

*In this chapter we explore how PP attachment ambiguities can be resolved to improve prosodic phrasing in synthetic speech. From a treebank of spoken Dutch we select instances of the attachment of prepositional phrases to either a noun or verb in the sentence. We train two machine learning algorithms (*MBL *and* RIPPER*) on making the distinction between noun and verb attachment on the basis of lexical information and a co-occurrence strength feature derived from the a very large database. The learned models are tested on the Spoken Dutch Corpus data by means of cross-validation experiments, and on held-out newspaper and e-mail data. The results indicate that the learned models have a reasonably stable performance on different kinds of data. Comparison with the reference transcription shows that having correct PP attachment information available improves the performance on prosodic phrase boundary allocation.*

[0]This chapter is based on van Herwijnen et al. (2003)

4.1 Introduction

One of the factors determining the acceptability of synthetic speech is the appropriate placement of phrase boundaries, realized typically and most audibly by pauses (Sanderman, 1996). Incorrect prosodic phrasing may impede the listener in the correct understanding of the spoken utterance (Sanderman and Collier, 1997). As we described in Chapter 2 a major factor that causes difficulties in appropriate phrase boundary placement is the lack of reliable information about syntactic structure. As discussed in Chapter 1, even if there is no one-to-one mapping between syntax and prosody, the placement of prosodic phrase boundaries is nevertheless to a large extent dependent on syntactic information (Selkirk, 1984; Bear and Price, 1990; van Herwijnen and Terken, 2001b). To cope with the lack of elaborate syntactic information several shallower strategies have been applied to allocate phrase boundaries. One strategy is to allocate phrase boundaries on the basis of punctuation only. In general, however, this results in too few phrase boundaries and some incorrect ones.

As we showed in Chapter 2, a clear example of information about syntactic structure being useful for the allocation of phrase boundaries is the attachment of prepositional phrases (PPs). When a PP is attached to the preceding NP or PP (henceforth referred to as noun attachment), as in example 4.1a, a phrase boundary at the juncture between *pizza* and *with* (indicated by []) is usually considered inappropriate. However, when a PP is attached to the verb in the clause (verb attachment), as in example 4.1b, an intervening phrase boundary between the PP and its preceding NP or PP (between *pizza* and *with*) is optional, and when implemented prosodically, usually judged appropriate (Marsi et al., 1997).

(4.1) (a) *Hij eet pizza [] met ansjovis.*
He eats pizza [] with anchovies.

(b) *Hij eet pizza [] met een vork.*
He eats pizza [] with a fork.

Deciding about noun versus verb attachment of PP's is a notoriously hard task in parsing, since it is understood to involve knowing lexical preferences, verb subcategorization, fixed phrases, but also semantic and pragmatic 'world' knowledge. Typical current parsers (e.g. statistical parsers such as developed by Collins (1996); Ratnaparkhi (1997); Charniak (2000)) interleave PP attachment with all its other disambiguation tasks. However, because of its interesting complexity, a line of work has concentrated on studying the task in isolation (Hindle and Rooth, 1993; Ratnaparkhi et al., 1994; Brill and Resnik, 1994; Collins and Brooks, 1995; Franz, 1996; Zavrel et al., 1997). The study described in this chapter can be seen as following these lines of isolated studies, pursuing the same process for Dutch.

We assume that at least two sources of information should be used as features in training data: (i) lexical features (e.g. head words), and (ii) word co-occurrence strength values (the probability that two words occur together, within some defined vicinity). Lexical features may be informative when certain individual words frequently, or exclusively, occur with either noun or verb attachment. This may hold for prepositions, but also for heads of the involved phrases, as well as for combinations of these words. We will illustrate that there is such a strong relation between the word identity and the type of attachment on the basis of prepositions. Co-occurrence strength values may provide additional clues to informational ties among words; when we investigate the co-occurrences of nouns and prepositions, and of verbs and prepositions, the co-occurrence strength value could also indicate whether the prepositional phrase is attached to the noun or to the verb in the syntactic tree.

In this study, we use two machine learning algorithms to decide on PP attachment. In line with the case study for English, introduced in Ratnaparkhi et al. (1994), we collect a training set of Dutch PP attachment instances from a syntactic treebank. Collection of this data is described in section 4.2. The relation between the identity of the preposition and the attachment is statistically analyzed in section 4.3, subsequently we extract lexical head features from the treebank occurrences, and enrich this data with co-occurrence information derived from a large text corpus (section 4.4). Using the same features, we analogously build a held-out test corpus for which prosodic labelling is available. The machine learning task, involving automatic parameter and feature selection, is described in section 4.5. In this section, we also give the results of the cross-validation experiments on the original data and on the held-out data. Employing the learned PP attachment modules for filtering phrase boundary allocation is discussed in section 4.6, where we test on the held-out written text corpus. We discuss our findings in section 4.7.

4.2 Selection of material

From the syntactic treebank of the Corpus Gesproken Nederlands (CGN, Spoken Dutch Corpus)[1], development release 5, we manually selected 1004 phrases that contain [NP PP] or [PP PP] sequences. Annotated according to protocol (van der Wouden et al., 2002), all PP's have been classified into noun or verb attachment. This classification yields 398 phrases (40%) with a verb attached PP and 606 phrases (60%) with a noun attached PP.

[1]The *Spoken Dutch Corpus* is a database of contemporary Dutch as spoken by adults in the Netherlands and Flanders. The project is funded by the Flemish and Dutch governments and the Netherlands Organization for Scientific Research NWO. Its homepage is `http://lands.let.kun.nl/cgn/ehome.htm`.

Additionally, as a held-out corpus for testing the efficacy of PP attachment information for prosodic phrasing, we selected 157 sentences from various newspaper articles and e-mail messages, of which part has been annotated to obtain the reference transcription (see Chapter 2). A held-out corpus is a corpus which is not used for training, but it is held apart for testing only. We selected this corpus because part of it had been annotated earlier on prosodic phrasing through a reference transcription of ten phonetic experts (as described in Chapter 2). All selected 157 sentences contain either [NP PP] or [PP PP] sequences. To obtain a "golden standard" we manually classified all PP's into NOUN and VERB attachment, according to the "single constituent test" (Paardekooper, 1977). This test states that every string of words that can be placed in front of the finite verb, forms a single constituent. Thus, if and only if an [NP PP] or [PP PP] sequence can be fronted (as in example 4.2a), it forms a single NP containing a noun attached PP. If an [NP PP] or [PP PP] sequence can not be fronted (as in example 4.2b), the PP is verb attached. This classification resulted in 66 phrases (i.e. 42%) with a verb attached PP and 91 phrases (i.e. 58%) with a noun attached PP.

(4.2) (a) *Pizza met ansjovis, eet hij.*
Pizza with anchovies, eats he.

(b) **Pizza met een vork, eet hij.*
*Pizza with a fork, eats he.

4.3 Relation preposition identity and PP attachment

We investigated the relation between the identity of the preposition and the attachment type. It appears that for instance the preposition *van* (from, of) is mainly noun attached (i.e. the preposition introduces a noun attached PP), whereas many other prepositions are mainly verb attached (i.e. the prepositions introduce a verb attached PP). For all 157 phrases of the held-out corpus we listed the preposition together with the attachment. The same was done for the 1004 instances of the Spoken Dutch Corpus (CGN) data. For every preposition we counted the number of times it was noun or verb attached. Figure 4.1 shows the percentages of noun and verb attachment per preposition. If a certain preposition is 100% noun attached, that preposition is 0% verb attached, since the two attachment categories are complementary.

Considering this classification, for some prepositions we are likely to obtain high performance scores (i.e. a high number of correct attachment predictions) for predicting the attachment on the basis of preposition identity alone. However, for prepositions which are almost as often noun attached as verb attached (such as *aan, in, met* (at, in, with)), we will need to add other features to be able to predict the attachment correctly.

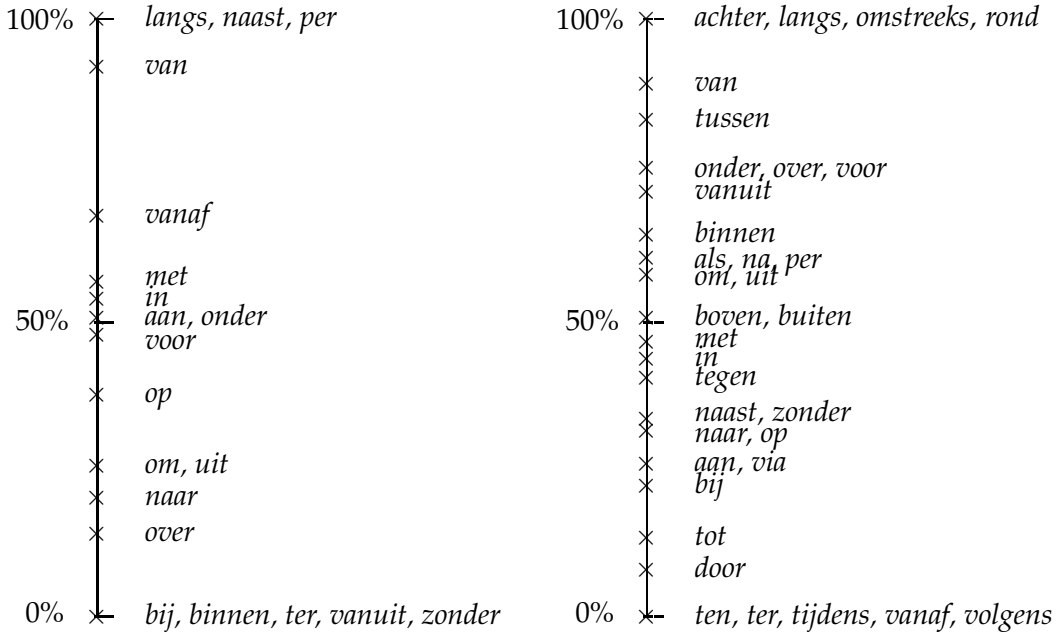


Figure 4.1: *Left panel: percentage of noun attachment per preposition for held-out data. Right panel: percentage of noun attachment per preposition for CGN data. 100% noun attachment means 0% verb attachment, since the two attachment categories are complementary.*

4.4 Feature engineering

4.4.1 Lexical features

Analogous to Ratnaparkhi et al. (1994), we (manually) selected as features the four lexical heads of the phrases involved in the attachment. We used the manually annotated phrasing and function labelling to determine the heads of all involved phrases. First, the noun head (N1) of the preceding NP or PP that the focus PP might be attached to; second, the preposition (P) of the PP to be attached; third, the verbal head (V) of the clause that the PP is in; and fourth, the noun head (N2) of the PP to be attached.

(4.3) *. . . dat zijn zoontje [NP de remmen] [PP van zijn fiets] [VP had gemaakt].*
N1 P N2 V

". . . that his little son had repaired the breaks of his bike."

In example 4.3 *fiets* attaches to the noun *remmen*. The example construction of an instance is thus stored in the data set as the following comma-separated 4-feature instance labelled with the NOUN attachment class:

```
remmen, van, fiets, gemaakt, NOUN.
```

4.4.2 Co-occurrence strength values

Several metrics are available that estimate to what extent words or phrases belong together informationally. Well known examples of such co-occurrence strength metrics are mutual information (Church and Hanks, 1991), chi-square and log likelihood (Dunning, 1993). Co-occurrence strength values are typically estimated from a very large corpus. Often, these corpora are static and do not contain neologisms and names from later periods. In this chapter, we explore an alternative by estimating co-occurrence strength values from WWW counts. The WWW can be seen as a dynamic corpus: it contains new words that are not yet incorporated in other (static) corpora. Another advantage of using the WWW as a corpus is that it is the largest freely and electronically accessible corpus for most languages (including Dutch). Consequently, frequency counts obtained from the WWW are likely to be much more robust than those obtained from smaller corpora. If co-occurrence strength values correlate with PP attachment, then the WWW could be an interesting robust background source of information. This reasoning was introduced by Volk (2000), who performed a study in which the WWW was used to resolve PP attachment. Following this, the second step in engineering our feature set was to add co-occurrence strength values for Dutch words, derived from WWW counts.

Explored methods

We explored three methods in which the co-occurrence strength value was used to predict the attachment for all 1004 phrases from the CGN. The first method is a replication of the study by Volk (2000). In this study co-occurrence strength values were computed for the verb within close vicinity of the preposition Cooc(VnearP) and for the noun within close vicinity of the preposition Cooc(NnearP). Volk (2000) assumes that the higher value of Cooc(VnearP) and Cooc(NnearP) decides the attachment. According to this assumption we say that if Cooc(VnearP) is the higher value, the PP is verb attached. If Cooc(NnearP) is the higher value, the PP is noun attached. When only Cooc(NnearP) was available (because the phrase did not contain a verb), the decision for noun or verb attachment was based on comparison of Cooc(NnearP) with a threshold of 0.50 (co-occurrence strength values are between 0.00 and 1.00). This is the threshold used by (Volk, 2000). For the Cooc(VnearP) and Cooc(NnearP) we used the advanced search function NEAR of the WWW search engine Altavista (Altavista, 2002). This function restricts the search to the appearance of two designated words at a maximal distance of 10 words, which is the default. The search is performed for both possible orders of appearance of the two designated words. We restricted the search to documents which were automatically identified by the search engine as being written in Dutch.

The second method uses only the Cooc(NnearP). This co-occurrence strength value is compared to a threshold. Co-occurrence strength values higher than the threshold predict that the PP is attached to the noun. First, we compare the Cooc(NnearP) to the threshold of 0.50, which is also used in method 1, when the Cooc(VnearP) is unavailable. Next, we determine an optimal threshold on noun attachment, since in an experiment on the perceptual costs of phrasing errors (described in Chapter 3) we found that incorrect phrase boundary insertions are less acceptable to the listener than incorrect boundary omissions. This means that we should optimize on noun attachment, which results in only few incorrect boundary insertions. We determined the optimal co-occurrence threshold by computing the performance measures for several thresholds, using 10-fold cross validation. 10-fold cross validation means that the whole corpus is divided in 10 partitions, 9 of these are used for training and 1 is used for testing. Training and testing is carried out 10 times, so that all partitions are used for testing once. We found that the optimal threshold for optimization on noun attachment should be 0.36. Co-occurrence strength values higher than the threshold predict that the PP is attached to the noun.

Thirdly, we tested a variant on the second method by computing the co-occurrence strength value of a noun immediately succeeded by a preposition Cooc(N P), because in our test set there cannot be a word between N1 and P. For the Cooc(N P) we searched for exact multi-word phrases: "<noun> <prep>". This function restricts the search to the appearance of the two adjacent words in the indicated word order. The Cooc(N P) is compared to a threshold of 0.50, where co-occurrence strength values higher than the threshold again predict a noun attached PP. Again, an optimal threshold is determined by computing the performance measures for several thresholds, using 10-fold cross validation. We found that the optimal co-occurrence threshold for optimization on noun attachment should be 0.07. Co-occurrence strength values higher than the threshold predict that the PP is attached to the noun.

The general formula for computing the co-occurrence strength value of two terms is given by function (4.4) as proposed by Volk (2000). This method is based on the respective frequency of X and the joint frequency of X with a given preposition; where P stands for preposition and X can be either a noun or a verb.

$$cooc(X\ P) = \frac{freq(X\ P)}{freq(X)} \tag{4.4}$$

The number of found documents according to the above described search methods was used for *freq(X P)*. The *freq(X)* was derived from the WWW by performing a separate search for the single word form. The notion co-occurrence strength value could also be referred to as relative frequency estimate of the conditional probability that a preposition co-occurs with a certain noun or verb.

Results for the respective methods

We compared the co-occurrence strength values for noun and verb attachment computed according to the second and third method, since these values will be compared to a threshold for predicting the attachment. The Cooc(NnearP) was significantly different for noun and verb attachment ($t = -11.65, p < 0.001, df = 1002$). The Cooc(N P) was also significantly different for noun and verb attachment ($t = -12.43, p < 0.001, df = 1002$). For the first method we did not compute the difference in co-occurrence strength values for noun and verb attachment, since we directly compare Cooc(VnearP) and Cooc(NnearP), where the higher value predicts the attachment.

Table 4.1 shows the performance measures accuracy, precision, recall and F_β-value, for both noun and verb attachment, for the 1004 instances derived from the CGN. Also a baseline is shown, which gives the performance measures when noun attachment was predicted for all 1004 phrases.

Table 4.1: *Performance measures for predicting PP attachment on the basis of three variants of co-occurrence strength values, with fixed and optimal thresholds.*

		NOUN attachment			VERB attachment		
method + threshold	**accuracy**	**precision**	**recall**	**$F_{\beta=1}$**	**precision**	**recall**	**$F_{\beta=1}$**
NnearP or VnearP	62	71	62	66	51	61	56
NnearP (th = 0.50)	61	83	44	58	51	86	64
NnearP (th = 0.36)	64	75	61	67	54	71	61
N P (th = 0.50)	46	96	11	19	42	99	59
N P (th = 0.07)	67	84	54	65	55	87	67
baseline	60	60	100	75	-	0	-

Table 4.1 shows that Method III with the optimal threshold of 0.07 has the best accuracy on PP attachment. Although it is not the best in all respects, we prefer this method, because it uses co-occurrence strength values for adjacent nouns and prepositions in the order in which they appear in the text, this in analogy with the fact that adjacency gives more information than that obtained with the NEAR function, in PP attachment. The difference between these two is illustrated in example 4.5. In sentence 4.5a the ambiguous PP is '*naar school*'. With the search function for adjacent N and P, the search for the combination '*auto - naar*' will not deliver this combination as found in sentence 4.5b, whereas with the NEAR function sentence 4.5b will be found. This sentence is not representative for the combination '*auto - naar*', since in sentence 4.5b the N and P are in the opposite word order and at a distance of 10 words. According to this reasoning we added the Cooc(N P) feature as the eleventh feature to our data sets for both corpora.

(4.5) (a) *Hij heeft de kinderen met de auto naar school gebracht.*
He has the children by car to school brought.
"He has brought the children to school by car."

(b) *Hij heeft naar de garage gebeld, om een afspraak voor de auto te maken.*
He has to the garage called, to an appointment for the car to make.
"He has called the garage to make an appointment for the car."

4.5 Machine learning

Machine learning algorithms extrapolate from the example to new input cases, either by extracting regularities from the examples for instance in the form of rules or decision trees, or by a more direct use of analogy in lazy learning algorithms such as memory-based learning. We chose to use two machine learning algorithms in our study: rule induction as implemented in RIPPER (Cohen, 1995) (version 1, release 2.4) and memory-based learning MBL (Aha et al., 1991; Daelemans et al., 1999), as implemented in the TiMBL software package (Daelemans et al., 2002).

Rule induction

Rule induction is an instance of "eager" learning, where effort is invested in searching for a minimal-description-length rule set that covers the classifications in the training data. The rule set can then be used for classifying new instances of the same task. RIPPER (Cohen, 1995) induces rule sets for each of the classes in the data, maximizing accuracy and coverage for each induced rule. The method starts with the ordering of all classes in the training data (for the experiments described here that is NOUN and VERB). The rule induction algorithm finds a rule set that separates the least frequent class from the remaining classes. All instances covered by the learned rule set are then removed from the data set, and the algorithm separates the next least frequent class from the remaining classes. This process is repeated until a single class remains. This class, which is the most frequent one, will be used as default class.

Memory-based learning

Memory-based learning, in contrast, is "lazy", meaning that learning is merely the storage of training examples in memory and it generalizes by using intelligent similarity metrics. The category of the most similar example(s) is used as a basis for extrapolating the category of the test example.

Memory-based learning treats a set of labelled (classified) training instances as points in a multi-dimensional feature space, and stores them as such in an 'instance base' in memory. An instance consists of a fixed-length vector of feature-value pairs, and an information field containing the classification of that particular instance. After the instance base is stored, new (test) instances are classified by matching them to all instances in memory, and by calculating with each match the 'distance' between the in-

stance in memory and the new instance. The classification of new material in MBL essentially follows the k-nearest neighbor classification rule (Cover and Hart, 1967) of searching for nearest neighbors in memory, and extrapolating their (majority) class to the new instance.

The strength of memory-based language processing is that it performs no abstraction, for instance through defining rules, which allows it to deal with productive but low-frequency exceptions (Daelemans et al., 1999). Taking these exceptions into account is useful, since it is difficult to discriminate between noise on the one hand, and valid exceptions and irregularities on the other hand.

4.5.1 Experiments

A central issue in the application of machine learning is the setting of algorithmic parameters; both RIPPER and MBL feature several parameters of which the values can seriously affect the bias and result of learning. Also, the particular features that are selected as well as the amount of data available will determine which parameters are optimal. Few reliable rules of thumb are available for setting parameters. To estimate appropriate settings, a big search space needs to be sought through in some way, after which one can only hope that the estimated best parameter setting is also good for the test material – it might be overfitted on the training material.

Fortunately, we were able to do a pseudo-exhaustive search (testing a selection of sensible numeric values where in principle there is an infinite number of settings), since the CGN data set is small (1004 instances). For MBL, we varied the following parameters systematically in all combinations (see Daelemans et al. (2002) for a description of these parameters):

- the k in the k-nearest neighbor classification rule: 1, 3, 5, 7, 9, 11, 13, 15, 19, 21, 25, 29, 35, 39, 45, 49, 55 and 65
- the type of feature weighting: none, gain ration, information gain, and chi-squared
- the similarity metric: overlap, or MVDM with back-off to overlap at levels 1 (no back-off), 2, and 5
- the type of distance weighting: none, inverse distance, inverse linear distance, and exponential decay with $\alpha = 1$, $\alpha = 2$ and $\alpha = 4$

For RIPPER we varied the following parameters:

- the minimal number of instances to be covered by rules: 1, 2, 5, 10, 25, 50
- the class order for which rules are induced: increasing and decreasing frequency
- allowing negation in nominal tests or not
- the number of rule set optimization steps: 0, 1, 2

We performed the full matrix of all combinations of these parameters for both algorithms in a nested 10-fold cross-validation experiment. First, the original data set was split in ten partitions of 90% training material and 10% test material. Second, nested 10-fold cross-validation experiments were performed on each 90% data set, splitting it again ten times. To each of these 10×10 experiments all parameter variants were applied. Per main fold, a nested cross-validation average performance was computed; the setting with the average highest F-score on noun attachment is then applied to the full 90% training set, and tested on the 10% test set.

4.5.2 Results

First, we report on the results obtained directly from the nested cross-validation experiment on the Spoken Dutch Corpus data. Second, we report on applying the best overall parameter settings of RIPPER and MBL to the external validation corpus of newspaper and e-mail data.

Internal results: Spoken Dutch Corpus data

First, we carried out experiments using MBL and RIPPER to obtain the performance score per feature (i.e. for the lexical features and the co-occurrence strength value). Table 4.2 for MBL and Table 4.3 for RIPPER show that for optimizing on noun attachment the scores for all five features are reasonably robust and add information. However, the performance measures for the best parameter setting for using all features are considerably higher. For testing on single features the performance is lower, especially on verb attachment. For MBL and RIPPER only the P-feature obtains a performance on verb attachment that approximates the performance for testing on all features. For MBL the co-occurrence feature also shows a reasonable performance on both noun and verb attachment. Appendix D gives the results for MBL for testing on combinations of two features. These results show that the performance on using two features is better than on single features, but not as well as on using all features.

Table 4.2: *Performance measures in percentages for predicting PP attachment in the CGN material (1004 instances) by* MBL.

MBL		**NOUN attachment**			**VERB attachment**		
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
all	77	81	81	81	71	69	70
N1	67	69	85	76	63	39	48
P	73	81	72	76	63	73	67
N2	62	64	86	74	52	24	32
V	59	62	82	71	46	22	30
Cooc(N P)	68	74	73	74	59	60	59
baseline	60	60	100	75	-	0	-

Table 4.3: *Performance measures in percentages per feature for predicting PP attachment in the CGN material (1004 instances) by* RIPPER.

RIPPER		NOUN attachment			VERB attachment		
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
all	70	74	83	77	52	50	49
N1	64	63	98	77	83	11	18
P	69	74	81	76	64	52	53
N2	66	64	98	78	87	17	27
V	62	61	99	76	55	4	7
Cooc(N P)	65	65	93	76	49	21	27
baseline	60	60	100	75	-	0	-

The performance measures for both algorithms are considerably higher than the baseline which indicates the performance when always noun attachment is predicted. MBL produces the highest accuracy, 77%, which is significantly higher than the accuracy of RIPPER, 70% ($t = 2,87, p < 0.05, df = 18$). MBL also produces the highest F-score, 81%, which is significantly higher than that of RIPPER, 77% ($t = 2,97, p < 0.05, df = 18$).

The best overall cross-validated setting for MBL was no feature weighting, $k = 25$, MVDM, and exponential decay distance weighting with $\alpha = 2$. It has been argued in the literature that high k and distance weighting is a sensible combination (Zavrel et al., 1997). More surprisingly, no feature weighting means that every feature is regarded equally important.

For RIPPER, the best overall cross-validated parameter setting is to allow a minimum of one case to be covered by a rule, induce rules on the most frequent class first (noun attachment), allow negation (which is, however, not used in the end), and run one optimization round. The most common best rule set is the following:

1. if P = *van* then NOUN
2. if Cooc(N P) > 0.07 then NOUN
3. if P = *voor* then NOUN
4. if there is no verb then NOUN
5. else VERB

This small number of rules test on the presence of the two prepositions *van* (from, of) and *voor* (for, before) which often co-occur with noun attachment (i.e. on the whole data set 351 out of 406 occurrences of the two prepositions), a value of Cooc(N P) similar to the optimal co-occurrence threshold reported earlier (0.07), and the absence of a verb (which occurs in 27 instances).

External results: newspaper and e-mail data

We evaluated the results of applying the overall best settings on the held-out data (i.e. the 157 sentence external newspaper and e-mail material). Performance measures for MBL are given in Table 4.4 and for RIPPER in Table 4.5. These results roughly correspond with the previous results (i.e. the proportions are the same). Performance measures are again considerably above baseline, although lower than for CGN data. MBL attains lower precision but higher recall than RIPPER on noun attachment. Again, for testing on single features the performance is lower, especially on verb attachment. For RIPPER only the P-feature obtains a performance on verb attachment that approximates the performance for testing on all features. For MBL the same is true for both the P-feature and the co-occurrence feature.

Table 4.4: *Performance measures in percentages per feature for predicting PP attachment in the newspaper and e-mail material (157 instances) by* MBL.

MBL		**NOUN attachment**			**VERB attachment**		
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
all	66	69	74	72	62	55	58
N1	61	60	94	74	69	16	27
P	64	68	70	69	58	57	58
N2	57	57	96	72	43	4	8
V	55	57	90	70	36	7	12
Cooc(N P)	65	68	74	71	60	52	56
baseline	58	58	100	73	-	0	-

Table 4.5: *Performance measures in percentages per feature for predicting PP attachment in the newspaper and e-mail material (157 instances) by* RIPPER.

RIPPER		**NOUN attachment**			**VERB attachment**		
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
all	66	70	71	71	61	60	60
N1	58	58	100	73	100	1	3
P	64	67	72	70	58	52	55
N2	57	57	99	73	50	1	3
V	57	57	100	73	0	0	0
Cooc(N P)	62	62	91	74	67	24	35
baseline	58	58	100	73	-	0	-

4.6 Contribution to phrase boundary allocation

In this experiment we investigated the added value of having PP attachment information available in a straightforward existing prosodic phrasing algorithm for Dutch (described in Chapter 2). This phrasing algorithm uses syntactic information and sentence length for the allocation of prosodic phrase boundaries. For a subset (44 phrases) of the held-out corpus, we compared the allocation of boundaries according to the phrasing algorithm, and according to the same algorithm complemented with PP attachment information, to the reference transcription of ten phonetic experts (see Chapter 2). This reference transcription was not available for all 157 phrases of the newspaper and e-mail data (i.e. the held-out corpus).

Table 4.6: *Performance measures in percentages for the revised phrasing algorithm complemented with PP attachment information from* MBL *and* RIPPER *(on 44 instances).*

phrasing algorithm	**accuracy**	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
phrasing	91	65	81	72
phrasing + MBL	92	70	79	74
phrasing + RIPPER	92	71	80	75
phrasing + golden standard	93	72	81	77

Table 4.6 shows the performance measures for this comparison, indicating that the improvement from PP attachment information is largely in precision. Indeed, blocking certain incorrect placements of phrase boundaries improves the precision on boundary placement. MBL and RIPPER attain an improvement of five or six points in precision. Although they incorrectly prevent three or two *intended* phrase boundaries (when compared to the manual classification mentioned in section 4.2), they do in fact correctly prevent *unintended* boundaries in eleven other cases. Some instances of the latter are given in example 4.6, where [] indicates the location of the prevented boundary.

(4.6) (a) . . . *afschaffing [] van het laatste recht* . . .
. . . abolition [] of the final right . . .

(b) . . . *het grootste deel [] van Nederland* . . .
. . . the biggest part [] of the Netherlands . . .

(c) . . . *de straatlantaarns [] langs de provinciale weg* . . .
. . . the street lights [] along the provincial road . . .

Table 4.6 also shows the performance measures for the phrasing algorithm complemented with the "golden standard". These results indicate the maximally attainable improvement of the phrasing algorithm using correct PP attachment information. The results obtained with MBL and RIPPER come close to this maximally attainable improvement.

4.7 Discussion

We have presented experiments on isolated learning of PP attachment in Dutch, and on using predicted PP attachment information for filtering out incorrect placements of prosodic boundaries. First, PP attachment was learned by the best optimized machine learner, MBL, at an accuracy of 77%, an F_β-score of 81% on noun attachment, and 70% on verb attachment. Since we found that incorrectly inserted phrase boundaries are less acceptable to the listener than incorrectly omitted ones (see Chapter 3), the machine learning algorithms were optimized (via nested cross-validation experiments and pseudo-exhaustive parameter selection) on noun attachment. That type of attachment typically prevents a prosodic boundary. We show that improvements are made in the precision of boundary allocation; a high precision means few incorrect inserted boundaries.

Comparing the eager learner RIPPER with the lazy learner MBL, we saw that RIPPER typically induces a very small number of safe rules, leading to reasonable recall (83%) but relatively low precision (74%). Although the recall for RIPPER is higher, MBL performs better on all other measures. The bias of MBL to base classifications on all training examples available, no matter how low-frequent or exceptional, resulted in higher performance measures, indicating that there is more reliable information in local matching on lexical features and the co-occurrence feature than RIPPER estimates. However, with a larger training corpus, we might not have found these differences in performance between the two learning algorithms.

In engineering our feature set we combined disjoint ideas on using both lexical features and co-occurrence strength values. The lexical features were sparse, since they only came from the 1004-instance training corpus, while the co-occurrence feature was very robust and "unsupervised", based on the very large WWW. Only the combination of these five yielded the best performance – individually the features do carry information, but always less than the combination. This suggests that it is essential to employ features that each add unique information, either on lexical identity or on co-occurrence strength.

It would be interesting to investigate ways of embedding our approach for predicting PP attachment within other, more general parsing algorithms. At present there are no parsers available for Dutch that disambiguate PP attachment, which leaves the comparison between PP attachment as an embedded subtask of a full parser with our approach as future work.

Disambiguation of argument and condition

5

From previous evaluation we learned that accents are not always allocated correctly. Especially accentuation of the sentence final verb is often incorrect. The identity of the preceding nominal constituent (whether it is an argument or a condition) is of importance for the accentuation of the verb. In this chapter we first discuss de definition of "argument" and "condition". Next, we describe machine learning experiments for predicting the identity of the nominal constituent. Finally, we discuss the merit of being able to discriminate between argument and condition for accent assignment in synthetic speech.

5.1 Introduction

One of the factors determining the acceptability of synthetic speech is the appropriate allocation of sentence accents. As we described in Chapter 2, a major factor that causes difficulties in the correct allocation of accent on sentence final verbs is the lack of reliable information about syntactic structure. Previous research showed that accentuation of this verb depends among other things on the nominal constituent that precedes the verb in the surface structure (Dirksen, 1992a). In this chapter we will use machine learning algorithms to predict the status of this nominal constituent, which can be either an argument (5.1) or a condition (5.2).

(5.1) *Hij heeft het +hele +boek -gelezen.*
He has the entire book read.
"He has read the entire book."

(5.2) *Hij heeft de +hele nacht +gelezen.*
He has the entire night been reading.
"He has been reading the entire night."

Based on the identity of the nominal constituent, we decide whether or not the sentence final verb phrase should be accented. In section 5.2, we discuss the classification into argument and condition and the implications for further research. In section 5.3, we describe machine learning experiments for predicting the identity of the nominal constituent (i.e. argument or condition). In section 5.4, we discuss the merit of using information about argument versus condition as a filter for the assignment of accent to sentence final verbs.

5.2 Argument versus Condition

5.2.1 SAAR as starting point

Our starting point for accent placement in the sentence final verb phrase is the Sentence Accent Assignment Rule (SAAR) (Gussenhoven, 1982, 1984). For this we distinguish three semantic sentence constituents: **Argument**, **Predicate** and **Condition**. SAAR is applied in two steps. The first step is allocation of focus domains (indicated with [ ]), and the second step is deciding on the exact location of the accent in that specific domain.

Domain allocation: A(X)P → [A(X)P]
P(X)A → [P(X)A]
Y → [Y]

where X and Y stand for Argument, Condition and Predicate.

We will concentrate on cases where the predicate is preceded by either a condition or an argument (as shown below).

ACP → [A][C][P]
AAP → [A][AP]

If preceded by a condition the predicate is a separate focus domain. If preceded by an argument, the predicate and the argument comprise one focus domain. To every focus domain [ ] at least one accent will be assigned. If an argument and a predicate comprise one focus domain, the accent will be on the argument (because it is strong in relation to the predicate). The scope of the accent concerns the whole predicative expression. If the nominal constituent is a condition, it constitutes a separate focus domain. Then, the scope of the accent does not concern the predicate and the predicate will receive an accent. In case of sentence 5.1 and sentence 5.2 SAAR predicts the accentuation correctly. In sentence 5.1 the verb is not accented, whereas in sentence 5.2 the verb is accented.

The general question is which constituents count as arguments and which as conditions. Before asking this question, we will first consider the validity of SAAR.

5.2.2 Validity of SAAR

The general validity of SAAR has been questioned before: Gussenhoven (1992) and Marsi (2001) mention some exceptions. In the following sections we will discuss these and other exceptions.

Topicalization

The first exception discussed is the topicalized argument, which can not lead to deaccentuation of the verb. Marsi (2001) shows this on the basis of sentence 5.3. In example 5.3a the pronoun "Jan" is not topicalized, whereas it is in example 5.3b.

(5.3) (a) *+Jan -slaat me.*
John hits me.
"John hits me."

(b) *+Jan +sla ik.*
John hit I.
"It's John I hit."

However, the conclusion that topicalized arguments can not lead to deaccentuation of the verb might be incorrect. We presume that the artificiality of sentence 5.3b plays an important role, and that accentuation of the verb in example 5.3b might be due to rhythmical aspects (as discussed by Schmerling (1976) and Baart (1987)). If we con-

struct a more natural example containing a topicalized argument, the verb has to be deaccented. Sentence 5.4 is such an example.

(5.4) *(Wat moest ik nog doen?) Oja, Bob +Hartman moest ik nog bellen.*
(What should I still do?) Oh, Bob Hartman should I still call.
"(What was it that I still had to do?) Oh, I still had to call Bob Hartman."

Extraposition of arguments

The second exception mentioned is the extraposed argument. Marsi (2001) states that deaccentuation of the verb is often more inappropriate with an extraposed argument. In sentence 5.5a the verb should not be accented, in accordance with SAAR, because it is preceded by an argument '*op de stoptrein naar Schiedam Centrum*'. In sentence 5.5b the verb has to be accented according to Marsi. However, if we assume that focus domain allocation is blocked by prosodic boundaries, SAAR would not apply to sentence 5.5b. In fact, if we delete the boundary, as in sentence 5.5c, our intuition is that leaving the verb unaccented is acceptable and this is in fact correctly predicted by SAAR.

(5.5) (a) *U hebt +acht +minuten / om op de +stoptrein naar +Schiedam*
You have eight minutes for to the slow-train to Schiedam
+Centrum over te stappen ///
Centre to transfer.
"You have eight minutes to transfer to the slow train to Schiedam Centre."

(b) *U hebt +acht +minuten / om +over te stappen / op de +stoptrein*
You have eight minutes for to transfer to the slow-train
naar +Schiedam +Centrum ///
to Schiedam Centre.

(c) *U hebt +acht +minuten / om over te stappen op de +stoptrein naar*
You have eight minutes for to transfer to the slow-train to
+Schiedam +Centrum ///
Schiedam Centre.

Regular order of constituents

Another example can be constructed that is not in accordance with SAAR. In sentence 5.6a the predicate is immediately preceded by a condition. Thus, according to SAAR the predicate should be accented. However, in this example the predicate is not accented. When we have a closer look at sentence 5.6a we notice that this is an instance of irregular, or marked word order. By marked order we mean that the order of the constituents '*een ijsje*' and '*bij de molen*' is not the most natural: the order of the constituents in sentence 5.6b and sentence 5.6c will come up more often.

(5.6) (a) *We hebben een ijsje bij de molen -gekocht.*
We have an ice cream at the mill bought.
"We have bought an ice cream at the mill."

(b) *We hebben bij de molen een ijsje -gekocht.*
We have at the mill an ice cream bought.

(c) *We hebben een ijsje -gekocht bij de molen.*
We have an ice cream bought at the mill.

Deviations from the regular order often result from a contrast effect. By changing the order of the constituents a certain constituent can be emphasized. As we mentioned before, in the study reported here we leave contrast effect aside. Therefore, we will only discuss examples with a regular order of constituents. In apparent counterexamples we will investigate whether it is a matter of marked order of constituents.

Semantic predictability

Another apparent counterexample has to do with the semantic predictability of an argument to a verb. If the sentence final verb phrase is preceded by an argument, we expect the verb to be unaccented and when the verb is preceded by a condition we expect the verb to be accented. However, semantic and lexical aspects also play a role in the accentuation of the verb (Kruyt, 1985). Certain verbs (such as '*waarschuwen, shockeren, verzekeren, dreigen*' (to warn, to shock, to insure, to threathen)) have a higher accentability, while other verbs have a lower accentability (such as '*houden, bereiken, spelen, controleren*' (to keep, to reach, to play, to inspect)) Kruyt (1985, Fig. 4.2). Sentence 5.7 and 5.8 are examples with an accented sentence final verb, whereas in sentence 5.9 and 5.10 the verb is not accented.

(5.7) *De regering heeft de leiders van de staking +gewaarschuwd.*
The government has the leaders of the strike warned.
"The government has warned the leaders of the strike."

(5.8) *De spaarbank heeft de ontwikkelaars van woningbouwprojecten +verzekerd.*
The savings bank has the developers of housing construction projects insured.
"The savings bank has insured the developers of housing construction projects."

(5.9) *De politie heeft alle auto's op versleten banden -gecontroleerd.*
The police has all cars for worn out tires -inspected.
"The police has inspected all cars for worn out tires."

(5.10) *Het residentie-orkest heeft Nederlandse avant-garde -gespeeld.*
The residential orchestra has Dutch avant-garde played.
"The residential orchestra has played Dutch avant-garde."

In sentences 5.7 and 5.8 the verb is preceded by an argument. Yet, in contrast with SAAR the verb is accented. In sentences 5.9 and 5.10 the verb is preceded by a condition. Yet, in contrast with SAAR the verb is not accented. These deviations from SAAR indicate that there might be an effect of the identity of the verb.

We were able to reason away many of the apparent exceptions, however there remain some cases (such as verb identity) that require closer investigation. For the research in this chapter we will nevertheless hold on to the correctness of SAAR, stating that the predicate verb will not be accented if preceded by an argument, whereas the predicate will be accented if preceded by a condition within the same intonational domain. In the next section we will address the question of which constituents are arguments and which constituents are conditions.

5.2.3 Distinction argument - condition

Various tests have been described for making the distinction between argument and condition (Gussenhoven, 1984; Baart, 1987; Marsi, 2001). In general, constituents that can be left out (can be deleted) and are not subcategorized for by the verb are conditions. All other constituents are arguments. A well known test to decide whether or not the constituent can be removed from the matrix phrase is the so called "*en wel...*" test. The nominal constituent is a condition if the constituent can be removed from the sentence and can be placed after "*en wel*" (*and more specifically*), and if the resulting sentence forms a semantically and syntactically correct sentence. If the resulting sentence is not syntactically correct (indicated by * in the examples) the nominal constituent is an argument. Instance 5.11 is an example of this test: sentence 5.11a concerns a condition and sentence 5.11b concerns an argument. This is in accordance with the analysis of sentences 5.1 and 5.2.

(5.11) (a) *Hij heeft gelezen, en wel de hele nacht.*
He has read, and more specifically the entire night.
"He has been reading, and more specifically the entire night."

(b) **Hij heeft gelezen, en wel het hele boek.*
He has read, and more specifically the entire book.
"He has read, more specifically the entire book."

From the test for "deletability" information about subcategorization frames of verbs can be derived. If the constituent is an argument, the verb subcategorizes for a certain constituent. If the constituent is a condition, the verb does not subcategorize for that constituent. Subcategorization information is not for all verbs available from a corpus. Besides, some verbs have several subcategorization frames, which introduces an extra ambiguity to resolve, namely which frame is the one that applies to a specific appearance of that verb.

Arguments

As argued before, arguments do not induce accentuation of the sentence final verb phrase. Below, there are some examples of an argument preceding the verb phrase. These examples illustrate that indeed the predicate should not be accented when preceded by an argument. This is in accordance with SAAR (see section 5.2.1).

(5.12) (direct object)

Hij heeft een boek van Wolkers -gelezen.
He has a book by Wolkers read.
"He has read a book by Wolkers."

(5.13) (subject)

Morgen wordt de piano -bezorgd.
Tomorrow will the piano be delivered.
"Tomorrow the piano will be delivered."

(5.14) (subject)

In Zeist is een instrumentenfabriek -afgebrand.
In Zeist did an instruments factory burn down.
"In Zeist an instruments factory did burn down."

(5.15) (indirect object)

Ik heb het boek aan mijn vader -gegeven.
I have the book to my dad given.
"I have given the book to my dad."

(5.16) (prepositional object)

Guus heeft naar de paasvakantie -verlangd.
Guss has for the Easter holidays longed.
"Guss has longed for the Easter holidays."

(5.17) (prepositional object)

Karel heeft urenlang op zijn broer -gewacht.
Charles has for hours for his brother been waiting.
"Charles has been waiting for his brother for hours."

Prepositional objects constitute a special type of arguments, because superficially they resemble conditions. Since we want to decide between argument or condition on the basis of the surface structure, indirect objects (such as example 5.15) and prepositional objects (such as examples 5.16 and 5.17) may give complications in the machine learning experiments. Sentences 5.16 and 5.17 convincingly demonstrate that prepositional objects induce deaccentuation of the predicate.

Conditions

As mentioned above, constituents that can be left out and are subcategorized for by the verb are conditions. Typically, these are adverbs. When we change the order of the constituents of examples 5.12–5.14 for arguments in such a way that a condition precedes the verb (as in examples 5.18–5.20), we see that the accentuation status of the verb changes due to the fact that it is now preceded by a condition instead of an argument.

(5.18) *De boeken van Wolkers worden nog heel vaak +gelezen.*
The books by Wolkers are still very often read.
"The books by Wolkers are still read very often."

(5.19) *De piano wordt morgen +bezorgd.*
The piano will tomorrow be delivered.
"The piano will be delivered tomorrow."

(5.20) *In Zeist is een instrumentenfabriek door brand +verwoest.*
In Zeist was an instruments factory by a fire destroyed.
"In Zeist an instruments factory was destroyed by a fire."

Other examples in which the sentence final verb is preceded by an adverb and in which SAAR correctly predicts the accentuation are given below.

(5.21) (predicative adverb)
De man is ellendig +gestorven.
The man has miserably died.
"The man died miserably."

(5.22) (adverb of time)
De rekening is vorige week +betaald.
The bill has last week been paid.
"The bill has been paid last week."

(5.23) (adverb of manner)
De conferentie is zonder resultaat +geëindigd.
The conference has without results ended.
"The conference ended without results."

(5.24) (adverb of aspect)
Zij heeft haar doel ondanks alle tegenslag +bereikt.
She has her goal despite all bad luck reached.
"She has reached her goal despite all bad luck."

(5.25) (adverb of person)

Het nieuwe boek van Hermans werd door de recensent +besproken.
The new book by Hermans was by the critic discussed.
"The new book by Hermans was discussed by the critic."

However, there are some apparent counterexamples. The verb in sentence 5.26 is not accented (against our expectations).

(5.26) *Het kersverse bruidspaar heeft in een hotel -gelogeerd.*
The fresh bridal couple has in a hotel stayed.
"The fresh bridal couple has stayed in a hotel."

In this example the nominal constituent that precedes the verb is a locative. In the next section we will consider accentuation patterns in sentences with locative expressions more closely.

Locatives
The impression from the examples above is that the sentence final verb is not accented if preceded by a locative. The examples below support this impression.

When we compare sentence 5.27 to sentence 5.6a, we see that in these sentences the order of the constituents is identical. However, whereas instance 5.6a is an example of marked order, instance 5.27 cannot be explained in this manner. We suppose that the verb '*zetten*' (to put) subcategorizes for an object and a locative, and that the verb '*kopen*' (to buy) subcategorizes for an object, but not for a locative.

(5.27) *Hij heeft de tas naast de auto -gezet.*
He has the bag next to the car put.
"He has put the bag next to the car."

(5.28) *De kat heeft uren onder de tafel -gezeten.*
The cat has for hours under the table been.
"The cat has been under the table for hours."

(5.29) *Hij heeft in de tuin -gespeeld.*
He has in the garden been playing.
"He has been playing in the garden."

For these examples we argue that the locatives behave like arguments instead of conditions, because (i) they do not induce accentuation of the predicate, and (ii) they can not be left out. However, there exist locatives that can be left out and that do induce accentuation of the verb (see example 5.30).

(5.30) *Moeder heeft de hele middag in de tuin +gelezen.*
Mother has the entire afternoon in the garden been reading.
"Mother has been reading the entire afternoon in the garden."

This example implies that we can not state that in general locatives behave like arguments. We suppose that accentuation status of the verb is connected with the identity of the verb. Verbs like '*zetten*' and '*zitten*' (to put, to sit) that express an action that intrinsically requires a certain location, subcategorize for the constituent expressing that location. In such cases the verb will not be accented, since the verb subcategorizes for the nominal constituent preceding the verb.

In examples such as 5.29 the location is not intrinsically required by the verb, because the locative can be deleted (as in example 5.31). However, we assume that accent on adverbs has an integrative function (see Baart, 1989), so that the expression "has been playing in the garden" as a whole expresses the predicate. Obviously the suitability of locative-verb combinations constituting a predicate will depend on the particular items to be combined. This is a topic for further research.

(5.31) *Hij heeft de hele middag gespeeld.*
He has the entire afternoon played.
"He has played the entire afternoon."

5.2.4 Implications for further research

When we leave discourse context and contrast effects out of consideration, the review reported above shows that arguments and some locatives do not induce accentuation of the sentence final verb phrase. All other adverbs, including some locatives which are not subcategorized for by the verb, are conditions, and they do induce accentuation of the verb phrase. We conclude that overall SAAR is useful for predicting the accentuation of the sentence final verb phrase. The rules we use for the remainder of this chapter then are:

⇒ *The sentence final verb phrase is not accented*
if it is preceded by an argument
(including locatives that are subcategorized by the verb).

⇒ *The sentence final verb phrase is accented*
if it is preceded by a condition.

We assume that proper accent assignment to sentence final verbs is possible when we are able to predict whether a nominal constituent preceding the verb, is an argument or a condition. In the experiments below we abstract away from context effects such as given-new information.

5.3 Machine learning experiments

As we saw in the previous chapter, state-of-the-art parsers do not provide a complete analysis of the syntactic structure. Certain syntactic ambiguities, such as PP attachment, remain unsolved. The ambiguity we want to resolve in this chapter is the status of the nominal constituent (which is an NP or nominal part of PP) preceding a sentence final verb phrase. Machine learning algorithms appeared to be a useful instrument for predicting PP attachment. Therefore, we now explore machine learning experiments for predicting argument versus condition, to use it as a filter in accentuation of the sentence final verb. In line with the experiments on PP attachment (as described in Chapter 4), we assume that two sources of information should be used for training data: (i) lexical features (e.g. the head words P, N and V), and (ii) a word co-occurrence strength value.

Two machine learning algorithms are applied for the classification of nominal constituents into argument and condition. We selected a training corpus from a syntactic treebank and a held-out corpus for which prosodic labelling is available. Collection of both corpora is described in section 5.3.1. We extracted lexical head features from the treebank occurrences, and we added co-occurrence information derived from the WWW as an extra feature (see section 5.3.2). The setup of the machine learning experiments is described in section 5.3.3. In section 5.3.4, we give the results of these experiments by means of the performance measures accuracy, precision, recall and F_β-value, for the training corpus and the held-out data. And in section 5.3.5, we discuss the merit of having information about argument versus condition, for accentuation in synthetic speech.

5.3.1 Selection of material

From the Corpus Gesproken Nederlands (CGN, Spoken Dutch Corpus), development release 6, we manually selected 1613 sentences that contain a sentence final verb phrase preceded by a nominal constituent. Classification of the nominal constituents into argument and condition was done according to protocol (van der Wouden et al., 2002) with manual correction. This classification yields 1348 sentences (84%) with an argument preceding the sentence final verb phrase and 265 sentences (16%) with a condition preceding that verb phrase.

Additionally, we selected a held-out corpus for testing the efficacy of information about the identity of the nominal constituent preceding the sentence final verb phrase for accent assignment. For this corpus we selected 61 sentences from various newspaper articles and e-mail messages. We selected this corpus because part of it had been annotated earlier on accentuation through the reference transcription (see Chapter 2). To obtain a "golden standard" we manually classified all nominal constituents into argument (ARG) and condition (COND), according to the criteria mentioned in section 5.2.4. This classification yields 46 sentences (75%) with an argument preceding the sentence final verb phrase and 15 sentences (25%) with a condition preceding that verb phrase.

5.3.2 Feature engineering

Lexical features

Analogous to the experiments on PP attachment, we selected the lexical heads of the phrases involved: the nominal constituent (whether or not part of a PP) and the sentence final verbal phrase. First, if available, we selected the preposition (P) preceding the nominal constituent; second, the noun head (N) of the nominal constituent; and third, the verbal head (V) of the sentence final verbal phrase.

(5.32) *. . . [PP naar de paasvakantie] [VP verlangd].*
P N V
". . . longed for the Easter holidays."

In example 5.32 *paasvakantie* is an argument to the verb *verlangd*. The example construction of an instance is thus stored in the data set as the following comma-separated 3-feature instance labelled with the argument class:

```
naar, paasvakantie, verlangd, ARG.
```

Co-occurrence strength values

Analogous to the experiments on PP attachment, we added a co-occurrence strength value derived from WWW counts as a fourth feature to our data sets for both corpora. In Chapter 4 we introduced three methods for computing the co-occurrence strength value. For the experiment on PP attachment we reasoned that the preferred value was the co-occurrence strength value for an adjacent preposition and noun. With respect to discrimination between argument and condition we expect to gain information from the co-occurrence strength value for preposition and verb. For the experiment described here we chose to compute the co-occurrence strength value for the preposition in close vicinity with the verb, since the preposition and the verb are not adjacent in the sentence, which means that a co-occurrence strength value for adjacent P and V is not a sensible one.

To obtain the Cooc(PnearV) we used the NEAR function in the WWW search engine Altavista (Altavista, 2002). This function restricts the search to the appearance of two designated words at a maximal distance of 10 words, which is default. The search is performed for both possible orders of appearance of the two designated words. The formula for computing Cooc(PnearV) is given by function 5.33. This method is based on the respective frequency of the verb, and the joint frequency of the preposition and the verb.

$$Cooc(PnearV) = \frac{freq(PnearV)}{freq(V)} \tag{5.33}$$

We restricted the search to documents which were automatically classified by Altavista as being written in Dutch.

5.3.3 Experiments

As we did for PP attachment, we chose to use two machine learning algorithms in this study: rule induction as implemented in RIPPER (Cohen, 1995) (version 1, release 2.4) and memory-based learning (MBL) (Aha et al., 1991; Daelemans et al., 1999) as implemented in the TiMBL software package, version 4.3 (Daelemans et al., 2002). To obtain the best setting of algorithmic parameters we performed a pseudo-exhaustive search as we did for PP attachment. Again, we were able to do this because of the relatively small data set (1613 instances). For MBL we systematically varied the following parameters in all combinations:

- the k in the k-nearest neighbor classification rule: 1, 3, 5, 7, 9, 11, 13, 15, 19, 21, 25, 29, 35, 39, 45, 49, 55 and 65
- the type of feature weighting: none, gain ration, information gain, and chi-squared
- the similarity metric: overlap, or MVDM with back-off to overlap at levels 1 (no back-off), 2, and 5
- the type of distance weighting: none, inverse distance, inverse linear distance, and exponential decay with $\alpha = 1$, $\alpha = 2$ and $\alpha = 4$

For RIPPER we systematically varied the following parameters in all combinations:

- the minimal number of instances to be covered by rules: 1, 2, 5, 10, 25 and 50
- the class order for which rules are induced: increasing and decreasing frequency
- allowing negation in nominal tests or not
- the number of rule set optimization steps: 0, 1, 2

We performed the full matrix of all combinations of these parameters for both algorithms in a nested 10-fold cross-validation experiment (see Chapter 4). We also performed the experiments on the basis of every single feature (P, N, V and Cooc(PnearV)).

5.3.4 Results

Spoken Dutch Corpus data

Table 5.1 lists the performance measures produced by MBL on the CGN data. Table 5.2 lists these performance measures produced by RIPPER. When using all features for both algorithms, we obtained performance measures above the baseline that shows the performance measures when always ARG is predicted.

Table 5.1: *Performance measures in percentages on* ARG *versus* COND *prediction in the CGN material (1613 instances) by* MBL

MBL		ARG			COND		
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
all	89	90	97	93	77	50	61
P	82	83	100	90	35	1	3
N	89	91	96	93	75	55	63
V	80	83	96	89	21	5	8
cooc	80	83	95	89	25	8	12
baseline	84	84	100	91	-	0	-

Table 5.2: *Performance measures in percentages on* ARG *versus* COND *prediction in the CGN material (1613 instances) by* RIPPER.

RIPPER		ARG			COND		
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
all	89	90	97	93	79	48	60
P	81	83	97	89	36	8	12
N	89	90	97	94	81	50	61
V	82	82	100	90	0	0	0
cooc	82	82	100	90	0	0	0
baseline	84	84	100	91	-	0	-

For MBL and RIPPER the results for using only N, are as good as the results for all features. Thus N is as informative as all features together. This means that the noun feature contains necessary trigger words which are typical for conditions (such as '*uur, maand, week*' (hour, month, week). The other single features perform below baseline, indicating that these features do not add information. From this we conclude that addition of other features than only the noun does not increase the performance. When testing on combinations of two or three features, we also obtain the best performance if the noun feature is part of the combination (see Appendix E). The results for testing on these combinations, however, do not exceed the results for using only the noun feature.

The best overall cross-validated parameter setting for MBL using all features, turns out to be the overlap metric, with $k = 1$ and no weighting. No feature weighting means that all features are regarded equally important. No distance weighting means that all neighbors have the same weight.[1]

For RIPPER the best overall cross-validated parameter setting is to allow a minimum number of 1 case covered by a rule, induce rules for the less frequent class first (COND),

[1]Note that with $k = 1$ for the nearest neighbor classification rule, distance weighting is not a sensible setting.

allow negation and run 2 optimization rounds. An example of the resulting rule set is the following:

1. if N = *keer* then COND
2. if N = *beetje* then COND
3. if N = *jaar* then COND
4. if Cooc(PnearV) > 0.0059 and < 0.1955 and P = *met* then COND
5. if Cooc(PnearV) > 0.0049 and < 0.3136 and > 0.3114 then COND
6. if Cooc(PnearV) > 0.0049 and < 0.1266 and P = *na* then COND
7. else ARG

Newspaper and e-mail data

We evaluated the results of applying overall best settings on the 61 sentence newspaper and e-mail data. Table 5.3 lists the performance measures produced by MBL on these 61 sentences of the held-out corpus. Table 5.4 shows these measures produced by RIPPER. The results for both algorithms are a good deal lower than the performance measures for the CGN data. When using all features, the results are slightly above the baseline. The results for every single feature are equal to or below the baseline. Unlike for CGN data, for MBL using only the noun feature results in lower performance measures than using all features. For RIPPER the noun feature and the preposition feature perform better than the combination of all features.

Table 5.3: *Performance measures in percentages on* ARG *versus* COND *prediction in newspaper and e-mail material (61 instances) by* MBL.

MBL		ARG			COND		
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
all	77	79	96	86	60	20	30
P	75	75	100	86	-	0	-
N	75	77	96	85	50	13	21
V	75	76	98	86	50	7	12
cooc	74	76	96	85	33	7	11
baseline	75	75	100	86	-	0	-

This considerable difference in performance between CGN data and newspaper and e-mail data, can be a consequence of the difference in type of data, which is spoken data versus written data. Moreover, the small number of instances in the newspaper and e-mail data deliver a rather poor results table, so consequently we can only draw some minor conclusions.

Table 5.4: *Performance measures in percentages on* ARG *versus* COND *prediction in newspaper and e-mail material (61 instances) by* RIPPER.

RIPPER		ARG			COND		
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
all	75	76	98	86	50	7	12
P	80	79	100	88	100	20	33
N	77	76	98	87	67	13	22
V	75	75	100	86	0	0	0
cooc	75	75	100	86	0	0	0
baseline	75	75	100	86	-	0	-

5.3.5 Contribution to accentuation

In this final experiment we assessed the added value of using information about argument versus condition in PROS-3 (Dirksen, 1994). We compared the accentuation of the sentence final verb phrase according to PROS-3 and according to PROS-3 complemented with information about the identity of the preceding nominal constituent, to the reference transcription (mentioned in section 4.2). We did this for a subset (38 phrases; 27 with an argument and 11 with a condition) of the held-out corpus, because the reference transcription was not available for all 61 phrases of the newspaper and e-mail data.

Table 5.5 shows the performance measures for this comparison, indicating that the accentuation of the sentence final verb is slightly improved when using information about the status of the nominal constituent (ARG vs. COND). The results for PROS-3 complemented with this information derived with MBL are better than PROS-3 solely. This improvement is mainly in the precision. The results for PROS-3 complemented with the information derived with RIPPER is only better in precision. Indeed, blocking certain incorrect placements of accents improves the precision on accentuation.

Table 5.5: *Performance in percentages on accentuation of the sentence final verb by PROS-3 (for 38 instances), complemented with information about argument versus condition from* MBL *and* RIPPER *for all features, and with a "golden standard".*

	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
PROS-3	80	63	84	72
PROS-3 + MBL (all)	81	64	83	72
PROS-3 + RIPPER (all)	80	64	81	72
PROS-3 + golden standard	81	65	85	74

MBL attains the best improvement. Although it incorrectly prevents three *intended* accents (when compared to the classification mentioned in section 4.2), it does in fact

correctly prevent *unintended* accents in six other cases. Two instances of the latter are given in example 5.34, where - indicates the prevented accent.

(5.34) (a) ... *vacatures kunnen -invullen.*
... vacancies can fill out.

(b) ... *een discussie -geboren.*
... a discussion arisen.

MBL incorrectly inserts two accents, where there was no accent *intended* (when compared to the classification mentioned in section 4.2), while it does also correctly insert *intended* accents in two other cases. These two latter instances are given in example 5.35, where + indicates the location of the inserted accent.

(5.35) (a) ... *twintig minuten +gebrand.*
... twenty minutes burned.

(b) ... *twee uur +stadten.*
... two hours shopping.

Table 5.5 also shows the performance measures for PROS-3 complemented with the "golden standard". These results indicate the maximal attainable improvement of accentuation when using information whether the nominal constituent preceding the sentence final verb is an argument or a condition. The results that we obtained for PROS-3 complemented with the information from MBL and RIPPER is worse than when complemented with the "golden standard".

5.4 Discussion and conclusion

We discussed the criteria for a nominal constituent preceding a sentence final verb phrase to be an argument or a condition. This classification is a factor in the accentuation status of the verb (based on SAAR (Gussenhoven, 1982, 1984)). We discussed that overall there are two rules that apply to this: (i) the sentence final verb is not accented if it is preceded by an argument (including locatives that are subcategorized by the verb), and (ii) the sentence final verb is accented if it is preceded by a condition.

The whole experimental design can be seen as being somewhat cyclic, because we start with defining what we consider as arguments and conditions, whereafter we investigate to what extent we can predict these classifications through performing machine learning experiments. However, we think we sufficiently argued why we classify certain instances of nominal constituents to be arguments or conditions, and we consistently use the specified classification rules. Therefore, we consider the experimental design to be legitimate.

The results of testing on the Spoken Dutch Corpus (CGN) data showed that machine learning experiments using lexical features and a co-occurrence feature (computed from WWW counts) are useful for the prediction of the status of the nominal constituent which precedes a sentence final verb phrase. The results on the held-out newspaper and e-mail data also showed that this method is useful, however the results are far less successful than for CGN data. This can be due to the small number of instances in the held-out corpus (only 38 sentences), and to the fact that we trained on spoken data and tested on written data (although we did not find such an effect for the experiments on PP attachment prediction). Besides the deviations (i.e. marked order of constituents and poor predictability of the constituent to the verb) for accentuation status of the verb, discussed in section 5.2, might also have been of influence.

The main conclusion from the machine learning experiments is that for disambiguation of arguments and conditions, the noun feature is the most important. The status of the nominal constituent preceding the sentence final verb can be predicted on the basis of the identity of the noun solely. It would also be interesting to perform machine learning experiments for directly predicting the accentuation status of the sentence final verb. From the discussion about the validity of SAAR (section 5.2.2) we might expect that for directly predicting the accentuation status, the verb feature, instead of the noun feature, would be the most important.

Evaluation of the new prosody module ECLIPSE

6

In this chapter we describe the evaluation of the new prosody module (ECLIPSE) that resulted from studies described in the previous chapters. The evaluation is dual, consisting of a objective evaluation through comparison with the reference transcription and a subjective evaluation by means of a perception experiment in which listeners had to indicate the acceptability of the different realizations of the same sentence. The results of the objective evaluation show that ECLIPSE performs considerably better than PROS-3. The results of the subjective evaluation show that ECLIPSE is preferred by the listeners and the experts over PROS-3 and that there is no significant difference between ECLIPSE and the reference transcription.

6.1 Introduction

The research described in the previous chapters was used to create a new prosody module (ECLIPSE). This module uses syntactic and lexical information for the allocation of phrase boundaries and accents. In this chapter we describe the investigation of the merit of using the revised algorithm for prosodic phrasing, using information about PP attachment and using information about the status of the nominal constituent that precedes a sentence final verb. First, we conduct a objective evaluation comparing the output of the prosody module with the reference transcription and the previously evaluated algorithm PROS-3. Next, we perform a perception experiment in which listeners have to indicate the acceptability of the prosodic structure (i.e. subjective evaluation). Finally, as a cross-check we compare the results from the perception experiment to quality judgements from three experts.

We hypothesize that ECLIPSE assigns a better and more acceptable prosodic structure than PROS-3. To test this hypothesis we need to make a fair comparison between the two algorithms. PROS-3 operates in tandem with a robust syntactic parser. This syntactic analysis is often incorrect. ECLIPSE uses syntactic information based on the Amazon[1] parser, which delivers a more correct syntactic analysis (which is considered as state-of-the-art). Comparing the performance of ECLIPSE with that of PROS-3 as such will give a distorted image of the differences. Therefore, we will also compare the performance of ECLIPSE to a version of PROS-3 that is based on syntactic information provided by Amazon. Henceforth, we will refer to this version of PROS-3 as PROS-3+.

6.2 Objective evaluation

We evaluate the prosody module ECLIPSE by comparing the output with the reference transcription of human experts, the output of PROS-3 (the prosody assigning algorithm we started from) and PROS-3+. For this, we compare the allocation by ECLIPSE of the different types of phrase boundaries (weak, medium, strong and no boundary) and the allocation of accents (+/- accent) with that by the reference transcription, PROS-3 and PROS-3+. The test material (24 sentences (see Appendix C)) contains the factors which turned out to be the major cause of errors in assigning accents and phrase boundaries, as was indicated by the error analysis described in Chapter 2. Apart from contextual effects (which we don't deal with in this project), these factors are noun attached PP, verb attached PP, long first major constituent, various punctuation marks, sentence final verb preceded by an argument and preceded by a condition. The 24 sentences are selected from the newspaper articles and e-mail messages that we used for computation of the reference transcription (see Chapter 2). Table 6.1 shows an example of prosody assignment according to the reference transcription, PROS-3, PROS-3+ and ECLIPSE for sentence 6.1.

[1] Amazon is a syntactic parser developed at Nijmegen University. Its homepage is: `http://lands.let.kun.nl/amazon/`.

(6.1) *Hoezeer er ook een verbod geldt op het lekken*
However much there also a ban is on leaking information
uit de ministerraad, toch is het beraad niet supergeheim ///
from the cabinet meeting, still is the meating not topsecret.
"However much there is a ban on leaking information from the cabinet meeting, the meeting is not topsecret."

Table 6.1: *Allocation of accents and phrase boundaries according to the reference transcription, PROS-3, PROS-3+ and ECLIPSE. An asterisk indicates that the word is accented. Slashes indicate that a boundary is allocated succeeding the word. The number of slashes indicates the boundary strength.*

word	**reference**	**PROS-3**	**PROS-3+**	**ECLIPSE**
hoezeer	*		/	
er				
ook				*
een			/	
verbod	*	* /	*	*
geldt			//	
op		*		
het				
lekken	*		* /	*
uit				
de				
ministerraad	* //	* //	* //	* //
toch	*	* /	* /	*
is				
het				
beraad		* /	* /	*
niet	*	* /	* /	*
supergeheim	* ///	* ///	* ///	* ///

These examples show that PROS-3 assigns too many boundaries and allocates boundaries at incorrect locations. The same is true for accentuation. PROS-3+ also assigns too many boundaries, but it does not allocate boundaries within syntactic constituents. The allocation of accents by PROS-3+ is somewhat different from that by the reference transcription, but it is better than that by PROS-3. If ECLIPSE is compared to the reference transcription we see that there is no discrepancy in boundary allocation, and that accent assignment is slightly different.

6.2.1 Phrasing

We counted the number of phrase boundaries assigned to the 24 sentences mentioned above (see Figure 6.1). These numbers give a first impression of the performance of ECLIPSE for phrase boundary allocation compared to the reference transcription,

PROS-3 and PROS-3+. From now on we consider PROS-3 and PROS-3+ as baselines to which we compare the performance of ECLIPSE.

We see that the number of weak boundaries for ECLIPSE is smaller than that of PROS-3 and PROS-3+, but larger than that of the reference transcription. For medium boundaries the same is true, however with smaller differences. The numbers of strong boundaries are almost equal. In total, both PROS-3 and PROS-3+ assign considerably more phrase boundaries than the reference transcription. ECLIPSE also assigns more boundaries, but this difference is far less substantial.

For more detailed results we computed the performance measures accuracy, precision, recall and F_β-value. Since a bimodal value is necessary for the computation of these performance measures we computed the number of incorrect insertions and omissions according to the two methods described in Chapter 2. Method 1 abstracts away from boundary strength (it only makes the distinction between boundary and no boundary). Method 2 does consider boundary strength.

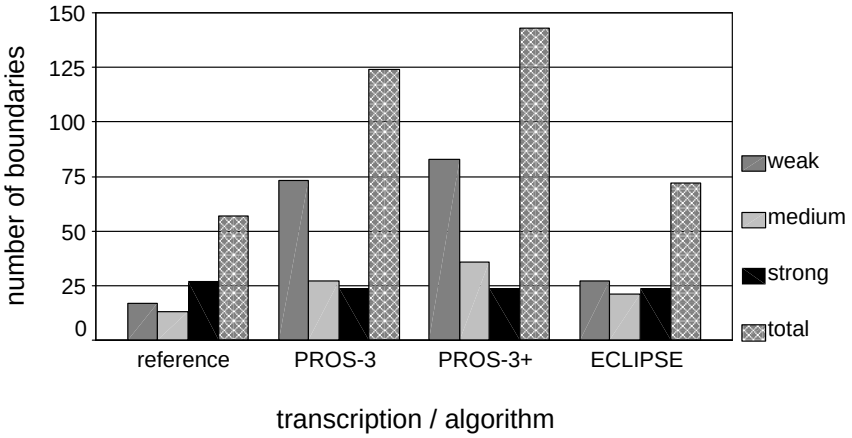


Figure 6.1: *Number of boundaries per type for the reference transcription, PROS-3, PROS-3+ and ECLIPSE.*

Table 6.2: *Performance measures in percentages for allocation of phrase boundaries, for PROS-3, PROS-3+ and ECLIPSE compared to the reference transcription.*

	method 1				**method 2**			
	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$	**accuracy**	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
PROS-3	79	38	82	52	77	30	71	42
PROS-3+	79	39	98	56	77	33	92	49
ECLIPSE	95	75	95	84	93	68	87	76

Table 6.2 shows that for all algorithms the performance for allocation of phrase boundaries according to Method 1 is better than according to Method 2. This could be expected, because Method 2 is more stringent. Although the recall is slightly lower for ECLIPSE than for PROS-3+, overall the performance measures for ECLIPSE according to both methods are considerably better than those for PROS-3 and PROS-3+. The large

improvement in precision is far more important than the slight decay in recall, since the results in Chapter 3 showed that incorrect boundary insertion (which ratio is indicated by precision) is more disturbing to the listener than incorrect boundary omission (which ratio is indicated by recall).

6.2.2 Accentuation

We counted the number of accents that are assigned to the 24 sentences. These numbers give a first impression of the performance of ECLIPSE for accent assignment compared to the reference transcription, PROS-3 and PROS-3+.

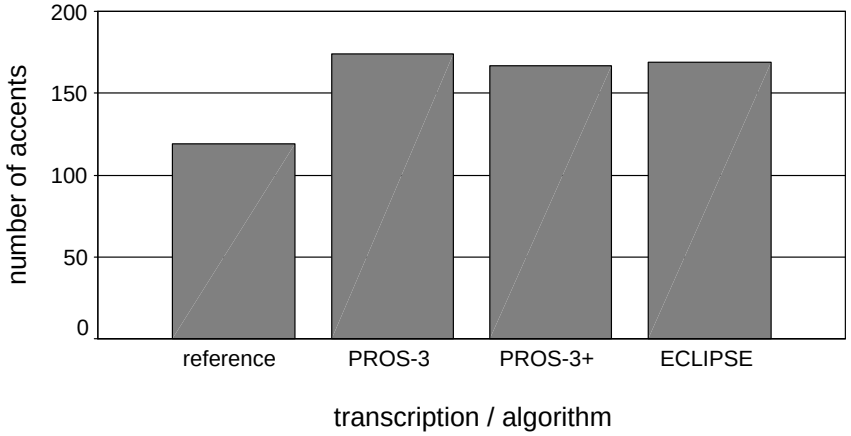


Figure 6.2: *Number of accents for the reference transcription, PROS-3, PROS-3+ and ECLIPSE.*

Figure 6.2 shows that the number of accents for ECLIPSE is almost equal to that of PROS-3 and PROS-3+ and considerably larger than that of the reference transcription. The performance measures given in Table 6.3 show that for allocation of accents the performance measures for ECLIPSE are better than those for PROS-3 and PROS-3+.

Table 6.3: *Performance measures for allocation of accents for PROS-3, PROS-3+ and ECLIPSE compared to the reference transcription.*

	accuracy	**precision**	**recall**	$\mathbf{F}_{\beta=1}$
PROS-3	77	56	82	67
PROS-3+	81	62	87	72
ECLIPSE	84	65	92	76

6.3 Subjective evaluation

As we did for the evaluation of the three Dutch TTS systems and PROS-3 under various conditions (see Chapter 2) we performed a perception experiment to assess the acceptability of the prosodic structure as assigned by ECLIPSE. The perception experiment presented here puts the results of the comparison between the assignment of

prosodic structure by ECLIPSE, the reference transcription and PROS-3+ to the test. We left PROS-3 out of this evaluation study, because it induces many prosodic errors due to improper syntactic information provided by the robust syntactic parser instead of the Amazon parser. We consider these errors not relevant for this evaluation (as mentioned in section 6.1). The evaluation was not split into an evaluation of the assignment of prosodic phrase boundaries and accents, since we believe that accentuation and prosodic phrasing can only be completely appreciated when the two factors are provided together. Moreover, analytic listening is hard for naive listeners (they are not used to listening to only one property of intonation and abstain from others).

By means of this perception experiment we put three hypotheses to the test, concerning the acceptability of the assigned prosodic structure. Our first hypothesis is that **listeners prefer the reference transcription over PROS-3+**, since in section 6.2 we found that the performance measures for both phrasing and accentuation are not satisfactory. Our second hypothesis is that **listeners prefer ECLIPSE over PROS-3+**, since we found that the performance measures for both phrasing and accentuation for ECLIPSE are considerably higher than for PROS-3+. Our third hypothesis is that **the difference between ECLIPSE and the reference transcription is smaller than the difference between PROS-3+ and the reference transcription**, because in the objective evaluation we found that the performance measures for ECLIPSE are rather good (when compared to the reference transcription).

6.3.1 Method

Experimental Design

The 24 sentences mentioned in section 6.2 were processed by the female voice of Calipso Text-to-Speech synthesis (which we also used for the experiment described in Chapter 3). Grapheme input was processed by this system, resulting in a phoneme representation, which was corrected manually. The prosodic structures resulting from ECLIPSE, the reference transcription and PROS-3+ were assigned. Thus, for each sentence three spoken versions were generated[2].

Accents were realized as so called pointed hats, consisting of an accent lending rise followed by an accent lending fall, where both rise and fall are realized on the same syllable. Four types of phrase boundaries were realized. No boundary, a weak boundary realized by a continued high pitch and pre-pausal lengthening, a medium boundary realized by a continuation rise followed by a 350ms pause and a strong boundary realized by a final pitch fall followed by a 500ms pause.

The utterances were presented pairwise to 20 subjects. Each pair consisted of two realizations of the same sentence, with two of the three prosodic structures. All three

[2]The realizations of these sentences can be obtained from `http://www.ipo.tue.nl/homepages/ovherwij/evaluation-eclipse.html`.

possible combinations were presented in both orders. This resulted in $24 * 3 * 2 = 144$ sentence pairs.

We distributed the sentence pairs over two subject groups according to a Latin square design. This means that group I was presented with the 24 sentence pairs, with all three possible combinations of two prosodic structures, but in only one order (realization X – realization Y), while group II was presented with the same 24 sentence pairs, with all three possible combinations of two prosodic structures, in the opposite order (realization Y – realization X). This resulted in 72 sentence pairs per subject. The sentence pairs were presented in random order.

The experiment was conducted in a sound treated room. The stimuli were presented over head phones, while at the same time the text of the sentence was displayed on a screen. The silence between the two utterances of a pair was 400ms. Each utterance pair was presented twice. After the second presentation, subjects were asked to indicate on a 7-point scale which utterance of the pair they preferred and to what extent. Subjects had to indicate their judgement by clicking with the mouse on a button on the screen. These buttons indicated the preferences for the first sentence or the second sentence. The scale ranged from -3 to +3, where -3 indicated a strong preference for the first utterance, 0 no preference for either utterance, and +3 a strong preference for the second utterance. The 7-point scale is the same as that used in the experiments on perceptual costs of errors (see Chapter 3).

Prior to the actual experiment there was a training phase where the subjects could get acquainted with the procedure. In this training phase subjects were presented with 6 utterance pairs that were not part of the actual experiment. After 36 sentence pairs subjects had a short break. The total duration of the experiment was about 1 hour.

All subjects were native speakers of Dutch and none of them reported hearing problems. They were all students in the age of 19 through 30 and they were not familiar with the research described in this thesis.

Statistical Design

As we did for the perception experiment on the perceptual costs of errors, we submitted the data to an analysis of variance for paired comparisons (Scheffé, 1952). This method is developed for experiments in which preferences are expressed on a scale of 7 points or more.

6.3.2 Results

The preference scores for assignment of prosodic structure, averaged over all subjects and both utterance orders, are given in Table 6.4. These scores indicate that there is a preference for the prosodic structure assigned by the reference transcription and ECLIPSE when compared to PROS-3+ (comparison A–B and B–C). The score for com-

parison C–B indicates that there is no substantial difference between ECLIPSE and the reference transcription. These results are also visualized in Figure 6.3, where a score of -3 indicates a maximal preference for the algorithm mentioned first in the comparison, and a score of +3 indicates a maximal preference for the second mentioned algorithm. The significance of the results will be discussed in following sections.

Table 6.4: *Mean preference scores for assignment of prosodic structure resulting from the comparison of utterances realized with the prosodic structures as assigned by PROS-3+ (A), the reference transcription (B) and ECLIPSE (C). The plus sign indicates a preference for the second utterance in the comparison.*

	comparison			**preference score**
A – B	PROS-3+	vs.	reference	+0.42
A – C	PROS-3+	vs.	ECLIPSE	+0.40
C – B	ECLIPSE	vs.	reference	+0.07

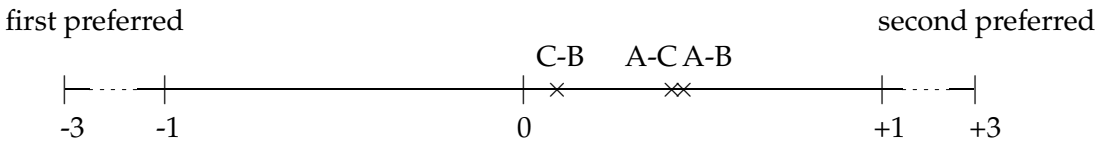


Figure 6.3: *Schematic representation of the preference scores for the comparisons given in Table 6.4. The negative score represents a preference for the realization mentioned first in the comparison, whereas the positive score represents a preference for the realization mentioned second in the comparison.*

Hypothesis 1: Listeners prefer the reference transcription over PROS-3+.

For testing this hypothesis we used the judgements for comparison A–B in Table 6.4. Judgements were summed over all subjects and all sentences. Analysis of variance for paired comparison shows that the difference between the two versions ($\alpha_1 - \alpha_2$) is 0.44, with a yardstick (Y) of 0.13. So, $\alpha_1 - \alpha_2 > Y$ is true, indicating that the difference between α_1 and α_2 is significant ($p < 0.05$) for this comparison. These results confirm our hypothesis that listeners prefer the utterance with the prosodic structure assigned by the reference transcription.

Figure 6.4 shows the preference scores per sentence length, for the comparison of PROS-3+ with the reference transcription. Sentence length does not exert a significant influence on the preference for the reference transcription ($r = -0.26$, $p < 0.05$).

One sentence (given in example 6.2), shows a substantial preference for PROS-3+. This sentence contains a long first major constituent. PROS-3+ allocates weak boundaries within the first major constituent (as in example 6.2a), while in general we assume that phrase boundaries should not be allocated before the end of this constituent (as in example 6.2b). In this specific sentence listeners do seem to prefer phrase boundaries within the first major constituent.

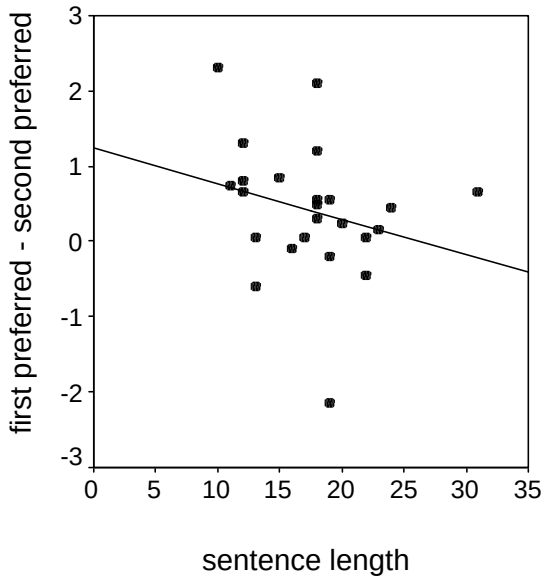


Figure 6.4: *Preference scores per sentence for comparison of PROS-3+ with the reference transcription. A negative score indicates a preference for PROS-3+, a positive score indicates a preference for the reference transcription. Sentences length is expressed in number of words.*

(6.2) (a) Vanaf de *invoering / van het *beginsel / van de *openbaarheid / van *bestuur / in *1980 // is dit type *voorstellen *regelmatig gedaan ///

(b) Vanaf de *invoering van het *beginsel van de openbaarheid van *bestuur in *1980 // is dit type voorstellen *regelmatig gedaan ///

"*As from the introduction of the principle of publicity of government in 1980, this type of propositions has been made regularly.*"

Hypothesis 2: Listeners prefer ECLIPSE over PROS-3+.

For testing this hypothesis we used the judgements of comparison A–C in Table 6.4. Judgements were summed over all subjects and all sentences. Analysis of variance for paired comparison shows that the difference between the two versions ($\alpha_1 - \alpha_2$) is 0.38, with a yardstick (Y) of 0.13. So, $\alpha_1 - \alpha_2 > Y$ is true, indicating that the difference between α_1 and α_2 is significant ($p < 0.05$) for this comparison. These results confirm our hypothesis that listeners prefer the utterance with the prosodic structure as assigned by ECLIPSE.

Figure 6.5 shows the preference scores per sentence length, for the comparison of PROS-3+ and ECLIPSE. For most sentences there is a clear preference for ECLIPSE. However, for one sentence (i.e. the same as discussed under hypothesis 1) there is again a clear preference for PROS-3+. Overall the results show that the preference for

ECLIPSE is larger for the shorter sentences. The correlation ($r = -0.39$) is significant at the 0.05 level.

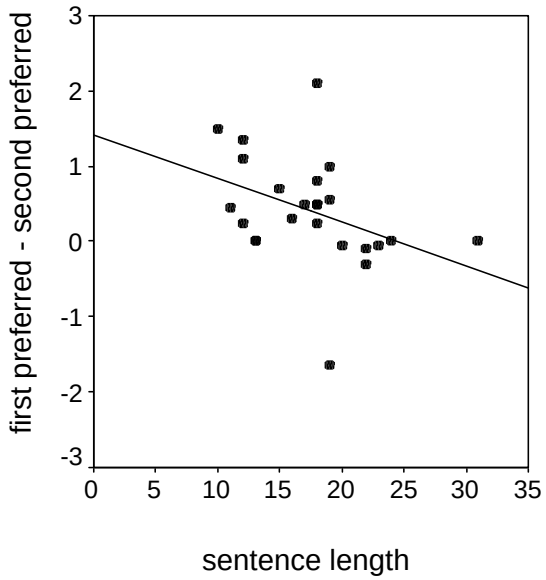


Figure 6.5: *Preference scores per sentence for comparison of PROS-3+ with ECLIPSE. A negative score indicates a preference for PROS-3+, a positive score indicates a preference for ECLIPSE. Sentences length is expressed in number of words.*

Hypothesis 3: The difference between ECLIPSE and the reference transcription is smaller than the difference between PROS-3+ and the reference transcription.

For testing this hypothesis we first performed an analysis of variance for paired comparisons on the judgements of comparison C–B in Table 6.4. Judgements were summed over all subjects and all sentences. Results show that the difference between the two versions ($\alpha_1 - \alpha_2$) is 0.05, with a yardstick (Y) of 0.13. This means that the difference between α_1 and α_2 is not significant ($p < 0.05$) for this comparison. These results show that there exists no difference in preference between ECLIPSE and the reference transcription, whereas the preference for the reference transcription is significant when compared to PROS-3+. From this we conclude that our hypothesis is correct.

Figure 6.6 shows the preference scores per sentence length, for the comparison of the reference transcription and ECLIPSE. For about half of the sentences there is a slight (but not substantial) preference for the reference transcription, whereas for the other half of the sentences there is a slight preference for ECLIPSE. There exists no significant effect of sentence length ($r = -0.11$).

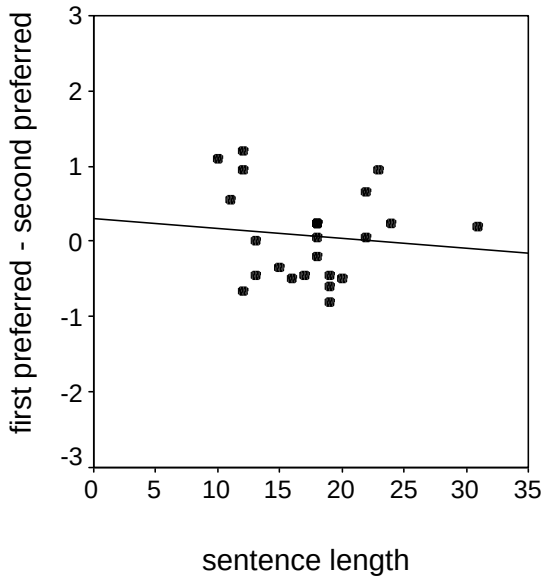


Figure 6.6: *Preference scores per sentence for comparison of ECLIPSE with the reference transcription. A negative score indicates a preference for ECLIPSE, a positive score indicates a preference for the reference transcription. Sentences length is expressed in number of words.*

To investigate the exact preferences for the three algorithms (PROS-3+, reference transcription and ECLIPSE), we also applied Thurstone's one-dimensional scaling technique (Thurstone, 1927; Torgerson, 1967). In order to construct the Thurstone scales we had to transform our 7-point scale data into a binary scale. The negative scores were collapsed to "first-preferred", the positive scores were collapsed to "second-preferred". The zero-score was equally divided over the two preference classes. Figure 6.7 is a schematic representation of the exact preferences for the three algorithms.

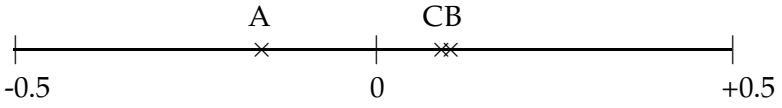


Figure 6.7: *Schematic representation of the preference scores for the algorithms PROS-3+, reference transcription and ECLIPSE.*

This scale shows that the reference transcription (B $= 0.09$) and ECLIPSE (C $= 0.08$) are preferred over PROS-3+ (A $= -0.17$) and that there exists only a slight preference for the reference transcription over ECLIPSE. This result is in accordance with the above mentioned Scheffé scales.

6.3.3 Correlation with expert judgements

In the previous section we learned that listeners prefer ECLIPSE and the reference transcription over PROS-3+, and that there is no preference for either ECLIPSE or the reference transcription. As a cross-check for investigation of the quality of the prosodic structure assigned by ECLIPSE, compared to that of PROS-3+ and the reference transcription, we collected quality judgements from three experts.

We asked the experts not only to judge the overall prosodic structure (accentuation and prosodic phrasing together), but also to judge the accentuation and prosodic phrasing separately. Judgements were given on a 4-point scale (where 3 is good and 0 is bad). This kind of judgements could not be obtained from the naive participants in the perception experiment, since analytic listening to separate factors (i.e. accentuation and phrasing) when perceiving overall prosodic structure, is a very difficult task for non-experts.

The average expert scores (see Table 6.5) indicate that the experts prefer the overall prosodic structure, accentuation and phrasing as assigned by ECLIPSE and the reference transcription over that by PROS-3+. Moreover, the scores for ECLIPSE and the reference transcription are comparable, meaning that the experts have no preference for either of these two algorithms. These results correspond to the results of the perception experiment described in section 6.3.2.

Table 6.5: *Average expert scores for the overall prosodic structure, accentuation and phrasing as assigned by PROS-3+, the reference transcription and ECLIPSE. (0 =bad, 3 =good)*

		expert score		
	algorithm	**overall**	**accent**	**bound**
A	PROS-3+	0.97	1.65	1.07
B	reference	2.01	2.03	2.22
C	ECLIPSE	1.99	1.96	2.32

On the basis of the expert scores we investigated whether accentuation or phrasing is the major contributing factor with respect to the quality of the overall prosodic structure. The results in Table 6.5 imply that the score for overall prosodic structure is dependent on the score of the 'weakest link'. If the score for accentuation is lower (as for ECLIPSE and the reference transcription), the overall score equals this score. If the score for phrasing is lower (as for PROS-3+) the overall score equals this score.

A multiple linear regression analysis (Rietveld and van Hout, 1993) shows that the variance in expert score for overall prosodic structure can be significantly explained by the expert scores for both phrasing and accentuation ($r = 0.88$, $p < 0.01$), where the score for phrasing weights slightly more than the score for accentuation. Thus, the quality of the overall prosodic structure depends on the quality of both accentuation and prosodic phrasing.

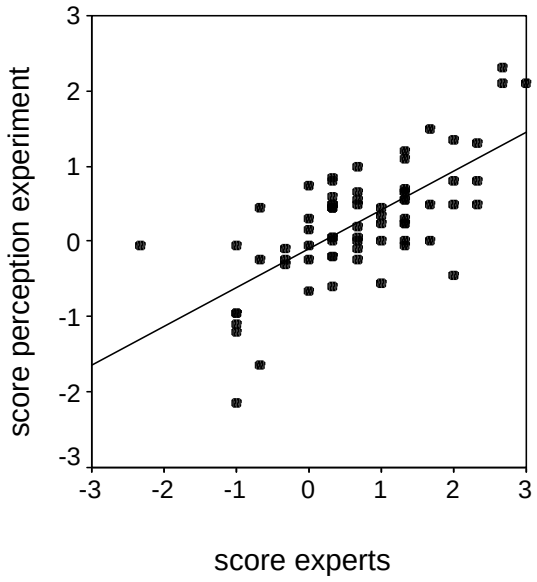


Figure 6.8: *Comparison of scores from preference scores from perception experiment and experts' difference scores for overall prosodic structure ($r = 0.69$).*

From the expert judgements we computed difference scores for every sentence in the three combinations of two versions (PROS-3+ vs. reference, PROS-3+ vs. ECLIPSE and ECLIPSE vs. reference). These difference scores for overall prosodic structure, accentuation and phrasing are compared to the preference scores from the perception experiment (see Figure 6.8). The correlation ($r = 0.69$) is significant at the 0.01 level. This means that the experts' difference scores on overall prosodic structure for the three combinations of two versions correspond with the preference scores from the perception experiment.

From this we conclude that both the perception experiment and the expert judgements indicate that the quality of the prosodic structure assigned by ECLIPSE and the reference transcription is higher than that of the prosodic structure assigned by PROS-3+.

6.4 Discussion and conclusion

From the objective evaluation we learned that there still is a difference in assignment of prosodic structure (i.e. phrase boundaries and accents) by a Text-to-Speech system and by human experts. However, the performance measures for ECLIPSE are rather good. The objective evaluation was performed on a small number of sentences that were selected from newspaper articles and e-mail messages. One might expect a different performance on a complete text. Since we performed the evaluation on sentences that contain the factors which turned out to be the major cause of errors in predicting

prosodic structure, we expect that ECLIPSE performs at least equally good on complete texts which contain non-problematic sentences next to the problematic ones we evaluated here.

From the subjective evaluation by means of a perception experiment, we learned that listeners prefer the new prosody module ECLIPSE over the older PROS-3, and that they find the prosodic structure assigned by ECLIPSE as acceptable as the reference transcription by human experts.

For the three categories (overall prosodic structure, accentuation and phrasing) there is also a substantial correspondence between the listeners' preferences and the experts' difference scores for the three combinations of two versions of prosodic structure. Experts also prefer ECLIPSE and the reference transcription over PROS-3+, and there is again no preference for either one of these two algorithms.

The reference transcription has been derived from annotations of complete newspaper texts and e-mail messages. If this transcription had been derived from annotations of sentences in isolation it would have contained somewhat more accents. In the current transcription some accents were omitted on the basis of contextual information. For 9 of the 24 sentences that we used for the subjective evaluation there was no effect of context. We computed the preference scores for assignment of prosodic structure for these 9 sentences, averaged over all subjects and both utterance orders. The results are given in Table 6.6.

Table 6.6: *Mean preference scores for assignment of prosodic structure resulting from the comparison of utterances realized with the prosodic structures as assigned by PROS-3+ (A), the reference transcription (B) and ECLIPSE (C). The plus sign indicates a preference for the second utterance in the comparison.*

	comparison	**preference score 24 sentences**	**preference score 9 sentences**
A – B	PROS-3+ vs. reference	+0.42	+0.61
A – C	PROS-3+ vs. ECLIPSE	+0.40	+0.82
C – B	ECLIPSE vs. reference	+0.07	-0.04

Scheffé's analysis of variance for paired comparisons showed the same results for the 9 sentences as for the 24 sentences. Listeners prefer the reference transcription over PROS-3+ ($\alpha_1 - \alpha_2 = 0.66$, $Y = 0.21$), they prefer ECLIPSE over PROS-3+ ($\alpha_1 - \alpha_2 = 0.77$, $Y = 0.21$), and there is no preference for either ECLIPSE or the reference transcription ($\alpha_1 - \alpha_2 = 0.10$, $Y = 0.21$). These results are only indicative since they are based on a very small number of sentences.

The results from the evaluation studies show that the revised phrasing algorithm in combination with information about PP attachment (noun or verb attachment) and information about the identity of the constituent (argument or condition) preceding a sentence final verb phrase, induces a considerable improvement in the automatic assignment of prosodic structure. This means that ECLIPSE delivers an acceptable prosodic structure for synthetic speech.

7 General discussion

The module for the assignment of prosodic structure (ECLIPSE) performs considerably better on accentuation and prosodic phrasing than existing Text-to-Speech systems for Dutch. Listeners and experts judge ECLIPSE as acceptable as the reference transcription of human experts. We therefore conclude that applying machine learning techniques, constrained by information about the perceptual costs of errors, is useful to obtain elaborate syntactic information. From this elaborate syntactic structure, in combination with lexical information, a perceptually appropriate prosodic structure can be computed for synthetic speech.

7.1 Recapitulation

In this thesis we have explored the use of correct syntactic and lexical information for improving the assignment of prosodic structure in synthetic speech, applying language engineering techniques that take into account psycholinguistic insights obtained through perception experiments. From an error analysis we learned that a major part of phrasing and accentuation errors is due to incorrect or insufficient syntactic information. We therefore started with investigating the importance of using proper syntactic information. Since this led to substantial improvement of the predicted prosodic structure, we adopted a language engineering approach to obtain more elaborate and correct information about syntactic relations (i.e. information that is not given by a state-of-the-art parser).

In automatic prosodic structure assignment there will always remain some errors. By means of perception experiments we assessed the perceptual costs of the different kinds of errors. We found that there exists a trade-off between allocation of too many prosodic phrase boundaries and too few. Listeners are more tolerant towards incorrect phrase boundary omissions than towards incorrect boundary insertions. Therefore, if forced to make a choice one should rather 'undergenerate' than 'overgenerate' prosodic phrase boundaries. For accentuation listeners are equally tolerant towards incorrect omissions and insertions. Therefore, it is important to allocate neither too many nor too few accents.

Taking these findings into consideration, we performed machine learning experiments for addressing two major aspects of assigning prosodic structure which require elaborate syntactic information: (i) resolving PP attachment ambiguities for correct allocation of phrase boundaries, and (ii) resolving argument – condition ambiguities for correct accentuation of sentence final verb phrases. Machine learning techniques appeared to be useful for deriving more elaborate syntactic information, and this information turned out to be valuable for the allocation of phrase boundaries and accents. Furthermore, we constructed an algorithm for prosodic phrasing that allocates a phrase boundary if the sentence length exceeds a previously established number of words. The precise location of the boundary is determined on the basis of the elaborate syntactic structure. This phrasing algorithm, which is embedded in the resulting prosody module ECLIPSE, performs considerably better in boundary allocation than the evaluated Text-to-Speech systems. As a side-effect of improved prosodic phrasing the allocation of accents also improved since accentuation is performed after the appointment of phrase domains, where accents are often allocated on the final content words in the phrase domains.

When comparing the output of ECLIPSE with that of the reference transcription by experts we obtain F_β-values of 84% for phrasing and of 76% for accentuation, which are substantially higher than those for the state-of-the-art TTS systems (being about 55% and 68% respectively). We especially gained a lot in the allocation of prosodic phrase

boundaries, on which we spent most effort through the formulation of a new phrasing algorithm and the resolution of PP attachment ambiguities through machine learning experiments. The allocation of sentence accents has also improved, however not as much as phrasing. Accentuation improved partly as a side-effect of correct prosodic phrasing, and through the prediction of the status of the nominal constituent preceding sentence final verb phrases.

Both untrained listeners and experts judged the overall prosodic structure assigned by ECLIPSE and by the reference transcription as being equally acceptable. From both the objective and subjective results we conclude that the use of elaborate syntactic and lexical information, derived through application of statistical techniques, leads to a much more acceptable prosodic structure.

7.2 Methodology and its limitations

Language engineering

The results showed that machine learning techniques, which use lexical information in combination with a co-occurrence strength value obtained from WWW counts, are valuable for deriving elaborate syntactic information about PP attachment and the distinction between argument and condition. For the research described in this thesis these results mean that we found a way to upgrade the syntactic information which we obtain from syntactic parsers. The elaborate information makes it possible to improve prediction of prosodic structure on the basis of syntactic information. It remains to be explored whether the machine learning approach is also useful for deriving elaborate syntactic information about other phenomena that are problematic for accentuation and prosodic phrasing (such as identification of reporting clauses, compound verbs and enumerations).

Co-occurrence strength values are now derived from WWW counts. However, for real-time prediction of prosodic structure using the WWW to obtain co-occurrence strength values could be a rather laborious solution; it requires an online connection with the WWW, when processing text for speech synthesis. The use of large static corpora would then be more convenient, although this would probably lead to a larger number of word pairs that are not found in the corpus, especially for processing newspaper articles since this text genre often contains new words that will not be included in static corpora. Further investigation of the relation between the size and the characteristics of the corpus and the performance should indicate whether other corpora than the WWW might be more useful for computation of co-occurrence strength values.

When searching the WWW for the combination of two words, we used the NEAR function of Altavista (Altavista, 2002), and a function that restricts the search to two adjacent words in the specified order. Both functions do not account for punctuation

marks. As a search result they also report instances of two words that are separated by a period or semicolon. The instances in the text that is processed do not contain these punctuation marks. Therefore, we expect that a more appropriate search function would give more relevant information.

Another disadvantage of the NEAR function is that the maximal distance between the two specified words is 10. It might be more worthwhile to experiment with different maximal distances (for instance of 5 words). It appears likely that the words often are usually close together in the text that is processed.

Psycholinguistics

Information about the perceptual costs of different kinds of errors in synthetic speech made it possible to carry out machine learning experiments in such a way that there is a bias for predicting the syntactic relation that has the least negative prosodic consequences when erroneously classified. Elaboration of syntactic information according to this strategy may be expected to lead to an increase in intelligibility, and thus to improved listeners' judgements about the acceptability of the synthesized utterances.

We investigated the perceptual costs of different kinds of phrasing and accentuation errors by means of pair-wise comparison of utterances with correct and incorrect phrasing and accentuation. We took the inverse of the strength of the preferences as an indicator for the perceptual costs of errors. The results showed that listeners are more tolerant towards incorrect phrase boundary omissions than towards incorrect boundary insertions, and they are equally tolerant towards incorrect accent insertions and omissions. We thus assumed that the perceptual costs of incorrect phrase boundary insertions are higher than those of boundary omissions, and that the perceptual costs of incorrect accent insertions and omissions are equal. Cross-validation of these results, by means of different psycholinguistic experiments should indicate whether our assumptions are valid.

Evaluation methods

We performed both objective and subjective evaluations of the Dutch Text-to-Speech systems, PROS-3 and ECLIPSE. For the objective evaluation we computed a reference transcription of 10 experts, to which we compared the prosodic structures assigned by the systems. This transcription denotes a 'mean' prosodic structure of the annotation of the experts.

We validated the reference transcription through comparison with a spoken reference of 3 experts, through computation of expert agreement, and by means of expert judgements of the reference transcription. There is a considerable amount of agreement between the reference transcription and the spoken reference, and the expert agreement is reasonable. The experts judged the reference transcription as being fairly good. From these results we conclude that the reference transcription is a valid transcription.

However, the reference transcription is an 'average' transcription. If the prosodic structure assigned by a system deviates from the reference, we call it a phrasing or accentuation error. We do not consider the gravity of a specific error, while we may expect that deviations from the reference transcription will be more serious when all experts agree upon the preferred prosodic structure, than when six out of ten experts agree. Applying other methods for assessing errors might have given different results. Besides, for any sentence probably a number of acceptable prosodic structures are possible. The locations of accents and phrase boundaries are not unrelated within a prosodic structure. In a 'mean' prosodic structure (such as the reference transcription) this relation is not accounted for. An alternative of the reference transcription would be to collect the annotations of the experts, select those for which there is consensus about their acceptability, and evaluate the systems' prosodic structures through comparison with these annotations. Another possibility would be to use a spoken version of text by one speaker as a reference. Or a consensus transcription can be derived of a number of experts, by having them actually sit together and agree upon one transcription.

For the subjective evaluation we performed perception experiments where untrained listeners had to indicate the acceptability of the different prosodic structures by means of pairwise comparison. In the evaluation of ECLIPSE we also obtained expert scores on the actual acceptability of a systems prosodic structure. For ECLIPSE and the reference transcription this score was about 2 (on a scale from 0 to 3), whereas this score was about 1 for PROS-3+. There exists a significant correlation between the expert judgements and the preference scores from the perception experiment. Therefore, we conclude that the preference judgements of the untrained listener are indeed an indication of the actual acceptability of the prosodic structures assigned by the different systems and that pairwise comparison is a useful method for assessing acceptability of prosodic structure. From this approach we can only obtain an idea about the preference for one system over another. This does not necessarily mean that the prosody is indeed fully satisfactory.

The acceptability of synthesized utterances depends for a large part on the assigned prosodic structure. Furthermore, it depends on the physical implementation of accents and phrase boundaries. The autosegmental phonology describes prosodic structure in terms of tone segments, in contrast with the IPO-description of Dutch intonation ('t Hart et al., 1990), which describes prosodic structure in terms of pitch movements. The pitch rise or fall on a specific word emphasizes that this word is prominent in relation to the other words in (that part of) the utterance. We adopted the IPO-description since it well suits our research. When speech is automatically generated, the prosodic structure is realized as a sequence of pitch movements, instead of sudden alternations between high and low pitch level. In this thesis the aspect of physical realization of prosodic structure has been disregarded. However, the Text-to-Speech system that we used for several comparison studies has been evaluated earlier (Terken, 1993). This evaluation showed that for isolated utterances the naturalness of the phys-

ical implementation of the intonation was as good as the human intonation. Therefore, we presume that the physical realization of the prosodic structure did not have a considerable effect on the expert judgements.

7.3 Applications

For application of ECLIPSE in real-time speech synthesis there are some requirements. First, a syntactic parser should be available to obtain the syntactic structure of a sentence. This syntactic structure should at least provide information about the major constituents in the sentence and their head words. Second, a large text corpus (such as the WWW) should be available for online computation of co-occurrence strength values. A requirement for the use of WWW counts is that the algorithm and system for computation of co-occurrence strength values are fast enough. Otherwise, online computation should be replaced by off-line computation or by database storage. Furthermore, the phrasing algorithm based on sentence length should be implemented, considering speech rate to establish the maximum number of words between two phrase boundaries. After prosodic phrasing an algorithm for accentuation (for instance as implemented in PROS-3) needs to be employed which uses phrase domains and information about head words as input information for the assignment of accents to specific words.

Using these sources the prosodic structure can be predicted, which will serve as one line of input for a Text-to-Speech system. The other line of input then is the text of the sentence, whether or not converted into a phoneme string. This approach seems particularly useful for TTS systems that use a separate module for prosodic structure prediction, such as diphone synthesis. Systems that perform concatenative synthesis, such as straightforward unit selection systems, do not gain from (parts of) ECLIPSE. These systems search for the longest phoneme strings that match (parts of) the utterance, disregarding the prosodic structure with which the strings are stored in the database. Other concatenative systems that do consider the prosodic structure, such as Festival (Taylor et al., 1998) can, however, profit from the several parts of ECLIPSE. For example, the results of the current research on the perceptual costs of errors can be taken into account when searching for the closest match in the speech database. If an exact match of phonemes is found, but that string contains a phrase boundary where the predicted prosodic structure does not, the system should search for a different matching phoneme string, or even for smaller phoneme strings that match the prosodic structure, since the perceptual costs of phrase boundary insertions are relatively high. If the found phoneme string doesn't contain a phrase boundary where there is one in the predicted prosodic structure, there is no need for the system to search for a different matching string, since the perceptual costs of phrase boundary omissions are relatively low.

7.4 Future research

In this thesis we only considered isolated sentences. If we applied ECLIPSE to whole newspaper texts or e-mail messages, we would still obtain adequate prosodic structures when only considering the syntactic structure of the text. However, in whole texts context or discourse information should be taken into account to obtain a more natural prosodic structure. Beforehand, we already stated that discourse information is beyond the scope of this project. When processing multi-sentence utterances, appending information about context (such as given-new relations and contrast effects) should result in an even more meaningful prosodic structure. It remains to be explored whether such contextual or semantic information that plays a role in the assignment of prosodic structure can be derived through computation of co-occurrence strength values from large corpora.

We used machine learning techniques to obtain more elaborate syntactic information about PP attachment and the distinction argument – condition. Training and testing of the learners in these experiments were performed on very small data sets. Therefore, these results can be regarded as being nothing more than a rough indication of the general usefulness of resolving these syntactic ambiguities. However, we showed that with such a small data set the performance in the machine learning experiments was reasonably good. And we obtained better results in assignment of prosodic structure using the information derived from the machine learning experiments. To assess the relevance of the experiments described here, we could cross-validate the results through performing these experiments on larger data sets. We expect that performing the experiments with a larger data set will not lead to different results, it would probably improve the assignment of accents and phrase boundaries.

As training corpora we used subsets of the syntactic treebank of the Spoken Dutch Corpus (CGN). For part of the CGN prosodic annotations will become available in the near future. Performing the same experiments on this corpus of prosodic annotations would be an interesting extension of the experiments described in this thesis.

Since the results showed that machine learning techniques are useful to resolve two specific syntactic ambiguities, it would be interesting to apply this approach also for deriving information about other phenomena that are problematic for accentuation (such as identification of reporting clauses) and prosodic phrasing (such as identification of enumerations).

7.5 Conclusion

Syntactic information about PP attachment and the distinction argument – condition, that is necessary for predicting prosodic structure for synthetic speech, can be obtained by means of machine learning experiments. Applying machine learning techniques taking into account information about the perceptual costs of errors, further improves the quality of synthesized utterances.

Bibliography

Abney, S. (1991). *Principle-Based Parsing*, chapter Parsing by Chunks. Kluwer Academic Publishers.

Aha, D., Kibler, D., and Albert, M. (1991). Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Altavista (2002). Advanced search cheat sheet. http://www.altavista.com/. Page visited September 2002 and May 2003.

Baart, J. (1987). *Focus, Syntax and Accent Placement: Towards a rule system for the derivation of pitch accent patterns in Dutch as spoken by humans and machines*. PhD thesis, University of Leiden, The Netherlands.

Baart, J. (1989). Focus and accent in a Dutch Text-to-Speech system. In *Proceedings of the 4th EACL*, pages 111–114, Manchester.

Bachenko, J. and Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in english. *Computational Linguistics*, 16(3):155–170.

Bachenko, J., Fitzpatrick, E., and Wright, C. (1986). The contribution of parsing to prosodic phrasing in an experimental Text-to-Speech system. In *Proceedings of the Association of Computational Linguistics conference*, pages 145–155.

Beach, C. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, 30:644–663.

Bear, J. and Price, P. (1990). Prosody, syntax and parsing. In *Proceedings of the Association of Computational Linguistics conference*, pages 17–22.

Beckman, M. (1996). The parsing of prosody. *Language and Cognitive Processes*, 11:17–67.

Birch, S. and Clifton, C. (1995). Focus, accent and argument structure. *Language and Speech*, 38(4):365–391.

Bolinger, D. (1989). *Intonation and its Uses: Melody in Grammar and Discourse*. Edward Arnold, London.

Brill, E. and Resnik, P. (1994). A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the 15th annual conference on Computational Linguistics*.

Carletta, J. (1995). Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 30(11):1–6.

Chafe, W. (1974). Language and consciousness. *Language*, 50:111–133.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of NAACL'00*, pages 132–139.

Church, K. and Hanks, P. (1991). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

Cohen, W. (1995). Fast effective rule induction. In *Machine Learning: Proceedings of the Twelfth International Conference*, Lake Tahoe, California.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*.

Collins, M. and Brooks, J. (1995). Prepositional phrase attachment through a backed-off model. In *Proceedings of Third Workshop on Very Large Corpora*, Cambridge.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27.

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.

Daelemans, W., van den Bosch, A., and Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine Learning, Special issue on Natural Language Learning*, 34:11–43.

Daelemans, W., Zavrel, J., Berck, P., and Gillis, S. (1996). MBT: A Memory-Based Part of Speech Tagger-Generator. In *Proceedings the Fourth Workhop on Very Large Corpora*, pages 14–27.

Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. (2002). TiMBL: Tilburg Memory Based Learner, version 4.3, Reference Guide. Available from `http://ilk.kub.nl/downloads/pub/papers/ilk0210.ps.gz`.

van Dinther, R. (2003). *Perceptual aspects of voice-source parameters*. PhD thesis, Eindhoven University of Technology, The Netherlands.

Dirksen, A. (1992a). Accenting and deaccenting: a declarative approach. *Proceedings Coling '92*.

Dirksen, A. (1992b). STP. Syntax-to-Prosody Conversion for the Polyglot TTS. Internal Report Version 1.1, Eindhoven University of Technology.

Dirksen, A. (1994). PROS-3 Syntax and Prosody for Text-to-Speech. User Manual Version 2.1a, IPO, Eindhoven University of Technology.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Fach, M. (1999). A comparison between syntactic and prosodic phrasing. In *Proceedings Eurospeech '99*, volume 1, pages 527–530.

Ferguson, G. and Takane, Y. (1989). *Statistical Analysis in Psychology and Education*. Psychology Series. McGraw-Hill International Editions.

Franz, A. (1996). *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, volume 1040 of *Lecture Notes in Artifical Intelligence*, chapter Learning PP attachment from corpus statistics, pages 188–202. Springer-Verlag, New York.

Gee, J. and Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:411–458.

Gussenhoven, C. (1982). Van fokus naar zinsaksent: een regel voor de plaats van het zinsaksent in het Nederlands. Technical report, Instituut Engels-Amerikaans, Nijmegen.

Gussenhoven, C. (1984). *On the Grammar and Semantics of Sentence Accents*. PhD thesis, Nijmegen University, The Netherlands.

Gussenhoven, C. (1992). Sentence accents and argument structure. In Roca, I., editor, *Thematic Structure: its Role in Grammar*, pages 79–106. Berlin: Foris.

't Hart, J. and Cohen, A. (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics*, 1:309–327.

't Hart, J. and Collier, R. (1975). Integrating different levels of intonation analysis. *Journal of Phonetics*, 3:235–255.

't Hart, J., Collier, R., and Cohen, A. (1990). *A perceptual study of intonation*. Cambridge University Press.

van Herwijnen, O. and Terken, J. (2000). Evaluation of automatic assignment of prosodic structure by Dutch TTS-systems. Technical report, Technische Universiteit Eindhoven.

van Herwijnen, O. and Terken, J. (2001a). Do speakers realize the prosodic structure they say they do? In *Proceedings of Eurospeech 2001 Scandinavia*, volume 2, pages 959–962.

van Herwijnen, O. and Terken, J. (2001b). Evaluation of PROS-3 for the assignment of prosodic structure, compared to assignment by human experts. In *Proceedings of Eurospeech 2001 Scandinavia*, volume 1, pages 529–532.

van Herwijnen, O., van den Bosch, A., Terken, J., and Marsi, E. (2003). Learning PP attachment for filtering prosodic phrasing. In *Proceedings of the EACL 2003, 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–146.

Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Hirschberg, J. and Prieto, P. (1996). Training intonational phrasing rules automatically for English and Spanish Text-to-Speech. *Speech Communication*, 18:281–290.

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings ICASSP '96*, volume 1, pages 373–376.

Keijsper, C. (1985). *Information Structure*. PhD thesis, University of Amsterdam, The Netherlands.

Klabbers, E. (2000). *Segmental and prosodic improvements to speech generation*. PhD thesis, Eindhoven University of Technology, The Netherlands.

Koehn, P., Abney, S., Hirschberg, J., and Collins, M. (2000). Improving intonational phrasing with syntactic information. In *Proceedings ICASSP '00*.

Kruyt, J. (1985). *Accents from Speakers to Listeners, An experimental study of the production and perception of accent patterns in Dutch*. PhD thesis, Rijksuniversiteit Leiden, The Netherlands.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Mallant, M. (1992). Een bolhoed of een platte hoed? Een intonatie onderzoek naar mogelijke factoren die de keuze tussen de platte hoed en andere melodische realisaties beïnvloeden. IPO Rapport 854, IPO, Eindhoven University of Technology.

Marsi, E. (2001). *Intonation in Spoken Language Generation*. PhD thesis, Nijmegen University, The Netherlands.

Marsi, E., Busser, B., Daelemans, W., Hoste, V., Reynaert, M., and van den Bosch, A. (2002). Combining information sources for memory-based pitch accent placement. In *Proceedings ICSLP, Denver, Co.*, pages 1273–1276.

Marsi, E., Coppen, P.-A., Gussenhoven, C., and Rietveld, T. (1997). Prosodic and intonational domains in speech synthesis. In van Santen, J. P. H., Sproat, R. W., Olive, J. P., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 477–493. Springer-Verlag, New York.

Marsi, E., Reynaert, M., van den Bosch, A., Daelemans, W., and Hoste, V. (2003). Learning to predict pitch accents and prosodic boundaries in dutch. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 2003*, pages 489–496.

Nooteboom, S. and Kruyt, J. (1987). Accents, focus distribution, and the perceived distribution of given and new information: An experiment. *Journal of the Acoustical Society of America*, 82(5):1512–1524.

Ostendorf, M. and Veilleux, N. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, 20(1):27–54.

Paardekooper, P. (1977). *ABN, Beknopte ABN-syntaksis*. Eindhoven, 5th edition.

Pan, S. and Hirschberg, J. (2000). Modeling local context for pitch accent prediction. In *Proceedings ACL-2000, Hong Kong*.

Pan, S. and McKeown, K. (1999). Word informativeness and automatic pitch accent modeling. In *Proceedings of EMNLP/VLC*.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT (Dist. by Indiana University Linguistics Club, Bloomington, IN.).

Pitrelli, J., Beckman, M., and Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings ICSLP '94*, volume 1, pages 123–126, Yokohama.

Prevost, S. and Steedman, M. (1994). Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139–153.

Ratnaparkhi, A. (1997). A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, EMNLP-2*, pages 1–10, Providence, Rhode Island.

Ratnaparkhi, A., Reynar, J., and Roukos, S. (1994). A maximum entropy model for prepositional phrase attachment. In *Proceedings of ARPA Workshop on Human Language Technology*, Plainsboro, NJ.

Rietveld, T. and van Hout, R. (1993). *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, Berlin - New York.

van Rijsbergen, C. (1979). *Information Retrieval*. Buttersworth, London, 2nd edition.

Sanderman, A. (1996). *Prosodic phrasing: production, perception, acceptability and comprehension*. PhD thesis, Eindhoven University of Technology, The Netherlands.

Sanderman, A. and Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech*, 40(4):391–409.

Scharpff, P. and van Heuven, V. (1988). Effects of pause insertion on the intelligibility of low quality speech. In *Proceedings of Speech '88, 7th FASE Symposium, Edinburgh*, pages 261–268.

Scheffé, H. (1952). An analysis of variance for paired comparisons. *Journal of the American Statistical Association*, 47:381–400.

Schmerling, S. (1976). *Aspects of English Sentence Stress*. Texas University Press, Austin.

Selkirk, E. (1984). *Phonology and Syntax: the Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook 3*, pages 371–405.

Selkirk, E. (1995). The prosodic structure of function words. In *University of Massachusetts occasional papers 18: Papers in Optimality Theory*, pages 439–469.

Shattuck-Hufnagel, S. and Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25(2):193–247.

Silverman, K., Kalyanswamy, A., Silverman, J., Basson, S., and Yashchin, D. (1993). Synthesizer intelligibility in the context of a name-and-address information service. In *Proceedings Eurospeech 1993*, volume 3, pages 2169–2172.

Taylor, P., Black, A., and Caley, R. (1998). The architecture of the Festival speech synthesis system. In *Proceedings of the third ESCA/COCOSDA Workshop on speech synthesis*, pages 147–151, Jenolan Caves, Blue Mountains, Australia.

Terken, J. (1993). Synthesizing natural-sounding intonation for Dutch: rules and perceptual evaluation. *Computer Speech and Language*, 7:27–48.

Terken, J. and Nooteboom, S. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information. *Language and Cognitive Processes*, 2:145–163.

Thurstone, L. (1927). A law of comparative judgment. *Psychological Review*, 34:251–259.

Torgerson, W. (1967). *Theory and Methods of Scaling*. London: John Wiley and Sons, 7th edition.

Volk, M. (2000). Scaling up. Using the WWW to resolve PP attachment ambiguities. In *Proceedings of KONVENS-2000*, pages 151–156. Sprachkommunikation, Ilmenau, VDE Verlag.

Volkskrant (2000). PCM Uitgevers, Amsterdam, The Netherlands.

Willemse, R. and Boves, L. (1991). Context free wild card parsing in a Text-to-Speech system. In *Proceedings ICASSP 1991*, volume 2, pages 757–760.

van der Wouden, T., Hoekstra, H., Moortgat, M., Renmans, B., and Schuurman, I. (2002). Syntactic analysis in the spoken dutch corpus (CGN). In Gonzalez Rodriguez, M. and Paz Suarez Araujo, C., editors, *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 768–773.

Zavrel, J., Daelemans, W., and Veenstra, J. (1997). Resolving PP attachment ambiguities with memory-based learning. In Elison, M., editor, *Proceedings of Conference on Computational Natural Language Learning*, pages 136–144.

Reference transcription A

Text 1: Koningskwestie moet D66 weer smoel geven

*Koningsgezind *Nederland / verwijt hem aan *zieltjeswinnerij te doen /// Het *andere deel der natie / *omhelst hem /// Wat bezielde D66-leider Thom de *Graaf // om de *invloed van de *koningin *nu ter discussie te stellen /// Politiek *opportunisme / of *overtuiging ///

*Beide /// De Graaf is niet bij *toeval lid van D66 // de *partij die al *jaren pleit voor een *rechtstreeks gekozen *minister-president /// Hij studeerde *geschiedenis en *staatsrecht /// en was *lid van de commissie De *Koning /// *Die adviseerde om niet het *staatshoofd // maar de Tweede *Kamer een formateur te laten aanwijzen ///

Pakweg *drie jaar *geleden / opperde de *liberale jongerenorganisatie *JOVD // om van het *koningschap een louter *ceremoniële functie te maken /// Ook *toen greep De Graaf naar zijn pen // om te pleiten voor *afschaffing van het *laatste recht van de *koningin /// haar *rol bij *kabinetsformaties /// Voor *persoonlijke invloed is er nog *voldoende *ruimte // zei hij /// *Al haar *andere taken / vallen immers onder de *ministeriële *verantwoordelijkheid ///

*Allerminst *revolutionaire *geluiden /// Vanaf de *invoering van het beginsel van de *openbaarheid van *bestuur in *1980 // is dit type voorstellen *regelmatig gedaan /// *Nu gaat De Graaf een stapje *verder /// Hij beperkt zich *niet tot de *formatie // hij vindt *ook dat de koningin uit de *regering en de Raad van *State moet /// met de *uitdrukkelijke *vermelding / dat hij van haar *geen *siervogel wil maken ///

*Opmerkelijker dan deze uitspraken / zijn de *reacties /// De *media maakten er *nieuws van /// *Voor- en *tegenstanders werden van stal gehaald // en in pakweg

*24 uur *tijd / waren een *rel en een *discussie geboren /// *Timing is het *halve *werk /// weet iedereen in de politiek /// De *Graaf heeft de tijdgeest *goed aangevoeld ///

*Journalisten spelen een *belangrijke *rol /// *Nederland kent een *steeds *transparantere *democratie /// Wie *complotten vermoedt // komt vaak *bedrogen uit /// *Plannen lekken *uit / nog *voordat ze van de *tekentafel zijn /// *Hoezeer er ook een *verbod geldt op het *lekken uit de *ministerraad /// *toch is het beraad *niet *supergeheim /// De *premier doet *elke week *verslag // en via *politici of *ambtenaren komt er voor het *overige *voldoende naar buiten ///

Alleen het *hof is nog een te nemen veste /// *Anders dan in het *verleden / zijn journalisten *allang niet meer bereid te *zwijgen /// Het *geheim van *Noordeinde is *uitdagend en *inspirerend /// de *moeite van het *onderzoeken en *vermelden *waard /// Vanwege de *ministeriële verantwoordelijkheid / is *alles *politiek / en *dus van belang voor het openbaar *bestuur /// Dus *ook de opvattingen en handelwijze van de *koningin ///

Text 2: Bewolking bederft zicht op eclips

Vanwege de *bewolking / is in het *grootste deel van *Nederland / woensdag *weinig te zien geweest / van de *laatste *zonsverduistering van de *eeuw /// *Veel toeschouwers merkten *alleen iets van de eclips // doordat het *enkele graden *kouder werd // en doordat het begon te *schemeren ///

Het *best was het zicht in de strook van *Zeeland tot *Arnhem /// In *Zuid-Limburg waar de zon voor *97 procent werd afgedekt door de *maan // bleef het nogal *bewolkt ///

*Echt donker werd het *nergens /// De *straatverlichting die de *energiebedrijven in *grote delen van Nederland uit *voorzorg hadden *ontstoken /// bleek *overbodig te zijn ///

Energiebedrijf *NUON // *Gelderland / *Flevoland / en *Friesland // besloot tot *woede van de *eclipsliefhebbers als *eerste / om in de drie provincies de *lantaarns te ontsteken /// De lampen hebben *twintig minuten *gebrand /// Het *effect van de *niet-complete zonsverduistering is er *niet door *bedorven /// concluderen de sterrenwachten achteraf ///

In *Utrecht mochten gemeenten *zelf aangeven of ze *straatverlichting wensten /// *Veertien gemeenten // waaronder de stad *Utrecht // hebben daarvan *gebruikgemaakt /// Ook de *provincie Utrecht liet de lantaarns langs de *provinciale wegen ontsteken /// Dat leverde *discussie op met de gemeente *Renswoude // die juist

absolute *duisternis wilde /// In *Renswoude zijn de straatlantaarns langs de provinciale weg *uit gebleven ///

*Oogklinieken van academische *ziekenhuizen / hebben gistermiddag *tien mensen behandeld / die *last hadden van hun *ogen / na het kijken naar de *eclips ///

*Geen van hen had *blijvend *letsel // maar volgens *specialisten op het gebied van de *oogheelkunde / manifesteert de *voornaamste schade zich vaak pas enkele dagen *na het kijken in de felle zon ///

In *Gouda werd een *vrouw opgenomen in het Groene *Hart Ziekenhuis /// De *ernst van haar letsel is nog *onbekend ///

De *massale *aandacht voor de verduistering / leidde tot *grote *rust op de wegen en in *straten /// Op de *parkeerplaatsen langs de *snelwegen was het *wel iets *drukker dan normaal // maar het korps *Landelijke *Politiediensten hoefde *Nauwelijks automobilisten te berispen / die op de *vluchtstrook de verduistering afwachtten ///

*Files ontstonden pas aan het *einde van de middag // toen de *aanvoerroutes vanuit het *zuiden *dichtslibden met terugkerende *eclipstoeristen ///

Bij het *Belgische plaatsje *Virton / stond gisteravond *negentien kilometer /// in *Luxemburg en *Noord-Frankrijk stonden files op *alle *N-wegen /// en in *Duitsland werg op de *A *9 tussen *München en *Neurenberg / *dertig kilometer geteld ///

In *Nederland bleef het echter *rustig /// ter hoogte van *Maastricht was het alleen aan het *einde van de *middag wat *drukker dan *normaal ///

De *volgende totale zonsverduistering / is in Nederland te zien op *7 oktober *2135 ///

E-mail

mail01

Tijdens het laatste *verjaardagsetentje / is er bij de aanwezigen de *vraag gerezen // hoe wij *verder willen met de verjaardagen/// Er zijn een *paar *opties /// 1 / we blijven get in *deze vorm doen /// *een keer per jaar / koop je de *cadeautjes voor diegene die *na jou jarig is /// 2 / *per *verjaardag wordt er een rondje gemaild / wie er *mee wil doen met het *verjaardagscadeau // en *wie het gaat *kopen /// 3 / *natuurlijk zijn er ook *andere opties mogelijk /// Wil *iedereen naar *mij mailen / wat je ervan *vindt /// Dan zal ik *binnenkort de *uitslag naar iedereen terugsturen // dan is er tenminste *duidelijkheid bij de *volgende verjaardagen ///

mail02
Na een *paar *iteraties / is gebleken / dat bijna *alle betrokkenen *inderdaad *beschikbaar zijn / om op *27 *oktober vanaf *drie uur te *vergaderen op het *IPO // om de *voortgang van het *MATIS project te bespreken /// Bij *deze bevestig is dus de *geplande *bijeenkomst ///

mail03
Vandaag om half *twaalf / zullen *Patrick *Morley en *Leo *Coolen / enkele *mededelingen doen /// *Locatie /// *Gehoorzaal *Leidschendam / met een *videoverbinding naar de *Colloquiumzaal in *Groningen /// U bent van *harte *uitgenodigd ///

mail04
Refererend naar onze *afspraak van donderdag *24 *augustus // wil ik nog even *mededelen / dat naar *alle *waarschijnlijkheid / *Caroline *Wolsheimer vanuit *KPN *BU *TC / hierbij *aanwezig zal zijn /// Zij zal mij dit / via mijn *mobiele *voicemail laten *weten /// Aangezien ik woensdag *23 augustus *niet aanwezig ben // wilde ik je hiervan in *ieder geval even op de *hoogte stellen /// *Graag tot *donderdagachtend / *negen uur ///

mail05
We hebben een *afspraak staan voor *morgenmiddag / om te praten over mijn *project /// Ik heb net met *Jacques overlegd // en het lijkt ons *niet echt heel erg *zinvol / om deze bespreking *door te laten gaan // aangezien er van *mijn kant *weinig te *bespreken is // onder *andere / doordat ik in de *tussentijd nog op *vakantie ben geweest // en iedereen zijn tijd dan wel *beter kan besteden /// Ik zal *morgen een *verslag sturen / van de stand van *zaken op *dit moment /// En ik stel *voor / dat we een *nieuwe datum afspreken ///

mail06
Allereerst mijn *excuses / voor het *niet direct reageren op je *mail /// *Bedankt voor je *aanmelding /// Op *11 september *aanstaande / krijg je te horen of je *geplaatst bent of *niet /// er wordt een *evenredige *verdeling gemaakt / naar *sekse en *gebouwdeel // *C of *E-hal etcetera /// Hou er *wel *rekening mee / dat de *eerste *EHBO-les / al op *13 september aanstaande begint /// De cursussen worden *apart gegeven ///

mail07
Op *vrijdag *14 juli aanstaande / zal de *gemeente de weg *Winschoterdiep *oostzijde / *opnieuw asfalteren /// Dit betekent / dat vanaf *elf uur *'s morgens de / *in- en *uitrit van de *paviljoens / *niet meer gebruikt kunnen worden /// Vanaf *dat moment / kan *gebruik gemaakt worden van de *ingang welke *gecreëerd zal worden bij de *parkeerplaats aan de *achterzijde van de paviljoens /// Het *hekwerk // naast het *electriciteitshuisje // zal op *dat moment voor het *in- en *uitrijden *geopend zijn

/// Men kan de paviljoens dan *bereiken / door over de *Europaweg of de *parallelweg naast de Europaweg te rijden // en via de *Barkhuisstraat // in de *woonwijk / naar het *toegangshek te rijden /// *Excuses voor de eventuele *overlast ///

mail08
Degenen die *morgen bij het afscheid van *Jan komen *zingen // hebben *zojuist van mij een *exemplaar van het lied / *Loftrompet voor *Jantje *Pel / ontvangen /// Het druist in tegen *alle *rijmregels / en is *uitermate *oubollig // maar het is volgens mij *wel *zingbaar /// Het *voorstel is om *morgenmiddag om half *twee / even te *repeteren in *1.18 ///

mail09
Ik vind het *voorstel van onze *planologische *dienst / *buitengewoon *acceptabel /// In het *kader van de *logistieke *coördinatie van de gemiddelde *radiohead-bezoeker // ben ik *uiterst tevreden met onderstaande *aanpak /// Ik kan me er echter *niet van weerhouden / om *mede te delen / dat ik zal *trachten om fase *1 in de keten op *tijd te *realiseren // zodat we *meteen kunnen doorstoten tot de *vierde en *laatste fase /// *Deze fase is uiteraard het *genieten van een naar verwachting *fantastisch *concert / van de *legendarische band *Radiohead ///

mail10
Ik kan helaas vrijdag niet // en was van plan *donderdag te komen /// Maar dan is *Reinier er denk ik niet /// Kunnen we de *vergadering anders verplaatsen naar *dinsdag *19 *september ///

mail11
Je hebt *heel treffend beschreven hoe dat *gaat / als je van *vakantie terugkomt / en een *overvolle *e-mailbox aantreft /// Maar nu je het *zegt // kan ik me *ineens weer *herinneren / dat ik *onderstaand mailtje *gelezen heb // en was hij ook *zo *teruggevonden ///

*Gefeliciteerd dat je *beide *vacatures voor het *IDUSI-project hebt kunnen invullen /// *Kopieën van de *aanstellingsbrieven / en de *complete *contactgegevens van de *onderzoekers in *kwestie // zie ik te zijner tijd *graag tegemoet ///

De *voorzitter van de *BC *UCD // Joerka *Deen // Piet *Bögels // voorzitter *PC // en *ik // willen *graag een keer bij jullie *langskomen / om *nader kennis te maken /// *Half *november willen we een *eerste *begeleidingscommissievergadering houden // maar over *beide zaken volgt nog nader *bericht ///

*Rest mij om jullie *heel veel *succes en *plezier toe te wensen / bij de *uitvoering van het *project ///

mail12
Dat levert *vermoedelijk *problemen op voor *iedereen /// Als je een *half uur per *praatje rekent / voor *presentatie en *commentaar // heb je voor de *vier praatjes al *twee uur nodig /// Het praatje van *Jacques is eigenlijk *langer // dus twee uur is *krap / en dan hebben we nog *geen tijd aan de *posters besteed /// *Drie *alternatieven /// om *negen uur beginnen // maar dan moet *Jacques eerder weg /// om half *twee beginnen // maar dan komt Jacques *later /// om *negen uur beginnen // *ophouden als Jacques *weg moet // en weer *verder gaan als hij *terug komt /// *Graag *reactie ///

mail13
Het is weer *tijd / voor het *culinaire *hoogtepunt van het *academisch *jaar /// Op *woensdag *26 *juli / organiseert *IPO haar *jaarlijkse *barbecue /// *Iedereen is vanaf half *vijf *uitgenodigd / om dit *feest van *haute *cuisine en *slechte *tafelmanieren *bij te wonen /// Voor slechts *5 gulden per *persoon / kun je *genieten van het *beste wat *flora en fauna te bieden hebben // *geroosterd en van een *plastic *bord /// Als je *mee wilt doen // en dat *wil je // meld je dan zo *snel mogelijk aan bij *ondergetekende ///

mail14
Op *31 *oktober / hoop ik *27 *jaar te worden /// Aanleiding *genoeg voor een *bescheiden *feestje / dacht ik /// Mocht je *tijd en *zin hebben // dan staat er op *zaterdagavond *28 oktober / vanaf een *uur of *8 / een stuk *gebak op je te wachten / in de *Achterstraat *42 in *Lochem // voor de *meesten van jullie *geen onbekend *adres /// Mocht je *niet kunnen komen // laat dat dan alsjeblieft / *tijdig / even weten /// We hebben *ruimte *genoeg / als je wilt blijven *slapen // maar breng dan wel zelf *slaapspullen mee // *tenzij je *genoegen wilt nemen met een *koude *harde *ondergrond zonder *dekens /// *Hopelijk zie ik jullie *allemaal verschijnen de 28e ///

mail15
Ik weet inmiddels hoe laat ik *aankom in *Nijmegen // de 26e /// *Tenminste / als de *trein geen *vertraging heeft /// Maar dan ben ik er om *18.23 uur /// Komt dat een beetje *redelijk *uit /// Lijkt me *wel / he /// Kunnen we *mooi iets *eten // en daarna nog *twee uur *stadten /// Je weet toch wel *zeker / dat het op *donderdag *koopavond is /// Anders rekenen we daar voor *niks op ///

Sentences of experiments on perceptive cost of errors

B

Experiment on prosodic phrasing

Noun attached PP (NOUN)

1 De rector heeft gezegd // dat er veel weerstand [ ] tegen de plannen is geweest.

2 De rector heeft gezegd // dat het tijd is voor het culinaire hoogtepunt [ ] van het academisch jaar.

3 In het nieuwsblad staat // dat de volledige inhoud [ ] van de kassa is gestolen.

4 In het nieuwsblad staat // dat er een strenge grenscontrole [ ] op wapens wordt gehouden.

5 De verslaggever heeft gehoord // dat de dreiging [ ] van een terroristische aanslag is toegenomen.

6 De verslaggever heeft gehoord // dat interviews tot de meest gelezen stukken [ ] in een krant behoren.

7 De buurman beweerde // dat hij zijn vertrouwen [ ] in de toekomst is verloren.

8 De buurman beweerde // dat zijn zoontje de remmen [ ] van zijn fiets had gemaakt.

9 Moeder vertelde // dat het tegenvoorstel [ ] van de reisorganisatie acceptabel was.

10 Moeder vertelde // dat zij de oude dagboeken [ ] van Wolkers had gelezen.

Verb attached PP (VERB)

11 De rector heeft gezegd // dat de dronken medewerker [ ] naar het ziekenhuis is vervoerd.

12 De rector heeft gezegd // dat de gemeentelijke brandweer [ ] tot ontruiming heeft besloten.

13 In het nieuwsblad staat // dat de werkloosheid [ ] met dertien procent is gestegen.

14 In het nieuwsblad staat // dat de aangekondigde zonsverduistering [ ] tot grote rust heeft geleid.

15 De verslaggever heeft gehoord // dat de loco-brugemeester [ ] uit het gemeentebestuur moet stappen.

16 De verslaggever heeft gehoord // dat de Keniaanse volkspresident [ ] voor zijn leven heeft gevreesd.

17 De buurman beweerde // dat hij voor het kinderfonds [ ] naar Zuid-Amerika is geweest.

18 De buurman beweerde // dat alle buurtbewoners [ ] op het nieuwe terrasje wilden zitten.

19 Moeder vertelde // dat de brutale overvaller [ ] aan een onbezorgde toekomst heeft gedacht.

20 Moeder vertelde // dat de zieke wethouder [ ] in de behandelmethode heeft geloofd.

Experiment on accentuation

Verb should be deaccented (ARG)

1 De buurman heeft een boek gelezen.

2 Hij heeft de tas naast de auto gezet.

3 Het kind heeft in de tuin gespeeld.

4 Karel heeft op zijn broer gewacht.

5 De directeur heeft een toespraak gehouden.

6 Mijn neef heeft een spel bedacht.

7 Zij heeft haar pop op de tafel gelegd.

8 De man is uit het raam gesprongen.

9 Guus heeft naar de vakantie verlangd.

10 Mijn moeder heeft een trui gebreid.

Verb should be accented (COND)

11 De bewaker heeft die nacht gelezen.
12 De man heeft in de tent gerookt.
13 De trein is om drie uur gestrand.
14 De wandelaars zijn in de regen vertrokken.
15 Mijn moeder heeft verbaasd gereageerd.
16 De hond heeft die middag gebeten.
17 De leiding is door de kou gesprongen.
18 De prijzen zijn op zondag gestegen.
19 Het kind is met spoed geopereerd.
20 Haar zus heeft luidkeels gezongen.

Sentences of evaluation of ECLIPSE

C

Noun attached PP

1 Vanwege de bewolking is in het grootste deel van Nederland woensdag weinig te zien geweest van de laatste zonsverduistering van de eeuw.

2 Geen van hen had blijvend letsel, maar volgens specialisten op het gebied van de oogheelkunde manifesteert de voornaamste schade zich vaak pas enkele dagen na het kijken in de felle zon.

3 Vanaf dat moment kan gebruik gemaakt worden van de ingang welke gecreëerd zal worden bij de parkeerplaats aan de achterzijde van de paviljoens.

4 Iedereen is vanaf half vijf uitgenodigd om dit feest van haute cuisine en slechte tafelmanieren bij te wonen.

Verb attached PP

5 Vanwege de ministeriële verantwoordelijkheid is alles politiek en dus van belang voor het openbaar bestuur.

6 Files ontstonden pas aan het einde van de middag, toen de aanvoerroutes vanuit het zuiden dichtslibden met terugkerende eclipstoeristen.

7 In het kader van de logistieke coördinatie van de gemiddelde concertbezoeker ben ik uiterst tevreden met onderstaande aanpak.

8 Het is weer tijd voor het culinaire hoogtepunt van het academisch jaar.

Long first major constituent

9 Vanaf de invoering van het beginsel van de openbaarheid van bestuur in 1980 is dit type voorstellen regelmatig gedaan.

10 Hoezeer er ook een verbod geldt op het lekken uit de ministerraad, toch is het beraad niet supergeheim.

11 Tijdens het laatste verjaardagsetentje is er bij de aanwezigen de vraag gerezen hoe wij verder willen met de verjaardagen.

12 Aangezien ik woensdag 23 augustus niet aanwezig ben wilde ik je hiervan in ieder geval even op de hoogte stellen.

Various punctuation marks

13 Ook toen greep De Graaf naar zijn pen om te pleiten voor afschaffing van het laatste recht van de koningin: haar rol bij kabinetsformaties.

14 In Zuid-Limburg, waar de zon voor 97 procent werd afgedekt door de maan, bleef het nogal bewolkt.

15 Als je mee wilt doen (en dat wil je), meld je dan zo snel mogelijk aan bij ondergetekende.

16 Mocht je niet kunnen komen laat dat dan alsjeblieft (tijdig!) even weten.

Sentence final verb preceded by an argument

17 Ook de provincie Utrecht liet de lantaarns langs de provinciale wegen ontsteken.

18 Dat leverde discussie op met de gemeente Renswoude, die juist absolute duisternis wilde.

19 Vandaag om half twaalf zullen Harry Jansen en Thomas Bergman enkele mededelingen doen.

20 Gefeliciteerd dat je beide vacatures voor het IDUSI-project hebt kunnen invullen!

Sentence final verb preceded by a condition

21 De straatverlichting die de energiebedrijven in grote delen van Nederland uit voorzorg hadden ontstoken, bleek overbodig te zijn.

22 Hij beperkt zich niet tot de formatie, hij vindt ook dat de koningin uit de regering en de Raad van State moet.

23 Zij zal mij dit via mijn mobiele voicemail laten weten.

24 Het hekwerk (naast het electriciteitshuisje) zal op dat moment voor het in- en uitrijden geopend zijn.

Results experiments on PP-attachment

D

Table D.1: *Performance measures in percentages on PP-attachment prediction for the CGN material (1004 instances) by* MBL.

MBL		NOUN			VERB		
	accuracy	**precision**	**recall**	**F$_{\beta=1}$**	**precision**	**recall**	**F$_{\beta=1}$**
all	77	81	81	81	71	69	70
N1	67	69	85	76	63	39	48
P	73	81	72	76	63	73	67
V	59	62	82	71	46	22	30
N2	62	64	86	74	52	24	33
cooc	68	74	73	74	59	60	59
N1+P	73	81	73	77	63	72	67
N1+V	65	67	86	75	62	32	41
N1+N2	70	70	89	78	70	40	50
N1+cooc	71	80	71	75	61	71	65
P+V	73	77	79	78	66	63	64
P+N2	72	78	75	76	63	77	64
P+cooc	73	80	76	78	65	70	67
V+N2	65	65	91	76	63	23	34
V+cooc	68	74	74	74	59	59	59
N2+cooc	70	77	72	74	60	67	63
baseline	60	60	100	75	-	0	-

Results experiments on argument – condition

E

Table E.1: *Performance measures in percentages on* ARG *versus* COND *prediction for the CGN material (1613 instances) by* MBL.

MBL		ARG			COND		
	accuracy	**precision**	**recall**	**F$_{\beta=1}$**	**precision**	**recall**	**F$_{\beta=1}$**
all	89	90	97	93	77	50	61
P	82	83	100	90	35	1	3
N	89	91	96	93	75	55	63
V	80	83	96	89	21	5	8
cooc	80	83	95	89	25	8	12
P+N	89	90	97	94	78	49	60
P+V	81	83	97	90	27	6	10
P+cooc	82	83	99	90	37	3	5
N+V	88	90	97	93	79	46	58
N+cooc	89	90	96	93	76	52	61
V+cooc	82	83	99	90	13	2	4
P+N+V	88	90	97	93	77	49	60
P+N+cooc	89	90	97	93	79	49	59
P+V+cooc	83	83	99	90	52	3	6
N+V+cooc	88	90	96	93	76	51	61
baseline	84	84	100	91	-	0	-

Summary

The quality of synthetic speech in Text-to-Speech synthesis often sounds unnatural. This is partly due to the lack of proper prosodic structure: phrase boundaries and accents are missing or allocated incorrectly. Correct prosodic structure reduces the time and effort that it takes listeners to process and understand artificially generated speech. As described in Chapter 1, the aim of the research described in this thesis is to improve the assignment of prosodic structure on the basis of syntactic and lexical information. This project combines two lines of research: language engineering and psycholinguistics. We apply language engineering techniques, in combination with the psycholinguistic information which we have obtained through perception experiments, to enhance the syntactic information we have obtained from a state-of-the art syntactic parser.

In Chapter 2 we described an evaluation study which was carried out to investigate what the major error inducing factors are in automatic prosody assignment by existing Text-to-Speech systems for Dutch. For this evaluation we compared prosodic structure assigned by TTS systems to a reference transcription by 10 human experts. The results of the evaluation showed that a major problem for both phrasing and accentuation is incorrect or insufficient syntactic information. For phrasing a major problem is PP attachment ambiguity, whereas for accentuation of a main verb the major problem is the distinction between argument and condition. A second evaluation study and a perception experiment showed that proper syntactic information together with a revised phrasing algorithm improve the assignment of prosodic structure significantly. Another major problem for prosody assignment is the influence of context, but this is beyond the scope of this thesis.

In Chapter 3 we described two experiments that investigated the listeners' tolerance for different types of phrasing and accentuation errors. In the first experiment we showed that incorrect insertion of a phrase boundary at the juncture preceding an attached PP is less acceptable for the listener than incorrect omission. Thus, provided that it cannot avoid making mistakes, a TTS system rather allocates too few boundaries than too many. This implies that machine learning algorithms should have a bias for predicting noun attachment over verb attachment, because for noun attached PP's correct phras-

ing means that no boundary should be allocated preceding the PP, whereas for verb attached PP's there should. In the second experiment we showed that incorrect accent insertions on a sentence final verb are as bad as accent omissions. Thus, the system should find an optimum in accent allocation, such that there are as few accent insertions and omissions as possible. This implies that machine learning algorithms should be as good in predicting arguments as in predicting conditions.

In Chapter 4 we investigated how PP attachment ambiguities can be resolved to improve prosodic phrasing in synthetic speech. From a treebank of spoken Dutch we selected instances of the attachment of prepositional phrases to either a noun or verb in the sentence. We trained two machine learning algorithms (MBL and RIPPER) to make the distinction between noun and verb attachment on the basis of lexical information and a co-occurrence strength feature derived from the WWW. The trained models were tested on the Spoken Dutch Corpus data by means of cross-validation experiments, and on held-out newspaper and e-mail data. The results indicated that the trained models have a reasonably stable performance on different kinds of data. Comparison with the reference transcription showed that the availability of correct PP attachment information improves the performance on prosodic phrase boundary allocation.

From the evaluation in Chapter 2 we learned that accents are not always allocated correctly, especially in the case of the sentence final verb. The identity of the preceding nominal constituent (whether it is an argument or a condition) is of importance for the accentuation of the verb. In Chapter 5 we first discussed the definition of 'argument' and 'condition'. Next, we performed machine learning experiments for predicting the identity of the nominal constituent. We trained two machine learning algorithms (MBL and RIPPER) on making the distinction between argument and condition on the basis of lexical information and a co-occurrence strength feature derived from the WWW. Again, the trained models were tested on the Spoken Dutch Corpus data by means of cross-validation experiments, and on held-out newspaper and e-mail data. Comparison with the reference transcription showed that having information available about argument vs. condition improves the performance on accentuation.

In Chapter 6 we described the evaluation of our prosody module (ECLIPSE) which resulted from studies presented in the previous chapters. The evaluation is two-fold, consisting of an objective evaluation through comparison with the reference transcription and a subjective evaluation by means of a perception experiment in which listeners had to indicate the acceptability of the different realizations of the same sentence. The results of the objective evaluation showed that ECLIPSE performs considerably better than PROS-3 (an earlier developed system that assigns prosodic structure on the basis of a description of the word category of each word of the input text together with the relations between the words). The results of the subjective evaluation showed that the listeners and the experts preferred ECLIPSE over PROS-3 and that there is no significant difference between ECLIPSE and the reference transcription.

ECLIPSE performs considerably better on accentuation and prosodic phrasing than its competitors in existing Text-to-Speech systems for Dutch. Listeners and experts judge ECLIPSE as acceptable as the reference transcription of human experts. We therefore conclude in Chapter 7 that applying machine learning techniques, constrained by information about the tolerance for errors, is useful to obtain elaborate syntactic information. From this elaborate syntactic structure, in combination with lexical information, a perceptually appropriate prosodic structure can be computed for synthetic speech.

Samenvatting

Synthetische spraak klinkt vaak onnatuurlijk. Dit is deels het gevolg van het gebrek aan een goede prosodische structuur: frasegrenzen en accenten ontbreken of worden incorrect geplaatst. Goede prosodische structuur vermindert de benodigde tijd en de moeite die luisteraars moeten doen voor het verwerken en begrijpen van automatisch gegenereerde spraak. Zoals vermeld in hoofdstuk 1 is het doel van het in dit proefschrift beschreven onderzoek is het verbeteren van de toekenning van prosodische structuur op basis van syntactische en lexicale informatie. In dit project worden twee onderzoeksgebieden gecombineerd: taaltechnologie en psycholinguïstiek. We zullen gebruik maken van taaltechnologische technieken, om de syntactische informatie uit te breiden die we verkregen hebben door middel van automatische analyse. Hierbij wordt rekening gehouden met de psycholinguïstische informatie die we verkregen hebben door middel van perceptie-experimenten.

In hoofdstuk 2 hebben we een evaluatie beschreven die uitgevoerd is om te achterhalen wat de belangrijkste factoren zijn die fouten veroorzaken in automatische toekenning van prosodische structuur door bestaande Tekst-naar-Spraak (TTS) systemen voor het Nederlands. Voor deze evaluatie hebben we de prosodische structuur die toegekend wordt door TTS-systemen vergeleken met een referentietranscriptie van 10 experts. De resultaten van deze evaluatie laten zien dat een belangrijk probleem voor frasering en accentuering incorrecte of ontoereikende syntactische informatie is. Een belangrijk probleem voor frasering is de ambiguïteit van aanhechting van een PP (voorzetselfrase), en voor accentuering van het hoofdwerkwoord is een belangrijk probleem het maken van onderscheid tussen argument en conditie. Een tweede evaluatie en een perceptie experiment laten zien dat goede syntactische informatie in combinatie met een herzien fraseringsalgoritme de toekenning van prosodische structuur significant verbeteren. Een ander belangrijk probleem voor prosodie-toekenning is de invloed van context, maar dat probleem valt buiten het bereik van dit proefschrift.

In hoofdstuk 3 hebben we twee experimenten beschreven die de tolerantie van luisteraars voor verschillende typen fouten in de frasering en accentuering te bestuderen. In het eerste experiment laten we zien dat het ten onrechte toekennen van een frasegrens voorafgaand aan een aangehechte PP minder acceptabel is voor de luisteraar dan het

ten onrechte weglaten. Een TTS-systeem kan dus beter te weinig grenzen toekennen dan te veel. Dit betekent dat getrainde algoritmen een voorkeur moeten hebben voor aanhechting van de PP aan het zelfstandig naamwoord, boven aanhechting van de PP aan het werkwoord. Aanhechting aan het zelfstandig naamwoord betekent namelijk dat er bij correcte frasering geen grens wordt geplaatst voor de PP, terwijl dat bij aanhechting aan het werkwoord wel het geval is. In het tweede experiment laten we zien dat het ten onrechte toekennen van een accent op een zinsfinaal werkwoord even erg is als het ten onrechte weglaten van een dergelijk accent. Een TTS-systeem moet dus een optimum vinden voor accentplaatsing, zodanig dat er zo weinig mogelijk ten onrechte toegekende en weggelaten accenten voorkomen. Getrainde algoritmen (die onderscheid maken tussen argumenten en condities die voorafgaan aan het zinsfinale werkwoord) moeten dus even goed zijn in het voorspellen van een argument als van een conditie.

In hoofdstuk 4 hebben we getoetst hoe PP-aanhechting voorspeld kan worden om frasering in synthetische spraak te verbeteren. Uit het Corpus Gesproken Nederlands hebben we frasen geselecteerd die een PP bevatten die aangehecht is aan een zelfstandig naamwoord of een werkwoord. We hebben twee lerende algoritmen (MBL en RIPPER) getraind voor het maken van onderscheid tussen de twee typen PP-aanhechting. Deze systemen doen dat op basis van lexicale informatie en informatie over de frequentie van het samen voorkomen van het voorzetsel en het zelfstandig naamwoord of het werkwoord in een groot corpus zoals het WWW. De getrainde modellen zijn getest op data van het CGN door middel van cross-validatie experimenten, en op krant- en e-maildata. Uit de resultaten blijkt dat de getrainde modellen een tamelijk stabiel resultaat opleveren op verschillende typen data. Vergelijking met de referentietranscriptie laat zien dat het gebruik van informatie over PP-aanhechting de toekenning van frasegrenzen verbetert.

Door de evaluatie in hoofdstuk 2 zijn we te weten gekomen dat accenten niet altijd correct worden toegekend, met name in het geval van zinsfinale werkwoorden. De identiteit van de voorafgaande nominale constituent (argument of conditie) is van belang voor de accentuering van het werkwoord. In hoofdstuk 5 hebben we de definitie van argumenten en condities besproken. Vervolgens hebben we experimenten met lerende systemen beschreven voor het voorspellen van de identiteit van de nominale constituent. We hebben twee lerende systemen (MBL en RIPPER) getraind voor het maken van onderscheid tussen argument en conditie. Deze systemen doen dat op basis van lexicale informatie en informatie over de frequentie van het samen voorkomen van het voorzetsel en het werkwoord. De getrainde modellen zijn wederom getest op data van het CGN door middel van cross-validatie experimenten, en op krant- en e-maildata. Uit vergelijking met de referentietranscriptie bleek dat de toekenning van accenten verbetert wanneer gebruik wordt gemaakt van informatie over argument versus conditie.

In hoofdstuk 6 hebben we de evaluatie beschreven van onze prosodie module (ECLIPSE) die het resultaat is van het onderzoek dat gepresenteerd is in voorgaande hoofdstukken. De evaluatie is tweeledig en bestaat uit een objectieve evaluatie door middel van vergelijking met de referentietranscriptie en een subjectieve evaluatie door middel van een perceptie experiment. In dit experiment moesten luisteraars de acceptabiliteit beoordelen van de verschillende realisaties van dezelfde zin. De resultaten van de objectieve evaluatie laten zien dat ECLIPSE beter is dan PROS-3 (een eerder ontwikkeld systeem dat prosodische structuur toekent op basis van een beschrijving van de woordklasse van elk woord uit de invoertekst en relaties tussen de woorden). Uit de resultaten van de subjectieve evaluatie bleek dat luisteraars en experts een voorkeur hebben voor ECLIPSE en de referentietranscriptie boven PROS-3, en dat er geen significant verschil is tussen ECLIPSE en de referentietranscriptie.

ECLIPSE levert aanzienlijk betere resultaten op voor frasering en accentuering dan zijn concurrenten in bestaande Tekst-naar-Spraak systemen voor het Nederlands. Luisteraars en experts vinden de prosodische structuur van ECLIPSE even acceptabel als de referentietranscriptie van experts. Daarom concluderen we in hoofdstuk 7 dat het gebruik van lerende systemen, waarbij rekening wordt gehouden met informatie over tolerantie voor fouten in frasering en accentuering, zinvol is voor het verkrijgen van uitgebreide syntactische informatie. Op basis van deze syntactische informatie, in combinatie met lexicale informatie, kan een perceptief correcte prosodische structuur worden berekend voor automatisch gegenereerde spraak.

Curriculum vitae

18 August 1974	Born in Amsterdam, The Netherlands
1986 – 1992	VWO ('pre-university education'), Jeroen Bosch College, 's-Hertogenbosch
1992 – 1993	Pharmacy, Utrecht University
1993 – 1994	Dutch Linguistics and Literature, Utrecht University, Foundation Course ('propaedeuse')
1994 – 1995	General Linguistics, Utrecht University
1995 – 1998	Phonetics, Utrecht University Master's Degree
1998 – 2000	Junior Researcher on European project COMRIS, IPO, Center for User-System Interaction, Eindhoven University of Technology
2000 – 2004	Graduate Student UCE group, Department of Industrial Design, Eindhoven University of Technology, and ILK group, Department of Computational Linguistics and AI, Tilburg University