



KTH Tal, musik och hörsel

Predicting Prosodic Phrase Boundaries for Speech Synthesis

Christina Ericsson



Supervisor: Beata Megyesi

Approved 2003-11-20 :

Stockholm
2003-11-20

Master thesis in Speech Communication

Department of Speech, Music and Hearing
KTH, Kungliga Tekniska Högskolan
100 44 Stockholm



KTH Tal, musik och hörsel

Examensarbete i talkommunikation

Predicering av prosodiska frasgränser för talsyntes

Christina Ericsson

Godkänt: . 2003-11-20

Examinator Rolf Carlson

Handledare Beata Megyesi

Sammanfattning

Det har visats att prosodiska frasgränser är viktiga i talproduktion och -perception. De signalerar meddelandets syntaktiska struktur, och hjälper lyssnaren att dela in talet i meningsfulla bitar. Syftet med denna uppsats är att utifrån en språkligt annoterad text automatiskt predicera prosodiska frasgränser i ett text-till-tal-system.

Prediceringarna görs med ett fåtal regler och statistik baserade på prosodiskt annoterat talmaterial.

Prediceringarna utvärderades genom en jämförelse med prosodiskt annoterat talmaterial och ett litet lyssningstest med syntetiskt tal. Resultaten visar att det är möjligt att predicera prosodiska frasgränser, även om träningsmaterialet är förhållandevis litet.



**KTH Speech, Music
and Hearing**

Master Degree Project in Speech Communication

Predicting Prosodic Phrase Boundaries for Speech Synthesis

Christina Ericsson

Approved: 2003-11-20

Examiner Rolf Carlson Supervisor Beata Megyesi

Abstract

It has been shown that prosodic phrase boundaries are important both in speech production and perception. They signal the syntactic structure of the message, and help the listener to divide the speech into meaningful chunks. The purpose of this thesis is, given a linguistically annotated text, to predict prosodic phrase boundaries in a text-to-speech system. The predictions are done with a handful rules and statistics based on prosodically annotated speech material. The predictions were evaluated by a comparison with new prosodically annotated speech material, and a small listening test with synthetic speech was performed. The results show that it is possible to successfully predict prosodic phrase boundaries, even if the training material is relatively small.

1. Introduction	5
2. Prosodic phrase boundaries	6
2.1. Location of prosodic phrase boundaries in different communicative situations	6
2.2. Acoustic realisation of prosodic phrase boundaries	7
2.3. Perception of prosodic phrase boundaries	9
2.4. Automatic prediction of prosodic phrase boundaries	10
2.4.1. Punctuation symbols	10
2.4.2. Content words and function words	10
2.4.3. Part-of-speech tags	10
2.4.4. Syntax	12
3. Predicting prosodic phrase boundaries	13
3.1. Data preparation	16
3.1.1. Audio data	16
3.1.2. Linguistic analysis	17
3.1.3. The tagger - TnT	18
3.1.4. The parser - SPARK	18
3.1.5 Other linguistic features	21
3.2. Statistical calculations on the training material	21
3.3. Rules derived from the statistics	22
3.4. Validation procedure	24
3.4.1. Prediction using statistics	24
3.7. Acoustic realisation of prosodic phrase boundaries	30
3.8. The final system used for prediction and realisation of prosodic phrase boundaries	31
4. Evaluation test	35
4.1. Test method	35
4.2. Test results	35
5. Discussion and future research	38
6. References	40
7. Appendix	42
A. Mapping table of the linguistic features	42
B. Examples from the statistics	43
C. Test form	44
D. Synthesised sentences	45

1. Introduction

The purpose of the thesis is to predict the locations and (to a certain degree) the realisations of prosodic phrase boundaries in order to improve synthetic speech. The study presents an empirical investigation about how prosodic phrase boundaries can be automatically predicted given a linguistically annotated text.

Chapter 2 presents a definition of prosodic phrase boundaries, and their syntactical, perceptual and acoustic functions are explained. Prosodic phrase boundaries reflect the syntactic structure of the message, and can signal that a word is more important than the others are. The boundaries in a text-to-speech system are typically acoustically realised by final (or pre-boundary) lengthening, silence intervals (pauses), and f_0 -variations (Black & Taylor, 1998). Speakers use prosodic phrase boundaries to signal the structure of the message. At the same time they are provided with some extra time to plan the rest of the message. Also physical restrictions as the size of the speaker's lungs have an effect of where the prosodic phrase boundaries are located (Black and Lenzo, 1999). For the listener, the boundaries help when structuring the message into sentences, phrases and clauses (Carlson et al., 2002).

Chapter 3 describes the linguistically, acoustically and perceptually analysed material used for calculating statistics in order to predict prosodic phrase boundaries. The linguistic analyses include tagging, parsing and other linguistic analyses, while the acoustic analyses mainly concern durational and f_0 -phenomena. Furthermore, a perceptual annotation of where the prosodic boundaries were located was done. The pre-analysed recordings consist of two news recordings, one used as training material, and one used as validation or testing material. A recording of spontaneous speech from a radio interview was used as validation or testing material, to see how the statistics and rules elaborated for news text performed on this type of speech.

The method for predicting prosodic phrase boundaries was done in two steps:

1. Statistics were calculated from a pre-analysed training news recording, and rules mainly derived from the statistics were written.
2. The statistics and rules were tested on a similar pre-analysed news recording in order to see which statistical data and which rule that gave the best results when compared to the perceptual annotations.

Chapter 4 describes the listening test, where 16 sentences of different types and complexity. The sentences were synthesised in two different ways; once with prosodic phrase boundaries acoustically realised only at delimiters, and once by using the statistics and rules that gave the best results in the validation session. These were presented in random order to ten test subjects, of which five were used to synthetic speech and five were inexperienced listeners and they were asked to grade the groupings of the words on a scale 1-5.

A discussion about further improvements and suggestions about other investigations can be found in chapter 5. To improve the prediction of prosodic phrase boundaries, more linguistically, perceptually and acoustically annotated text would be desirable, as well as a bigger listening test.

The work was carried out within the GROG project at the Department of Speech, Music and Hearing at KTH. GROG stands for 'Gräns och gruppering – Strukturering av talet i olika kommunikativa situationer' (Boundaries and groupings – the structuring of speech in different communicative situations), and, as the name suggests, aims to model the prosodic structuring of speech in terms of boundaries and groupings (Carlson et al., 2002). The GROG project is supported by the Swedish Research Council (VR).

2. Prosodic phrase boundaries

This chapter describes the function of prosodic phrase boundaries in speech production and perception, how they are distributed in different communicative situations, how they are acoustically realised, and finally how they can be automatically predicted.

A prosodic phrase boundary can be described as the speaker's use of certain acoustic cues to signal the internal structure of the spoken message. These cues are acoustically realised as slower speech before a prosodic phrase boundary, a change in the fundamental frequency, and/or as a shorter or a longer silence interval. These changes help the listener to divide the message into chunks, as well as interpreting other prosodic information such as emphasis or focus (Gee & Grosjean, 1983). However, it is not obvious what should be regarded as a prosodic phrase boundary, nor how many levels of boundary strength that should be used. The insufficient research within this area is reflected by the many different prosodic transcription systems (Carlson et al., 2002).

When discussing the wider notion of prosodic phrasing, not only the demarcative cues that signal prosodic phrase boundaries are of interest. The coherence cues signalling that the words belong together as a group, are as important as the boundary cues (Bruce et al., 1992). However, this thesis will concentrate on the demarcative cues of the phrase boundaries.

2.1. Location of prosodic phrase boundaries in different communicative situations

Prosodic phrase boundaries are differently distributed in different communicative situations. A communicative situation can be described as the environment in which the speech takes place. Usually, we differ between spontaneous and non-spontaneous speech, but there are also different levels of spontaneity between these two extremes. The genuine non-spontaneous speech is for example fluent read-aloud speech, without any extra words, hesitations, coughs, interruptions or other disfluencies. An example is a professional reader reading a book aloud in a silent room. On the other extreme we have the total spontaneous speech, where the message is structured online and all of the above mentioned disfluencies occur frequently.

When reading aloud, a stronger prosodic phrase boundary usually occurs sentence-final, and a weaker boundary after certain clauses or phrases, while the stronger boundaries in spontaneous speech more often (than in non-spontaneous speech) occur before disfluencies such as hesitations, or prosodic emphasis (Carlson et al., 2002).

Hansson (1998) showed that even if the pauses in spontaneous speech have a freer distribution than in read-aloud speech, they are not inserted randomly. In her investigation of a 9 minutes long monologue by a Swedish speaker, the perceived pauses tended to occur at sentence boundaries (25%), after discourse marker or conjunction (30%), before accented content words (18%), and 27% at other locations. It also showed that only 57% of the total number of analysed sentences were preceded by a pause.

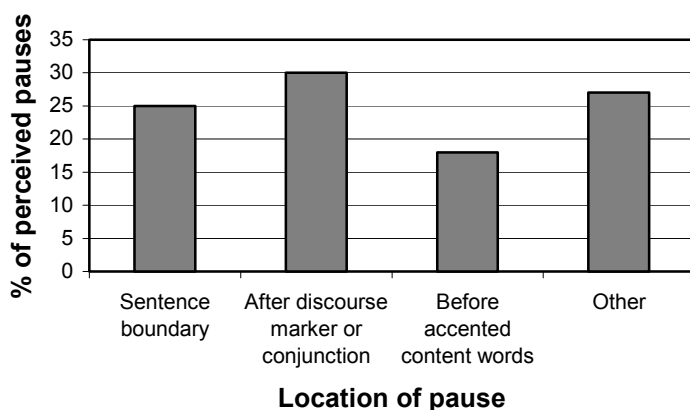


Figure 2.1. Pause distribution in spontaneous speech. Other locations are at clause boundaries, before adverbial phrases and phrases expressing negations, between repeated words, and before appositions and lexical hedges or fillers (Hansson, 1998).

2.2. Acoustic realisation of prosodic phrase boundaries

The acoustic representation of a prosodic phrase boundary and its adjacent words shows primarily three different acoustic phenomena: final (or pre-boundary) lengthening, silence intervals, and changes in the fundamental frequency. Each of the three phenomena is briefly described below. Other acoustic phenomena can also be involved, for example creaky voice and changes in amplitude (Wang and Hirschberg, 1992).

The length of the silence interval, final lengthening and f_0 -changes highly depend on how strong the boundary is. Also speaker depending features such as speaking rate, how much time the speaker needs to plan the rest of the message and so on have impact on the durations.

Final lengthening

The final lengthening is characterised by a lower speaking rate at the end of the word preceding the boundary. Usually, it is the final syllable rhyme that is longer than normal, i.e. the segments (or phones) from the last vowel to the end of the word.

The final lengthening is influenced by boundary strength. Horne et al. (1995) measured the duration of the word ‘procent’ preceding boundaries of different strengths, and found that the duration differed between approximately 520 and 600 msec. Furthermore, final lengthening is progressive, meaning that the final consonant was lengthened more than the preceding vowel. However, the greatest differences in the duration of the whole word appear when comparing focus and non-focus, where the focal word was between 30 and 125 msec longer than the non-focal word, depending on boundary strength.

An investigation on how lengthening is distributed in focal words (compared to non-focal words) has been done by Heldner & Strangert (2001). These data can to some extent be transferred to final lengthening. The whole word duration is about 25% longer than in non-focal position, and it is the stressed syllable that are most lengthened. Accent II syllables are usually more lengthened than accent I syllables. Table 2.1 below shows that it is the long speech segments (both vowels and consonants) that are most affected:

Accent	Syllable type	Approximate lengthening in focal position	
I	V: C	V: 24%	C: 16%
I	V C:	V: 11%	C: 23%
II	V: C	V: 18%	C: 26%
II	V C:	V: 6%	C: 37%

Table 2.1. Lengthening of stressed syllables in non-focal and focal position (Heldner & Strangert, 2001).

Furthermore, Heldner and Megyesi (2003) found that word-final rhymes before weak boundaries were longer than before long boundaries in Swedish.

Pauses

A silence interval (or a pause) marking a prosodic phrase boundary has usually a duration of 100 or more msec. The length of the pauses differs among speakers according to speaking rate, speaking style a.s.o. An investigation done by Fant et al. (2003) showed that pauses make up 25% of the total reading time when reading prose, while it is only 10% in news reading.

An experiment performed by Strangert (1990) showed that if the length of the silence intervals at paragraph boundaries is 1, the silence at sentence boundaries is 0.6, at clause boundaries 0.2, and at phrase boundaries somewhat lower than for clause boundaries.

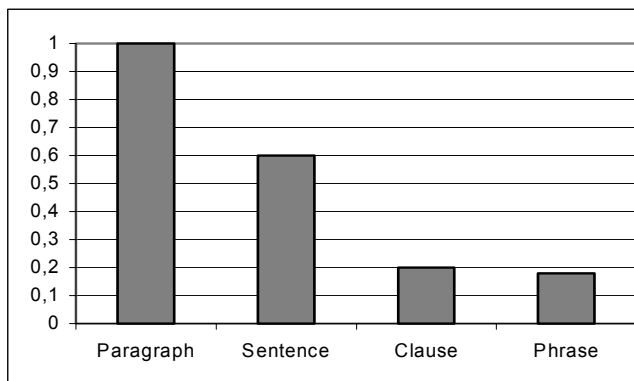


Figure 2.2. The relative length of silence intervals at different syntactic boundaries (Strangert, 1990).

She also showed that speakers differentiate a lot in their use of silence intervals. In a text with 810 words read by four different speakers, one of them used a silence interval every 18th word (totally 145), and another only every 9th word (totally 73).

Fant et al. (2003) point out that there are larger individual differences in how people pause *within* sentences than *between* sentences when reading aloud. Furthermore, the number of syllables has impact on the length of the silence interval. In their experiment, where 5 subjects read prose text, the average duration of a silence interval *between* sentences was 900 ms at 10 syllables' length, and 1300 ms at 40 syllables. Sentence *internally*, a sentence with less than 16 syllables was generally produced without any pauses, but if the sentence consisted of more than 20 syllables, there was usually one or more pauses. Clause boundaries had a pause with an average of 400-500 ms, and boundaries at lower syntactical levels around 100-200 ms, when reading prose text.

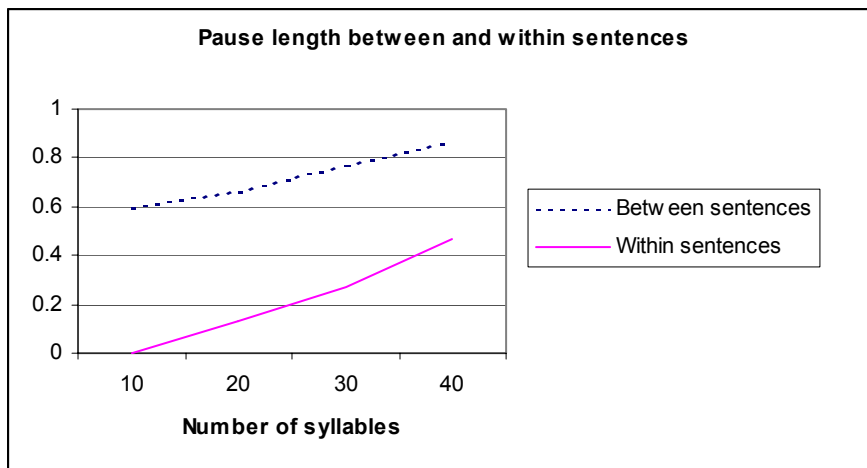


Figure 2.3. The relative pause length between and within sentences depending on number of syllables in the sentence (Fant et al., 2003).

Horne et al. (1995) also investigated how the boundary strength influenced the durations of the silence intervals. Here, focus did not affect the duration of the pause as it did for final lengthening. However, the silence interval increased from about 100 to 900 msec depending on boundary strength.

Fundamental frequency variation

Variations in the fundamental frequency occur both before and after the actual boundary. The pre-boundary f_0 -variation differs depending on the type of prosodic phrase boundary. So does the post-boundary f_0 -change, but most frequently it is a resetting, i.e. f_0 is the same as when starting an utterance, of the fundamental frequency, directly after the boundary. The pre-boundary change can be realised as different tones, high (H), mid (M) or low (L). [Reference](#).

2.3. Perception of prosodic phrase boundaries

According to Strangert (1990), both pre- and post-boundary information are important cues when perceiving prosodic phrase boundaries, even if most people use the silence interval itself as the strongest cue. People perceive silence intervals even where there is no such silence. In another experiment, she showed that it is possible to differentiate between different types of syntactic boundaries on the basis of the pre- and post-boundary cues alone (Strangert, 1992). The subjects listened to sentences containing the same words, but with different prosodic structures, and were asked to identify which kind of boundary (sentence/clause/phrase) they heard. The results showed that the subjects made 76% correct categorisations when all three cues were present, but it is also possible to differentiate between different boundaries on the basis of pre-boundary information alone (63% correct categorisations). When the three types of information were mixed up, so that they were in conflict with each other, the results showed that the silence interval itself was the strongest cue, but all three of them were taken into account when trying to categorise the boundary. There was also a difference across the subjects; some of them responded more on pre- and post-boundary cues, even though most of them used the silence interval as the strongest cue. Thus, only pre-boundary information cues alone give a fairly good categorisation, but it is the silence interval that is the most important cue, even when it is in conflict with other prosodic cues.

Bruce et al. (1992) performed another interesting experiment, showing that subjects may use different listening strategies when locating prosodic phrase boundaries. The majority of the

test subjects seemed to rely on a combination of f_0 and durational cues, but some primarily relied on the durational cues and others on f_0 -cues. There were more “duration-minded” subjects than there were “ f_0 -minded”. When one subject from each of the two groups were asked to produce sentences with the same meanings, the “durational-minded” subject used durational cues in a greater extent than the “ f_0 -minded” subject did. However, there was no such difference in the use of f_0 . This might suggest that regarding durational cues, there is a correspondence of people’s production and perception, even if there is no clear evidence for talking about to “prosodic types” of speakers.

2.4. Automatic prediction of prosodic phrase boundaries

In this section, some earlier approaches of automatic prediction of prosodic phrase boundaries will be presented. The results of automatic boundary prediction are highly dependent on the size of the training material, as well as of communicative situation. The best results have, not surprisingly, been obtained by studies of read speech by professional speakers (Wang and Hirschberg, 1992).

2.4.1. Punctuation symbols

At a first thought, one might think that prosodic phrase boundaries are equivalent to punctuation symbols, for instance a weak boundary at an orthographic comma, and a strong boundary at an orthographic full stop. Black and Lenzo (1999) point out that rules based solely on punctuation symbols can in fact be a good predictor of prosodic phrase boundaries in English and probably many other languages, since a punctuation symbol almost always includes a boundary. However, many prosodic phrase boundaries occur where no punctuation exists. Furthermore, the transcriptions of spontaneous speech do not usually contain any punctuation symbols. This means that a predictor based only on punctuation is not likely to over-generate, but the boundaries at other places with punctuation symbols will be ignored. Furthermore, the use of punctuation symbol is not homogenous among writers, and the boundary perception may depend on the reader’s individual punctuation habit (Steinhauer and Friederici, 2001). Their psycholinguistic investigation explored how the human brain processes commas, and how the individual use of commas has impact on their boundary perception.

2.4.2. Content words and function words

The distinction between content and function words are frequently used to predict prosodic phrase boundaries for speech synthesis. The words are divided into content words (usually nouns, proper nouns, verbs, adjectives, and sometimes participles) and function words (the remaining categories.). Many speech synthesis systems rely mainly on only this distinction, for example Black and Lenzo (1999), who suggest very simple rules to predict utterance internal boundaries. In English, as in Swedish, most function words are located before the words they relate to, which means that a boundary is more likely to occur between a content word and a function word. They suggest a rule that inserts a prosodic phrase boundary between a content word and a function word if it is located more than five words from the last punctuation symbol. Silverman (1987) used an even simpler rule, where a phrase boundary was inserted before every function word that follows a content word.

Heldner and Megyesi (2003) showed that the stronger the boundary is, the higher is the probability that the word preceding the boundary is a content word.

2.4.3. Part-of-speech tags

Sanders and Taylor (1995) used a British English corpus (SEC, the Spoken English Corpus) annotated with part-of-speech tags and prosodic boundaries to develop a statistical model for

phrase boundary prediction. The experiments were based on PoS trigrams, and it turned out that these were most useful for predictions of boundaries between the second and third word in the trigram, i.e. between the current and the following word. They used a tag set with 10 different tags (adjective, adverb, noun, determiner, subjunction & conjunction, preposition, pronoun, auxiliary verb, main verb and other). The baseline in this corpus is 78%, i.e. 78 out of 100 words were not followed by a prosodic boundary.

Five methods were tried out. The first method used only the PoS trigram, inserting a phrase break if the probability was over a certain threshold. This resulted in an accuracy of 89%, and 22 boundaries were inserted.

In the second method, the fact that the location of the last prosodic boundary has impact on the predictions was taken into account. The probabilities of the distance to the last prosodic boundary were added to the part-of-speech trigrams (phrase length / total number of phrases). The probabilities for a prosodic boundary to occur increases with the length of the phrase; the more words since the last boundary, the higher probability for that a boundary should be inserted. The trigram and distance probability is then multiplied, and if they are over the threshold, a boundary is inserted. 27 boundaries were now inserted, but the accuracy decreased to 87%.

The third method looks L trigrams ahead, where L is a relatively long sequence of words. The trigram with the highest probability is found, and if it is higher than the threshold, a prosodic boundary is inserted and the procedure starts again. This method gives an accuracy of 89%, and 16 boundaries were inserted.

The fourth method is a quite computational expensive one. All possible combinations (except between the first two words) of boundaries and non-boundaries in a sentence are calculated. This means that a sentence of 22 words will cause over a million computations, although only the part-of-speech trigram probabilities are used. This results in 14 inserted boundaries, and an accuracy of 90%.

The fifth and last method first inserts prosodic phrase boundaries at every punctuation mark. Next, the same exhaustive method as in the fourth method is performed, but now in a smaller area. The accuracy is now 87% and there were 26 inserted breaks.

The authors point out that the trigram and distance probabilities can be combined in many different ways; to simply multiply them are not optimal for the purpose.

Taylor and Black (1998) also investigated how part-of-speech tags can be used in order to predict prosodic phrase boundaries. The baseline in the test corpus was approximately 80%, and they worked with the distinction boundary or no boundary. Using only punctuation symbols for prediction resulted in an accuracy of 90.76%. Adding the rule that a boundary should be inserted after every function word that follows a content word highly decreased the accuracy, to 70.29%, suggesting that this rule tend to over-generate. When they used part-of-speech bigrams of current word and next word, as well as the rules for inserting boundaries at punctuation symbols, the accuracy was 91.10%. Note that the performance of the tagger is 94.4%. Using an n-gram of 6 part-of-speech tags decreased the accuracy to 88.01%. Using both the punctuation symbol rule and the content – function rule, plus the part-of-speech bigram probabilities gave an accuracy of 91.10%, i.e. the same as without the content – function rule. However, replacing the bigram with an n-gram of 6 part-of-speech tags resulted in a higher accuracy than without the content-function rule, 89.39%.

Phrase boundaries occur at regular intervals, making the probability for a boundary to occur higher the more number of words there are between the location and the last boundary (Taylor and Black, 1998). Many people have investigated this phenomenon, but since the last boundary might be incorrectly located, it is not always possible to compute the distance.

The size of the tag set is also an important question (Taylor and Black, 1998). If using part-of-speech trigrams to predict prosodic boundaries, it is necessary that the tag set is small enough or working with a training material of satisfactory size to give enough occurrences of each trigram.

Heldner and Megyesi (2003) used the content/function word distinction together with part-of-speech tags to predict prosodic phrase boundaries in a Swedish radio interview (spontaneous speech). The accuracy of no boundary was 72.6%, and of weak and strong boundaries 14.0 and 63% respectively. This makes a total of 66.3% correctly predicted words.

2.4.4. Syntax

Also the syntactic parsing can be used for prosodic boundary predictions. However, Black and Lenzo (1998) point out that even if syntax might be a strong correlate of prosodic phrasing, it is often not worth using it, since most parsers are not reliable enough for the purpose.

Lindström et al. (1996) used a dependency-based syntactic representation for predicting prosodic phrase boundaries, where the head of the phrase is central. A part-of-speech tagger is used to build the dependency graphs, and a dependency-oriented grammar identifies the head-modifier relationships, which contributes to disambiguate the utterance. They do not present any results of how this system performs.

3. Predicting prosodic phrase boundaries

The preceding chapter described earlier research within the area. This chapter presents the work that was done for the present study; how the prosodic phrase boundaries were predicted given a linguistically analysed text, validated and tested, and finally implemented in a system that takes any text as input and produces a sound file as output.

This first section gives a general introduction to the work with the prosodic phrase boundary prediction, since it is necessary to have an overall impression of how the different procedures work before we go into details. The work with the prediction of prosodic phrase boundaries mainly consisted of three parts:

1. *Statistical calculations and rule production:* Statistics over the localisations of prosodic phrase boundaries was calculated from a pre-analysed training material. In addition, a rule set with rules derived from the statistics for boundary prediction was written.
2. *Validation session:* The statistical output and the rules from (1) were validated on material similar to the training material to find out which statistical data and which rules that were most useful for boundary prediction.
3. *Test session:* A small listening test with synthetic speech and the acoustically realised predicted boundaries was performed to evaluate whether the insertion of the prosodic phrase boundaries made the synthetic speech more intelligible. This part of the study is described in more detail in chapter 4.

Before any of these sessions could begin, a lot of effort was put into collecting speech data, transcribing the audio data and annotating prosodic phrase boundaries, performing acoustic and linguistic analyses, and so on. These preparations are presented in section 3.1. The statistical calculations and the rules in step 1 above are described in section 3.2 and 3.3 respectively. In section 3.4, the validation procedure of step 2 is presented, and the results from the validation session are shown in section 3.5. The acoustic realisations of the predicted prosodic phrase boundaries are described in section 3.6. Finally, section 3.7 presents the program that takes any text as input, and produces a sound file with synthetic speech and the predicted prosodic phrase boundaries acoustically realised as output. Figure 3.1 on the next page shows the procedures the input text string passes on its way to the final output, a sound file. Each step and the components involved are briefly described below. The tagger (TnT), parser (SPARK), some of the scripts used to compute the other linguistic analyses, lexicon (CentLex) as well as the tools used for synthesising (MBROLA¹) and creating the intermediate file for manipulating the durations (Ivxtalk) were developed before this study. All the other procedures are developed in connection with this study.

1. *Text normalisation*

A very simple text normalisation, mainly tokenisation.

2. *Tagger*

The normalised string is sent to tagger (TnT) and each word is tagged with its part-of-speech category and morphological analysis.

3. *Parser*

The tagged text is sent to a rule-based chart-parser, SPARK, and each word is marked with the parse category the word belongs to at each node of the parse tree.

¹ MBROLA is a free tool for concatenative synthesising of recordings of any language and voice (Filipsson & Bruce, 1997).

4. *Other linguistic analyses*
This includes features as part-of-speech bigrams, phrase depth, content/function word, number of words since last boundary etc. Every word is now tagged with about 15 different linguistic features.
5. *Predict boundaries*
Statistics and rules are used to predict the localisations and, to some extent, the acoustic realisations.
6. *Lexicon lookup*
The words are looked up in a lexicon (CentLex) to get their phonemic representations.
7. *.PHO file generation*
A file where it is possible to change the durations of phones and silence intervals is generated by Ivxtalk, a tool from Infovox.
8. *Insert durations*
A Perl script lengthens phones that should have final lengthening and inserts silence intervals as predicted. This is done by increasing the total durations of the phones, and by inserting silence of various durations between the phones. This results in a new .PHO file.
9. *Synthesise*
The new .PHO file is synthesised by MBROLA.

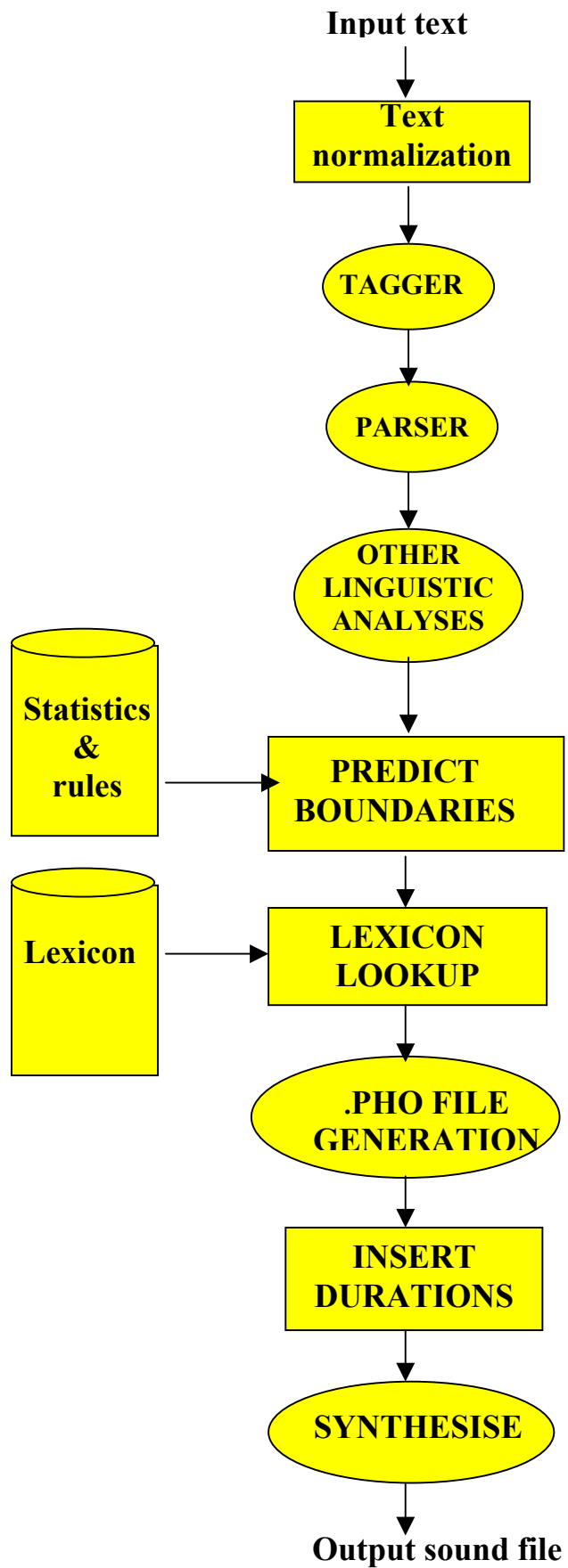


Figure 3.1. The procedures from text file to sound file with synthetic speech and prosodic phrase boundaries realised.

3.1. Data preparation

This section describes the training, validation and test material. The training material and the radio interview used for validation was collected and for the most part analysed before, but the validation and test material was prepared for this study.

3.1.1. Audio data

The training material, on which the statistics was calculated, consists of 2422 tokens from a news broadcasting from Sveriges Radio read by professional readers, distributed on 25 parts of monologue speech.

The news used for validation consists of 956 tokens, and was collected from www.sr.se/ekot. It is a 7 minutes long news broadcasting sound file. The radio interview used for validation consists of 2743 words distributed as 23 chunks of monologue-like speech from a radio interview (“Lördagsintervjun”). This orthographic transcription does not contain any delimiters at all.

These three recordings were orthographically transcribed and linguistically analysed (see 3.1.2 below). An acoustic analysis was done on the training news and the interview recordings, measuring durational and intonational phenomena. Three experienced transcribers listened to the recordings and marked where they perceived a weak or strong boundary. The majority vote of the three transcribers is considered as the gold standard² when comparing the predicted boundaries with the perceived boundaries in the validation procedure.

The test news is not a recording, just 62 sentences of the same type as the training and validation news, taken from www.sr.se/ekot. Some of them will be synthesised with and without the new boundary rules and used for a listening test. The text is only linguistically analysed. In the testing procedure, some sentences designed to be of various complexity are also used.

Table 3.1 below shows some statistics of the different recordings/texts. Note that the number of perceived weak and strong boundaries are almost the same in the non-spontaneous news texts, while the number of perceived weak boundaries in the radio interview are clearly higher than the strong boundaries. Furthermore, the number of disfluencies in the interview differs from the news data, confirming that hesitations are more common when the speech is less planned. The 3-way agreement for perceived boundaries shows for how many words and delimiters the three transcribers perceived the same boundary type (weak, strong or no boundary respectively).

² The material the predictions will be compared to when evaluating the predictions in the validation session, where the predicted boundaries are compared to the perceived ones. *Gold standard* means that it is not certain that the answers are correct; in this case it is impossible to decide if the transcribers perceived the boundaries ‘correctly’.

Text	Training News	Validation News	Validation Interview	Test News
No of words	2232	885	2743	1045
No of sentences	162	50	23 ³	62
No of speaking chunks	39	13	23	13
No of major delimiters	162	50	-	62
No of minor delimiters	28	21	-	37
No of tokens (words + delimiters)	2422	956	2743	1144
No of perceived strong boundaries	146	52	126	-
No of perceived weak boundaries	147	51	340	-
No of perceived no boundaries	2128	853	2277	-
No of disfluencies	0	2	70	-
No of disfluencies/no of words	0	0,022 ‰	2 ‰	-
No of words/sentence	13.78	17.7	119 ⁴	16.85
3-way agreement for perceived boundaries	92%	94%	91%	-
Annotation	acoustic, linguistic, perceptual	linguistic, perceptual	acoustic, linguistic, perceptual	linguistic

Table 3.1. Overview annotated recordings. No of words is without punctuation symbols.

3.1.2. Linguistic analysis

The linguistic analysis consists of 15 linguistic features that were expected to be usable for boundary prediction. Most features are derived from a tagged or parsed text, and a couple of features show how far from the last boundary the word is. Table 3.2 shows an overview on the linguistic features, which are described in more detail in the following sections.

content/function	Indicates if the word is a content word (proper noun, noun, verb, adjective or participle) or a function word (remaining part-of-speech classes and punctuation symbols).
pos-L	The part-of-speech tag of the word to the left.
pos	Part-of-speech of current word.
pos-R	The part-of-speech tag of the word to the right.
pos-L0	Part-of-speech bigram: left context and current word.
pos-0R	Part-of-speech bigram: current word and right context.
pos-L0R	Part-of-speech trigram: left context, current word and right context.
phrase-parse	All nodes in the parse tree belonging to the current word.
phrase-depth	Number of nodes from the current word to the top node of the parse tree.
phrase-depth-R	Phrase depth of the right context's word.
phrase-depth-0R	Bigram of phrase depth of current word and right context.
phrase-depth-diff	The difference between the phrase depth of current word and right context.
last_any_boundary	Number of words since last boundary (weak or strong).
last_strong_boundary	Number of words since last strong boundary.
phrase_firstlast	The node closest to the word in the parse tree, followed by the node directly under S.

Table 3.2. Description of the linguistic features used in the analysis.

³ Since the spontaneous material does not contain any delimiters, this figure is the same as the number of speaking chunks (or utterances).

⁴ This figure shows the number of words per speaking chunk.

3.1.3. The tagger - TnT

The tagger is a data-driven part-of-speech (PoS) tagger developed by Thorsten Brandt 2000. It was adapted to Swedish in 2000 (Megyesi, 2002). The accuracy was then 95.31% for all words (96,53% for known words and 88.24% for unknown words).

The tags are later converted into other PoS-categories elaborated for the present purpose. These part-of-speech tags are presented in table 3.3.

Tag	Description	Example	Tag	Description	Example
AF	participle	ansedd	N		
AQ	adjective	fin	NC	common noun	mat
CC	conjunction	och	NP	proper noun	Sven
CIS	infinitive marker	att	PF	personal pronoun	jag
CS	subjunction	att	PH	wh-pronoun	vem
DT	determiner	den	PS	possessive pronoun	min
DTR	relative determiner		Q	particle	in
FE	major delimiter	.	RG	adverb	inte
FI	minor delimiter	,	RH	relative adverb	när
I	interjection	hej	SP	preposition	på
MC	cardinal	tre	V	verb	springa
MO	ordinal	tredje			

Table 3.3. Part-of-speech tags.

3.1.4. The parser - SPARK

The parser used for the linguistic analysis is a rule-based chart-parser, developed by Megyesi (2002). The following categories are used:

Phrase type	Description	Example
ADVP Adverb Phrase	adverbs that can modify adjectives or numeral expressions	mycket (very)
AP Minimal Adjective Phrase	the adjectival head and its possible modifiers	mycket intressant (very interesting)
APMAX Maximal Adjective Phrase	more than one AP with a delimiter or a conjunction in between	mycket intressant och trevlig (very interesting and nice)
NUMP Numeral Expression	numerals with their possible modifiers, for example AP or ADVP	åtskilliga tusentals (several thousands)
NP Noun Phrase	the head noun and its modifiers to the left	Pilgers mycket intressanta och trevliga bok (Pilger's very interesting and nice book)
NPMAX Maximal Projection of NP	one or more NP(s) with following PP(s) and possible modifier	Pilgers mycket intressanta och trevliga bok om politik (Pilger's very interesting and nice book about politics)
PP Prepositional Phrase	one or several prepositions delimited by a conjunction and one or several NPs or NPMAXs, or in elliptical expressions an AP only	om politik (about politics)
VC Verb Cluster	a continuous verb group belonging to the same verb phrase without any intervening constituents	skulle ha varit (should have been)
INFP Infinitive Phrase	an infinite verb together with the infinite particle and may contain ADVP and/or verbal particles	att gå ut (to go out)
REL	relative clause	..., som kör bilen,... (who is driving the car)

Table 3.4. Phrase parse categories used by the SPARK parser for Swedish (Megyesi, 2002).

Three different kinds of output are available from the parser; indented output, tagged output, and a graphical output. The graphical output also contains the part-of-speech tags for each word. The parse tree for the sentence ‘*Pilgers mycket intressanta och trevliga bok om politik ligger på bordet.*’ (‘Pilger’s very interesting and nice book about politics is on the table that is green.’) is presented on the next page.

In the tagged version, each parsed word has a BIO tag, that is a tag that marks if the word is **B**eginning of the phrase, **I**nside the phrase or **O**utside the phrase. Examples of words that are outside the phrase are delimiters and interjections.

USA	NPB
tänker	VCB
öka	VCI
pressen	NPB_NPMAXB
på	PPB_NPMAXI
Iran	NPB_PPI_NPMAXI
i	PPB
kampen	NPB_NPMAXB_PPI
mot	PPB_NPMAXI_PPI
terrorismen	NPB_PPI_NPMAXI_PPI
.	O

Table 3.5. Parsing information of the sentence ‘USA will increase the pressure on Iran in the struggle against terrorism’ with IBO tags.

The error rate of the parser is estimated to between 6% and 11% with 98% confidence. Most errors are due to problems with prepositional phrases in maximal projections of NPs (Megyesi, 2000). For example, in the sentence in table 3.5, it can be discussed if the phrase ‘pressen på Iran’ (the pressure on Iran) really should belong to the same NPMAX, or if the correct parse is that ‘pressen’ is a noun phrase of its own and ‘på Iran’ just a prepositional phrase.

For boundary prediction, it is also useful to know if the word is the last in its phrase (assuming that a boundary is more likely to occur at the end of a longer phrase). To capture this, the end token “I” of the last word of a longer phrase (at least three words) was replaced with “E” for **E**nd. The parse tags are now BIOE:

USA	NPB
tänker	VCB
öka	VCI
pressen	NPB_NPMAXB
på	PPB_NPMAXI
Iran	NPB_PPI_NPMAXE
i	PPB
kampen	NPB_NPMAXB_PPI
mot	PPB_NPMAXI_PPI
terrorismen	NPB_PPI_NPMAXE_PPE
.	O

Table 3.6. Parsing information of the sentence ‘USA will increase the pressure on Iran in the XXX against terrorism’ with IBOE tags.

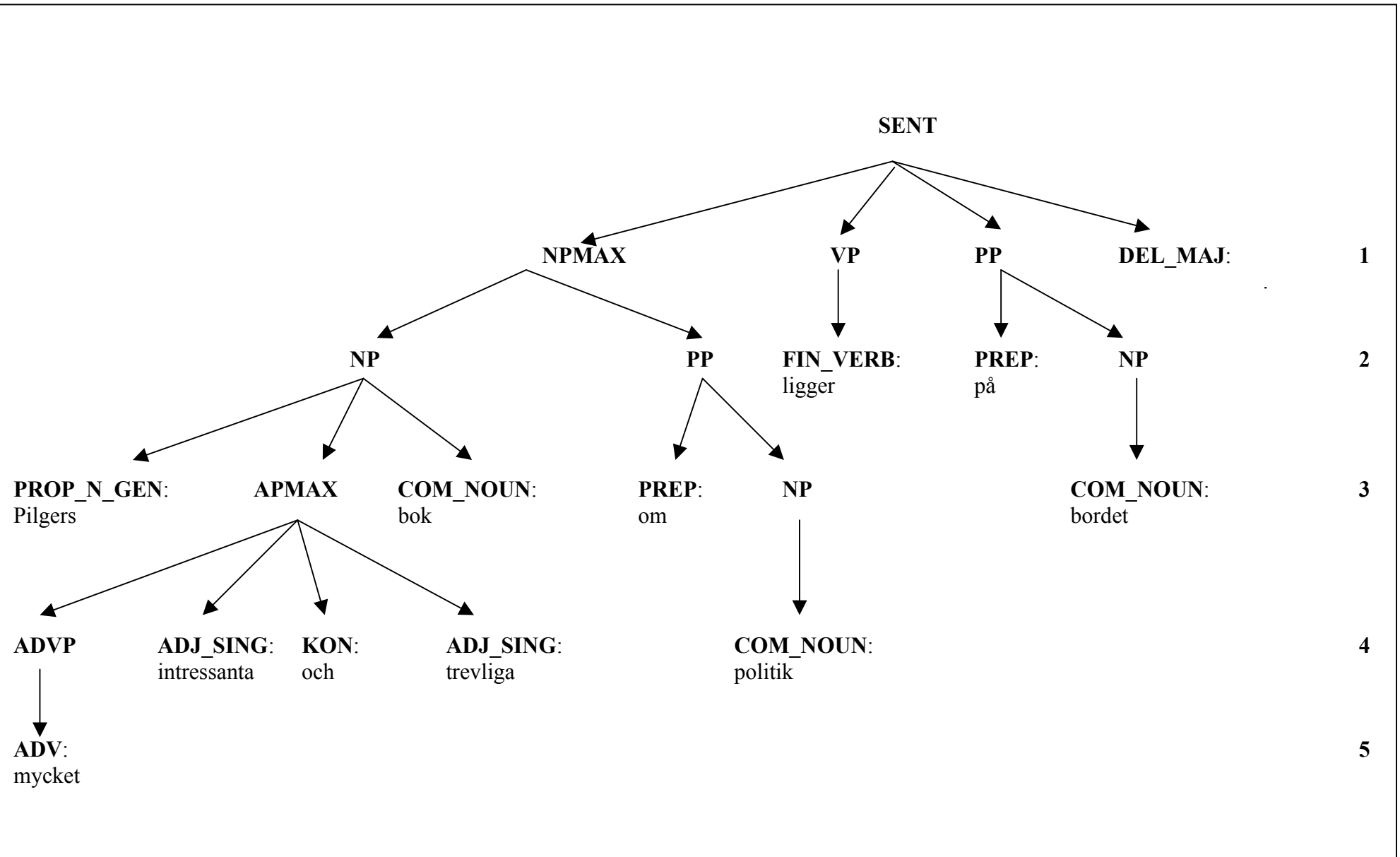


Figure 3.2. Phrase parse tree of the sentence "Pilgers mycket intressanta och trevliga bok om politik ligger på bordet". (Pilger's very interesting and nice book is on the table)

3.1.5 Other linguistic features

The other linguistic features are mainly derived from the output of the tagger or the parser.

From the tagged output, bigrams (L-0 and 0-R, where L stands for left context, 0 current tag and R right context) and trigrams (L-0-R) are computed. The part-of-speech tags are also converted into content or function words, where nouns, proper nouns, verbs, adjectives and participles are grouped into content words, and the remaining categories into function words.

From the parsed output, the phrase depth is computed. The phrase depth reflects how many nodes from the top node the word is. For example, the phrase depth for the word ‘*trevliga*’ (nice) in the parse graph above is 4, and for ‘*Pilgers*’ 3. There is also a field for the right context, and the bigram 0-R, and another for the difference of the phrase depth in the bigram (current word’s phrase depth – next word’s phrase depth). The hypothesis is that the greater differences between the phrase depths of the current word and next, the higher probability for a boundary to occur between them. The phrase depth is shown to the right in the parse tree above.

Not all features used for boundary prediction can be generated before the actual prediction starts. These features have mainly the function to prevent over-generation, i.e. that a boundary is inserted too soon after the last predicted boundary. The features `last_any_boundary` and `last_strong_boundary` count how many words there is between the last boundary (weak or strong) and the current location, and the distance from the last strong boundary and the current location. They are inserted on the fly; when a weak boundary is inserted, the `last_any_boundary` counter is set to 0, and when a strong boundary is inserted, both counters are set to 0.

3.2. Statistical calculations on the training material

Input: linguistically and perceptually annotated news text

Output: a file with the probabilities for a weak, strong or no boundary occurring with the different linguistic features

A Perl script that calculates the probabilities for the three boundary types occurring at certain linguistic features was used to create the statistics file, which will be used later for boundary prediction. The following formula describes how the probabilities are calculated where f is a certain linguistic feature value, for example the phrase depth 1, $\# perceived\ boundary\ type$ is the total number of strong, weak or no boundaries occurring after this feature value, and $\# tokens$ the total number of words and delimiters:

$$p^f = \frac{\# perceived\ boundary\ type}{\# tokens}$$

Table 3.7 below shows the probabilities for a certain boundary occurring with a content or a function word. The delimiters are treated as values of a type of their own, since they are not actually words. The first column contains all values that occurred in the training material for each feature, in this case content words, function words, major or minor delimiters. The second and third columns show the probability and the total number of occurrences in the training material when a strong boundary follows the function/content word or the minor/major delimiter. Column 4-7 contain the probabilities and occurrences for a weak or no boundary. The next column is to check that the sum of the three probabilities is 1, and the last column shows the total number of occurrences for this certain feature value.

1 Probabilities content/function

value	p strong	# strong	p weak	# weak	p no	# no	p sum	# sum
C	0.008	10	0.067	84	0.925	1166	1	1260
F	0	0	0.025	46	0.975	840947	1	971
FE	0.832	134	0.112	18	0.056	9	1	161
FI	0.065	2	0.677	21	0.258	8	1	31

Table 3.7. Probabilities and occurrences of the linguistic feature content/function words and the delimiters (rounded probabilities).

The table suggests that if the word is a content word (C), it is not likely to be followed by a boundary, and if it is, it is a small preference for weak boundaries. If the word is a function word (F), it was never followed by a strong boundary in the training material. The figures for the delimiters show that 80% of the major delimiters (FE) are followed by a strong boundary, and 68% of the minor delimiters (FI) are followed by a weak boundary.

The number of values for the different features in the training material differs from 4 for content/function word to 789 for part-of-speech trigrams (LOR). Obviously, many of the 789 part-of-speech trigrams occur only once in the training material, and are therefore not useful for the prediction.

When predicting boundaries from the statistics, the linguistic features can be combined and weighted in different ways. Furthermore, a lower limit for how many times a feature value must occur to be considered can be set, to prevent over-generations. This will be described in more detail in the section about the validation procedure below.

3.3. Rules derived from the statistics

As pointed out above, since the training material is of a small size, many of the linguistic feature values occur only once, which might lead to over-generations. Furthermore, all of the most common linguistic contexts are not present in the training material, which means that no boundaries will ever be predicted for unknown linguistic contexts. To make the predictions more reliable and homogeneous, some rules partly derived from the statistics were written. For some linguistic features, it is quite easy to see which linguistic feature values that differ from the others in the statistics file (see examples in Appendix B). The most obvious example is the delimiters, which almost always involve a weak or strong boundary. Other statistical data has too few occurrences, or does not show a clear pattern among the values, for example the part-of-speech feature, which is useful only in combinations with other features.

The rules are of two types; constraint rules, which prevents over-generations, and rules that predicts boundaries.

Many different rules and rule combinations, with and without the statistical calculations, were tried out. Below follows a description of the rules that are used in the final script. They are applied before the statistics, i.e. if none of the rules is used, the script makes predictions from the statistical data. The Perl code is somewhat simplified to more clearly show what is happening.

Delimiter rules

First, a rule that predicts a weak boundary at minor delimiters and a strong boundary at a major delimiter was applied, since they are the feature values that have the highest probabilities for prosodic phrase boundary insertion. The content/function feature or one of the part-of-speech features could be used to capture the delimiters, but I prefer to write a rule for this to keep the delimiters apart from the other values of the features. To avoid that a word immediately before a delimiter is assigned a boundary (which would cause two boundaries in a row), a constraint rule that prevents this was added.

Constraint rules

In order to prevent that a boundary is inserted between words that always belong together as a group, two constraint rules that were inserted before the other rules were added. The first rule says that there should never be a boundary if the word is part of an infinitive phrase.

The second rule prevents a boundary to be predicted if the part-of-speech bigram of the word to the left and the current word is V-Q, i.e. a verb followed by its particle.

Parsing rules

Next, rules using the information about the parse tree were written, suggesting that boundaries occur after the end of longer phrases (ending with 'E' in the parse, see above). To avoid boundaries after every single phrase, conditional rules checking how many words from last boundary the current word is, and how many words there are until the next delimiter (minor or major), where a boundary is already predicted by the first rule. Three words in both directions gave the best results.

End of relative clause

The first rule concerns relative clauses; *if* the word is the last in a relative clause, *and* we are more than three words from the last delimiter, *and* it is more than three words to the next delimiter, a weak boundary is inserted:

```
} elsif ($parse =~ /RELCL/ && $bcounter > 3 && $del > 3) {  
    $boundary = "WEAK";  
}
```

End of prepositional phrase and noun phrase

The second rule checks if the word is both the last in a prepositional phrase *and* a noun phrase. The phrase depth difference must be more than 1 to confirm the boundary relevance, and as in the previous rule, we must be at least three words from the last boundary and three words from the next delimiter:

```
} elsif ($parse =~ /(NPE.*PPE)|(PPE.*NPE)/ && $depthdiff > 1 && $bcounter >  
3 && $del > 3) {  
    $boundary = "WEAK";  
}
```

Conjunctions and subjunctions

The next rule inserts a weak boundary before a conjunction or subjunction *if* the phrase depth of the conjunction is 0 and we are more than three words from the last predicted boundary was added. This rule is useful for sentences with more than one main clause, where the clauses are connected with a conjunction such as 'och' (and) and 'men' (but). The phrase depth rule assures that the 'och' occur between clauses, and is not part of a noun phrase such as 'äpplen och päron' (apples and pears):

```
} elsif ($posR eq "CC" && $depthR == 0 && $bcounter > 3) {  
    $boundary = "WEAK";  
}
```

Phrase depth bigrams

The last rule mainly concerns the phrase depth bigram of the current word and the next. In the statistics file, the probabilities for any boundary are relatively high for the listed bigrams. Adding the constraints of how far we must be from the last boundary and until the next delimiter, and that the word must be a content word prevents too many over-generations:

```
} elsif ($depthOR =~ /^(00|01|02|03|41|51|61|40|50|60)$/ && $bcounter > 3  
&& $del > 3 && $contfunc eq "C") {
```

\$boundary = "WEAK";

If none of the rules is applied, the pure statistical data is used to do further predictions.

3.4. Validation procedure

Input:

- the statistics from the training procedure above (chosen linguistic features and weights), and/or the rules derived from the statistics or the more general rules
- linguistically and perceptually annotated text
- a comparison file with the words and perceived boundaries (the gold standard)

Output:

- the text marked with the predicted boundaries, as well as the perceived boundaries for comparison
- statistical information on the agreement of the predicted and perceived boundaries (accuracy), the precision, recall and f-scores

The validation procedure evaluates the statistics or rules, to see which method that gives the best results. The validation material is of the same type as the training material (news broadcasting) and consists of 885 words, linguistically and perceptually analysed. The predicted boundaries are compared to the perceived boundaries. Accuracy and f-scores are calculated, which is a common way to measure the performance of NLP systems such as information extraction, taggers and parsers.

The *accuracy* simply describes how many percents of the words that achieved the correct boundary (strong, weak or no boundary).

Precision shows how much of the answers from the system that was correct.

$$\textit{Precision} = \frac{\textit{number of correct answers given by the system}}{\textit{total number of answers given by the system}}$$

Recall shows how much relevant information that is covered by the system, i.e. how many of the perceived boundaries of this type were found by the system.

$$\textit{Recall} = \frac{\textit{number of correct answers given by the system}}{\textit{total number of correct answers in reference}}$$

F-score shows the balance between precision and recall. The β parameter is used to weight precision and recall (a value under 1 favours the precision, over 1 the recall). In this case, β is 1, which means that both values are equally important.

$$F_{\beta=1} = \frac{(\beta^2 + 1) * \textit{Precision} * \textit{Recall}}{\beta^2 * \textit{Precision} + \textit{Recall}}$$

3.4.1. Prediction using statistics

This section explains how the statistics over the training material is used in order to predict prosodic phrase boundaries.

Linguistic features and weights

As mentioned earlier, it is possible to choose which linguistic features the predictor will take into account when predicting prosodic phrase boundaries with the statistics, as well as how important these features should be. This can be done using weights, that decide the

importance of this feature. Given the linguistic features we would like to test as keys, and the weights for each of these features as values, the Cartesian cross product is calculated, and all possible combinations of features and weights are processed:

```
4 2
8 1
11 8.5
```

This hash table looks at feature number 4, 8 and 12, which correspond to the part-of-speech tag of the word to the right, the parse and the phrase depth bigram in the following combinations with their weights:

```
8-1 4-1 11-5
8-3 4-1 11-5
8-1 4-2 11-5
8-3 4-2 11-5
8-1 4-1 11-8.5
8-3 4-1 11-8.5
8-1 4-2 11-8.5
8-3 4-2 11-8.5
```

Since approximately 89% of the words (including delimiters) are *not* followed by a boundary, the probability for no boundary is almost always higher than for a boundary to occur after the word. When the probability for a boundary is higher, it is almost always a feature that occurs less than five times in the training material. To allow boundaries at other places, the probabilities must be weighted. This is done by increasing the number of occurrences of weak and strong boundaries in a certain context. For instance, the occurrences of a certain linguistic feature are 15, 25 and 60 for a strong, weak or no boundary respectively (hypothetical figures). To force the predictor to mark the word as being followed by a boundary, the summarised occurrences of weak and strong boundary are multiplied with its weight (let's say 3), and new probabilities are calculated. The probability for any boundary to occur with this feature value is now higher than before (0.67 instead of 0.40), and higher than for no boundary (0.67 compared to 0.33). The word is now predicted to be followed by a boundary. To decide if the boundary should be weak or strong, the original probabilities for a weak or strong boundary are compared ($0.15 < 0.25$). As one might suspect, it is almost always the weak boundary that has a higher probability than the strong in this case.

value	p strong	# strong	p weak	# weak	p no	# no	p sum	# sum
X	0.15	15	0.25	25	0.60	60	1	100
value	p strong + weak		# strong + weak		p no	# no	p sum	# sum
X	$120 / 180 = 0.67$		$15 + 25 * 3 = 120$		$60 / 180 = 0.33$	60	1	180

Table 3.8. Example of how the weights manipulate the probabilities. The occurrences of strong and weak boundaries have been multiplied with 3, and new probabilities are calculated. The original probabilities of weak and strong boundaries are then compared to decide which of them that should be predicted.

A weight of 1 does neither increase nor decrease the number of occurrences of boundaries, but the probabilities of the feature are taken into account. This has impact on the results for features where at least one of the values has a probability lower than 0.5 for no boundary, e.g. the delimiters in the content/function feature. Boundaries will be inserted when the probability of a value is above 0.5. If the probability for no boundary is higher than 0.5, the accuracy is the same as the baseline, since no boundaries at all are predicted. If the probability for no boundary is exactly 0.5, no boundary is predicted.

A weight that is lower than 1 consequently decreases the number of boundary occurrences. Again, this has no effect on features where the probability of no boundary is higher than 0.5

for all values. Instead, this low weight can be useful for a certain feature in combination with other features; it can function as a constraint where the other features tend to over-generate.

Occurrence limit for feature values

The number of occurrences of a certain context is also important to take into account when dealing with the statistics. One occurrence of a certain linguistic feature value where a strong boundary was perceived would exclusively predict strong boundaries for all future texts. For a larger training text, the lower limit for how many occurrences of a certain linguistic feature value there must be could be set higher, but since the training data is small, this limit was set to 5, possibly causing some over-generation. However, this limit can be changed when the training material is larger.

3.4.2. Results from the validation procedure using statistics

The statistical output shows the results of all chosen combinations of linguistic features and weights. Table 3.9 shows an example where the content/function feature has been weighted with 1, 5 and 50. The second column, Average F-score, shows the average of the f-scores for strong, weak or no boundary, to quickly get a hint of the overall results. The second column shows the accuracy, i.e. how many words that had correct predictions. The third, fourth and fifth columns show the f-scores for strong, weak and no boundary respectively, the following three columns precision and the last three the recall.

fe-we	Av. F	Acc	F - S	F - W	F - 0	P - S	P - W	P - N	R - S	R - W	R - 0
1-1	0.79	95.50	0.96	0.43	0.98	0.98	0.70	0.96	0.94	0.31	0.99
1-5	0.79	95.50	0.96	0.43	0.98	0.98	0.70	0.96	0.94	0.31	0.99
1-50	0.41	17.89	0.96	0.11	0.16	0.98	0.06	1.00	0.94	0.98	0.08

Table 3.9. Statistical output from the content/function feature with different weights. F = f-score, P = precision, R = recall, S = strong boundary, W = weak boundary, N = no boundary.

In this case, the weights 1 and 5 did not affect the predictions at all, since the figures are the same as predictions using the rules for delimiters only. Multiplying the number of occurrences for weak and strong boundaries with 50 leads to over-generations; all words are predicted to be followed by a weak boundary. The recall for weak boundaries is therefore much higher, while the recall figures for strong or no boundaries dramatically increase. The figures for strong boundaries are the same, since they still are inserted only at major delimiters.

Running all features alone with the weight 1 did not lead to better results for any feature. All features except three (pos-R, pos-0R and pos-L0R) led to the same results as running with the delimiter rules alone.

As could be expected, the largest problem is that the training material is too small. For example, the parse NPE_PPE_RELCLE_NPMAXE indicates that the word is the last in four phrases, which one might think should be a quite strong reason to insert a prosodic boundary. However, this parse occurs only four times in the training material, and hence no boundary is ever predicted for this context.

Finding the most useful features

All features were tried out independently and in combinations with other features. First, one feature at a time was tested with different weights, to see which features that could improve the results alone. Next, several combinations of features and weights were tested. Some of the features could not be used of their own, but could function as complements to other features. An example is the content/function feature, which used alone could insert weak and strong boundaries at minor and major delimiters, but is not useful if the weight is high enough to

insert boundaries at every content or function word. The best result was achieved when using the parse feature weighted 1.2 in combination with the phrase depth bigram with a relatively high weight, 9. The occurrences of a boundary to occur between two words with certain phrase depth are thus multiplied with 9. Combining these two features with the part-of-speech tag of the word to the right weighted 2 further improves the results. This suggests that a parsing analysis and the phrase depth feature derived from the parsing information is an important feature for automatic boundary prediction, even if the training material is small and do not have enough occurrences of certain values to yield for as many context as would be desirable.

Table 3.10 shows how the accuracy, f-score, precision and recall increased when using the delimiter rule and the statistics only.

	Av. F	Acc	F - S	F - W	F - 0	P - S	P - W	P - N	R - S	R - W	R - 0
1	0.33	89.23	0.00	0.00	0.98	0.00	0.00	0.97	0.00	0.00	1.00
2	0.81	95.71	0.96	0.48	0.98	0.98	0.68	0.96	0.94	0.37	0.99

Table 3.11. Average f-score, accuracy, f-score strong, f-score weak, f-score no, precision strong, precision weak, precision no, recall strong, recall weak, recall no, after adding the delimiter rules and statistical calculations.

3.5. Results from the combination of rules and statistics

Below follows some figures where the improvements in f-score, precision and recall are shown after adding certain rules or statistical calculations.

The numbers of the Y-axis show which rules or statistics that is used. Note that the rules are always applied in this order. The figures show how the results change after *adding* a rule to the rule set, not how this single rule alone changes the results.

- 1) Baseline
- 2) Delimiter rules
- 3) Constraint rules
- 4) End of relative clause
- 5) End of NP and PP
- 6) Next word is a conjunction and has phrase depth 0
- 7) Phrase depth bigram and content word
- 8) Statistics

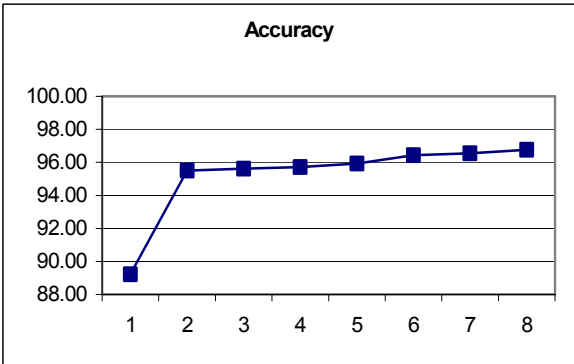


Figure 3.3. Accuracy for all words after adding rules and statistical calculations.

Figure 3.3 above clearly shows that the delimiter rules create the greatest improvements. The accuracy increases from approximately 89.23% (baseline) to 95.5%, meaning that 60 words in the validation material has now correct predictions. The other improvements increase the accuracy with between 0.10 to 0.52 percentage units, which corresponds to between one and five words. This might seem marginal, but the changes are relevant, and can be assumed to lead to better predictions in a larger validation material.

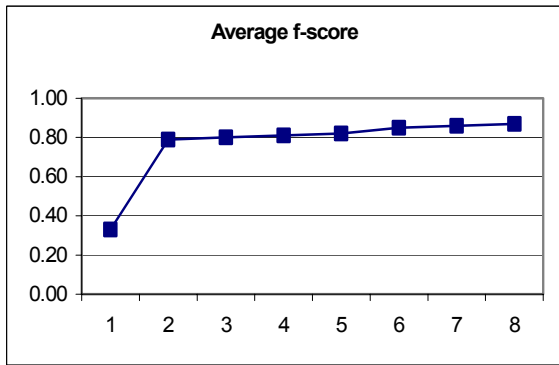


Figure 3.4. Average f-score after adding rules and statistical calculations.

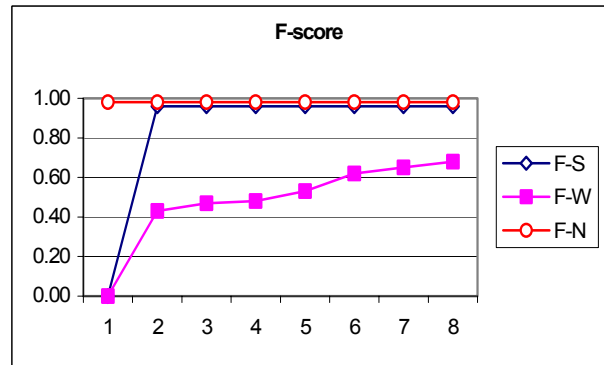


Figure 3.5. F-scores for strong, weak (F-W) and no (F-N) boundary after adding rules and statistical calculations.

Figure 3.4 above shows the average of the f-scores for each boundary type (strong, weak, no), which are shown individually in figure 3.5. It is undoubtedly the strong boundaries that had the greatest benefit from the delimiter rules. In fact, it is the only step where the f-score for strong boundaries were changed, since all the other rules only predict weak boundaries, and the probabilities for strong boundaries are rarely higher than for the weak boundaries. So, the f-score for the strong boundaries stays at 0.98, like the f-score for no boundary, which is 0.98 all the time. However, it is the weak boundary that is the most interesting boundary type. Inserting a boundary at every minor delimiter increases the f-score from 0 to 0.43. Especially the rule for words which finalise a prepositional phrase and a noun phrase improves the results (from 0.53 to 0.62).

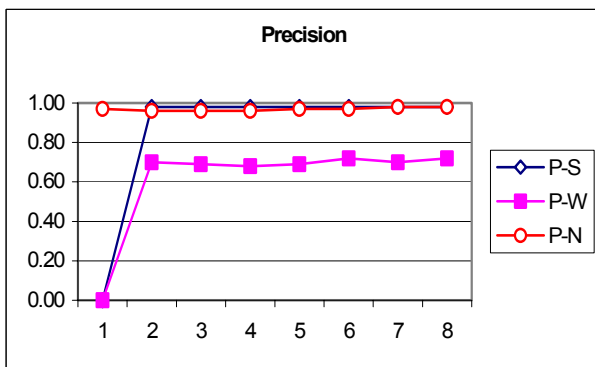


Figure 3.6. Precision for strong (P-S), weak (P-W) and no (P-N) boundary after adding rules and statistical calculations.

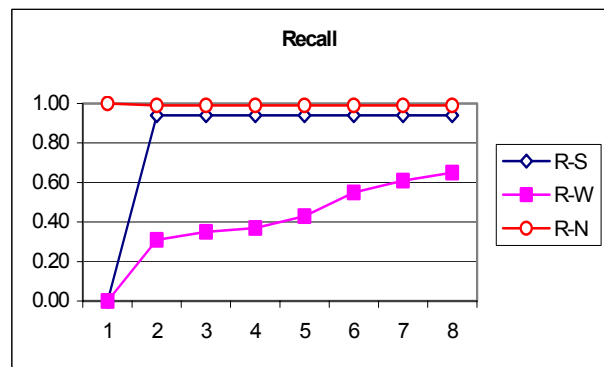


Figure 3.7. Recall for strong (R-S), weak (R-W) and no (R-N) boundary after adding rules and statistical calculations.

The tables above show the precision and recall for each boundary type. As explained earlier, the precision shows how many of the predictions that are correct. An example from figure 3.6 above is that approximately 70% (0.68-0.72) of the total amount of predicted weak boundaries

were correct. The remaining 30%, which were weak boundaries predicted where no boundary was perceived, consists mostly of weak boundaries inserted at minor delimiters.

The recall shows how much relevant information that is covered by the system, i.e. how many of the perceived weak boundaries were predicted as such. Figure 3.7 indicates that the system found more of the perceived weak boundaries as more rules and statistical calculations were added to the system (from 0 for baseline, 0.31 after the delimiter rules to 0.65 after the last rule).

The precision and recall for the strong boundaries increased to 0.98 and 0.94 respectively after the delimiter rules, and stayed there during the additional rules and statistical calculations. The precision for no boundary was 0.97 at baseline, and then changed between 0.96 and 0.98. The recall for no boundary started at 1.00 at baseline where no boundaries were inserted, and stayed at 0.99.

	Av. F	Acc	F - S	F - W	F - 0	P - S	P - W	P - N	R - S	R - W	R - 0
1	0.33	89.23	0.00	0.00	0.98	0.00	0.00	0.97	0.00	0.00	1.00
2	0.79	95.50	0.96	0.43	0.98	0.98	0.70	0.96	0.94	0.31	0.99
3	0.80	95.61	0.96	0.47	0.98	0.98	0.69	0.96	0.94	0.35	0.99
4	0.81	95.71	0.96	0.48	0.98	0.98	0.68	0.96	0.94	0.37	0.99
5	0.82	95.92	0.96	0.53	0.98	0.98	0.69	0.97	0.94	0.43	0.99
6	0.85	96.44	0.96	0.62	0.98	0.98	0.72	0.97	0.94	0.55	0.99
7	0.86	96.55	0.96	0.65	0.98	0.98	0.70	0.98	0.94	0.61	0.99
8	0.87	96.76	0.96	0.68	0.98	0.98	0.72	0.98	0.94	0.65	0.99

Table 3.11. Average f-score, accuracy, f-score strong, f-score weak, f-score no, precision strong, precision weak, precision no, recall strong, recall weak, recall no, after adding rules and statistical calculations.

The same rules and statistics were applied on the radio interview validation material. Since the interview does not contain any delimiters (except for the disfluencies, which are tagged as minor delimiters), the only rule that inserts strong boundaries at major delimiters were not used at all. Instead, all boundaries were considered weak. The baseline, where no boundaries at all are inserted, had a precision of 83.01. Applying the boundary predictions, this figure increased to 86.67. The incorrect predictions occurred frequently before conjunctions in utterances as ‘...medlemmar vågar ha en egen uppfattning **men** när man har att ta...’ (...members dare to have an own opinion **but** when you have to...), which actually could be a nice thing when generating synthetic speech.

	Av. F	Acc	F - S	F - W	F - 0	P - S	P - W	P - N	R - S	R - W	R - 0
1	0.30	83.01	0.00	0.00	0.91	0.00	0.00	0.83	0.00	0.00	1.00
8	0.38	86.67	0.00	0.23	0.91	0.00	0.58	0.85	0.00	0.14	0.98

Table 3.12. Average f-score, accuracy, f-score strong, f-score weak, f-score no, precision strong, precision weak, precision no, recall strong, recall weak, recall no, after adding rules and statistical calculations on the material from the radio interview.

Figure 3.8 show that the error rate for the news material increased from 10.77 to 3.24 when applying the rules and statistics for boundary predictions, i.e. an improvement of 70%. The error rate for the interview material decreased from 16.99 to 13.33, which improves the error rate with 78%. These figures are not comparable, since the perceived strong boundaries were replaced with weak ones in the interview. However, the figure might say that rules and statistics elaborated for news material are not as appropriate for spontaneous speech, but could function as a base for further investigations about predicting prosodic phrase boundary in spontaneous speech.

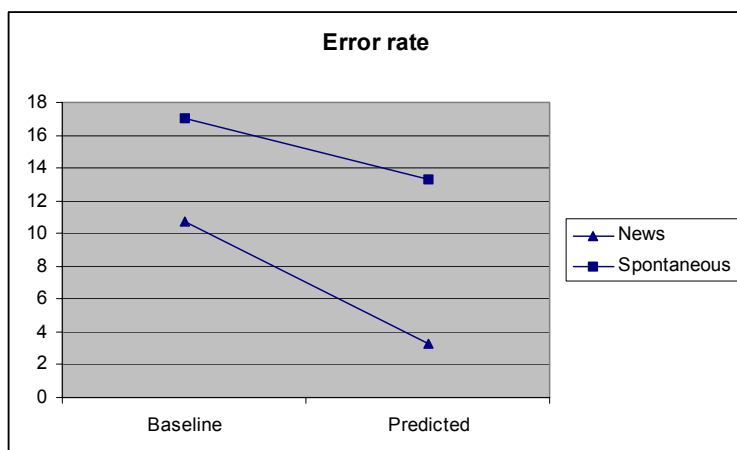


Figure 3.8. Error rate for news material and radio interview after adding rules and statistics.

The results show that the most important rule is the one that inserts prosodic phrase boundaries at delimiters. The following rules and the statistics mainly affect the weak boundaries, since strong boundaries are inserted only at major delimiters. The f-score for weak boundaries are improved from 0.43 after the delimiter rule to 0.68 when all rules and the statistics are applied. It is mainly the recall that is improved, i.e. how many of the weak boundaries that were found by the system. The same rules and statistics did also improve the results also on the material derived from spontaneous speech, though not as much as for the news material, which they were elaborated to handle.

3.7. Acoustic realisation of prosodic phrase boundaries

In this study, the acoustic realisations of the prosodic phrase boundaries only include final lengthening and silence intervals. The acoustic analysis of the training news text is not used, since it is not part of the task, and would clearly exceed the time limit. Instead, very simple realisations are used, to get a hint of how the final result could be.

The realisations of the prosodic phrase boundaries are added to the lexicon, and thus looked up in the same way as the ordinary phonetic transcriptions of the words. The output from the boundary predictor simply contains the orthographic words with codes for final lengthening and silence intervals marked. The phonematic transcription with codes for the prosodic phrase boundaries for the sentence ‘*En man som tidigare varit fängslad, har nu släppts fri.*’ (A man who earlier was put in prison, has now been released.) looks like this:

```
'EN M'AN S']M T"I:DIGARE0 V"A:RIT F"[NGSLAD FL_3 SI_100 H'A:R
N'U: SL'[PTS FR'I: FL_1 SI_400
```

FL stands for final lengthening, and is followed by its degrees: small or large. In this case the codes are *FL_3* and *FL_1*, meaning that the rhyme of the word should be lengthened. *FL_3* involves more lengthening than *FL_1*, and is currently used before weak boundaries, as suggested by Heldner and Megyesi (2003). *SI* shows that a short or long silence interval should be inserted, followed by the length of the silent interval. *SI_100* indicates a weak boundary and *SI_400* a strong boundary. When the fundamental frequency data has been processed, codes for the intonation changes can be added.

Before every silence interval, there is always a final lengthening of the last word. The rules for final lengthening are often very complex, but for this purpose, very simple rules have been used. They do not reflect how final lengthening is realised in the “real” rule file of the

synthesis. The final lengthening is applied on the rhyme of the word preceding the silent interval, i.e. it starts at the last vowel of the word and continue to the last phone of the word. The rules check if the phone is a vowel or a consonant, the length (short or long), if it is stressed or unstressed and finally, if the syllable has accent I, accent II or is unstressed. The original durations are multiplied with values derived from the lengthenings described in table 2.1 above. A long stressed accent II consonant achieves the largest lengthening, while a short stressed accent I vowel achieves the smallest lengthening. Again, these lengthenings are not claimed to be optimal, but are calculated to give a rough representation of final lengthening.

As described in section 2.2, final lengthening is greater if the word is in focus, but since there is no information about focus in the texts, this was left out of account. Perhaps lengthening content words more than function words could reflect this to some extent, but it seemed as a too general solution and was not tried out.

3.8. The final system used for prediction and realisation of prosodic phrase boundaries

Input:

- any text
- the statistical probabilities and the combination of linguistic features and weights
- the rule set

Output:

- linguistically annotated text with the predicted boundaries marked
- phonetically transcribed text with information about where and how the prosodic phrase boundaries should be realised
- sound file with the predicted boundaries realised

This section describes the steps in figure 3.1 in more detail, where a text string is converted into a sound file with synthetic speech and the predicted prosodic phrase boundaries are acoustically realised.

En man som tidigare varit fängslad, har nu släppts fri.

Text normalisation

This simple text normalisation does nothing but tokenising. For a real application, this step must be replaced with a better text normalisation module.

En man som tidigare varit fängslad , har nu släppts fri .

Tagging

The text string is converted to the correct input format for the tagger, one word/line, and the TnT tagger described in section 3.1.3 assigns each word its part-of-speech Parole tag.

En	DI@US@S
man	NCUSN@IS
som	PH@000@S
tidigare	RGCS
varit	V@IUAS
fängslad	AF0USNIS
,	FI
har	V@IPAS
nu	RG0S
släppts	V@IUSS
fri	AQPUSNIS
.	FE

Parsing

The SPARK parser parses the text, and produces an output file with the word followed by its PoS tag and parse:

```
En/DI@US@S_NPB_NPMAXB man/NCUSN@IS_NPI_NPMAXB
som/PH@000@S_NPB_RELCLB_NPMAXI tidigare/RGCS_ADVPB_RELCLI_NPMAXI
varit/V@IUAS_VCB_RELCLI_NPMAXI fängslad/AF0USNIS_APMINB_RELCLI_NPMAXI
, /FI_O har/V@IPAS_VCB nu/RG0S_O släppts/V@IUSS_VCB fri/AQPUSNIS_APMINB
./FE_O
```

Other linguistic analyses

The phrase depth of each word is calculated:

```
En      2
man     2
som     3
tidigare 3
varit   3
fängslad 3
,       0
har     1
nu      0
släppts 1
fri     1
.       0
```

The words are tagged as content or function word. The tags for delimiters are kept as FI or FE:

```
En      F
man     C
som     F
tidigare F
varit   C
fängslad C
,       FI
har     C
nu      F
släppts C
fri     C
.       FE
```

The other linguistic features are derived from the information we already have (PoS, parse, phrase depth and content/function). All data is summarised and printed to a file. Column 0 lists the word, 1 content/function, 2 the PoS of the left word, 3 PoS of current word, 4 PoS of the right word, 5 the PoS bigram of the left and current word, 6 the PoS bigram of current and right word, 7 the PoS trigram of left, current and right word. In column 8 is the parse, 9 the phrase depth, 10 the phrase depth of the right word, 11 the phrase depth bigram of current and right word, and column 12 the phrase depth difference. The 13th feature, the first and last parts of the parse is calculated later, as well as the features that counts number of words after the last boundary (any or strong).

0	1	2	3	4	5	6	7	8	9	10	11	12
En	F	0	DT	NC	0-DT	DT-NC	0-DT-NC	NPB_NPM AXB	2	2	22	0
man	C	DT	NC	PH	DT-NC	NC-PH	DT-NC-PH	NPI_NPMA XB	2	3	23	-1
som	F	NC	PH	RG	NC-PH	PH-RG	NC-PH-RG	NPB_REL LB_NPMA XI	3	3	33	0
tidigare	F	PH	RG	V	PH-RG	RG-V	PH-RG-V	ADVPB_R ELCLI_NP MAXI	3	3	33	0
varit	C	RG	V	AF	RG-V	V-AF	RG-V-AF	VCB_REL LI_NPMA XI	3	3	33	0
fängslad	C	V	AF	FI	V-AF	AF-FI	V-AF-FI	APMINB_R ELCLI_NP MAXI	3	0	30	3
,	FI	AF	FI	V	AF-FI	FI-V	AF-FI-V	O	0	1	01	-1
har	C	FI	V	RG	FI-V	V-RG	FI-V-RG	VCB	1	0	10	1
nu	F	V	RG	V	V-RG	RG-V	V-RG-V	O	0	1	01	-1
släppts	C	RG	V	AQ	RG-V	V-AQ	RG-V-AQ	VCB	1	1	11	0
fri	C	V	AQ	FE	V-AQ	AQ-FE	V-AQ-FE	APMINB	1	0	10	1
.	FE	AQ	FE	0	AQ-FE	FE-0	AQ-FE-0	O	0	0	00	0

Table 3.13. The summarised linguistic features of the sentence 'En man som tidigare varit fängslad, har nu släppts fri'.

Predict boundaries

Given the set of features, the prosodic phrase boundaries can be predicted. The file with the statistical data from the training material is read, and the rules and statistics are applied as described in section 3.4. The output is a text string with the final lengthening and silent intervals marked:

```
En man som tidigare varit fängslad ~large %weak har nu släppts fri ~small
%strong
```

Lexicon lookup

To get the phonemic transcriptions of the words, a TMH-internal lexicon, CentLex, is used. It consists of approximately 408 000 entries, and contains the graphemic representation of the word, part-of-speech tag and morphological analysis, information about frequency, and one or more phonematic transcriptions in RULSYS format.

The lexicon lookup function first tries to find the word with correct part-of-speech tag. If it is not found, it looks only for the graphemic representation. Since the phonematic transcriptions are sorted by the probability of pronunciation, it is always the first transcription that is chosen.

Some missing phonematic transcriptions have been added to the lexicon, and some incorrect transcriptions have been changed for the listening test. To get information about the boundaries, some entries for final lengthening and silent intervals were added:

```
%strong      100      PAUSE      SP_400
%weak        100      PAUSE      SP_100
~small       100      FINAL      FL_1
~large       100      FINAL      FL_3
```

These entries translate the boundary information in the text string into codes that later will be used to do the acoustic realisation of the prosodic phrase boundaries.

```
'EN M'AN S']M T'I:DIGARE0 V"A:RIT F"!NGSLAD FL_3 SP_100 H'A:R N'U: SL'!PTS  
FR'I: FL_1 SP_400
```

A hash list which indicates the localisations of the prosodic phrase boundaries is produced, to be used later when inserting the new durations.

.PHO file generation

The phonematic transcription without the codes for the boundaries is sent to Ivxtalk, which create a .PHO file. This file shows the individual phones together with their durations and some information about f_0 -movements.

Insert durations

A Perl script inserts the silent intervals and their durations into the .pho file. It also finds the rhyme of the words that should have final lengthening, and the durations of these phones are lengthened as described in section 3.7.

Synthesise

The new .PHO file is sent to mbrola, which creates a .wav sound file.

4. Evaluation test

An evaluation listening test was performed to find out whether the synthetic speech was easier to follow with prosodic phrase boundaries acoustically realised at more locations than only at delimiters. The test form is presented in appendix C, and the synthesised material in appendix D.

4.1. Test method

The test was performed on 10 listeners, divided into two groups. The five listeners in group 1 were familiar with synthetic speech and the five in group 2 were inexperienced listeners.

The test material consists of 16 sentences of different complexity and origin:

5 complex sentences from news text

5 complex sentences from Pisoni (an existing test consisting of short stories of different complexity)

2 short sentences from news text

2 short sentences from Pisoni

2 chunks of a radio interview

All sentences were synthesised twice; once with the prosodic phrase boundaries acoustically realised according to the rules and statistics described in section 3.2 and 3.3 and once with rules only for delimiters. The 32 sound files were randomised in two different ways, to avoid that the order in which the sound files were presented affects the judgements.

The short sentences contain between 8 and 13 words. The longer sentences contain between 21 and 40 words, and most of them differ in number of acoustically realised boundaries. Most of them contain one or more delimiters (in addition to the full stop at the end of the sentence). The two parts from the radio interview contain between 58 words each. Since there are no delimiters in the orthographic transcription of this kind of speech, the sound files synthesised with the delimiter rules only have no prosodic phrase boundaries at all acoustically realised.

The sound files were presented to the listener one at a time, and they were asked to grade the grouping of the phrases on a scale 1 to 5. They were told to try to disregard things like wrong f_0 -contour, over-explicit pronunciations and so on. To get used to the test, they were first presented to five sound files that were not part of the test.

4.2. Test results

Three utterances did not differ in number of phrase boundaries. Two of them were too short to achieve a boundary, and the third was a longer utterance which contained a comma, and thus had a boundary in both synthesising. Sometimes the same utterance had been graded as a 3 and 4, 4 and 5, but only one test subject had graded the same utterance with greater inconsistency; 3 and 1, and 3 and 5. However, since the average grade for each utterance was approximately the same, I chose not to take the inconsistency of one subject into account.

Figure 4.1 shows the average grades for the two different synthesising of short sentences, long sentences and interview text. Only the sentences where all rules and the statistics implied more prosodic phrase boundaries are included.

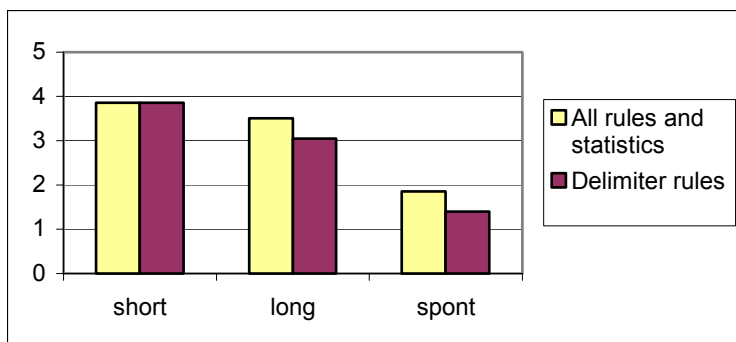


Figure 4.1. Average grades for the different types of sentences synthesised with all rules and statistics and only with delimiter rules.

Shorter utterances

Two short utterances differed when synthesised to the two versions. The average grade of the utterances synthesised in the standard way (with delimiter rules only) and all rules and statistics was exactly the same, 3.85. The experienced test subjects graded the standard version higher for both utterances, while the inexperienced test subjects thought that at least one utterance was better with one prosodic boundary realised.

Longer utterances

The benefits of more prosodic phrase boundaries are more clear when examining the results of the longer utterances. The average grade with all rules and statistics was 3.51, and 3.05 when the prosodic boundaries are realised only at delimiters. There is no significant difference between the experienced and inexperienced test subjects. Only one of nine utterances were judged lower when more boundaries were inserted (3.6 vs. 4.2), else the improvements for the utterances were between 0.25 and 1.75 grades higher. The utterance where the inserted boundaries made it sound worse included one comma, and hence a boundary in each versions. In the predicted version, one extra boundary was inserted, which apparently made it sound worse. The utterance which were most improved by more prosodic boundaries is the longest one (40 words), which in the standard version contained no boundary at all.

Utterances from spontaneous speech

The two utterances synthesised on the radio interview material were long and contained no punctuation marks. Even if the predictor was encompassed to news text, where punctuation marks are present, it improved the grades in the test. One of the utterances was graded 1.5 vs. 2.1, and the other 1.3 vs. 1.6. The grades are not high – none of the versions did not sound well – but at least it helped with some boundaries.

Figure 4.2 and 4.3 show how the experienced and inexperienced subjects differed in their judgement of the synthesising with all rules and statistics vs. only delimiter rules. There is a small tendency that the inexperienced listeners graded the synthesising with all rules and the statistics higher than the experienced listeners did.

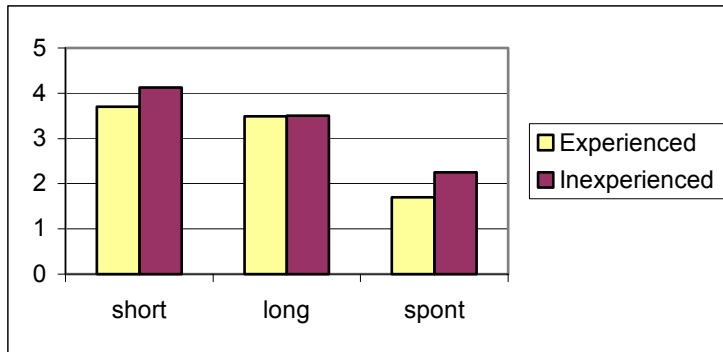


Figure 4.2. Difference between the experienced and inexperienced test subjects for sentences synthesised with all rules and statistics.

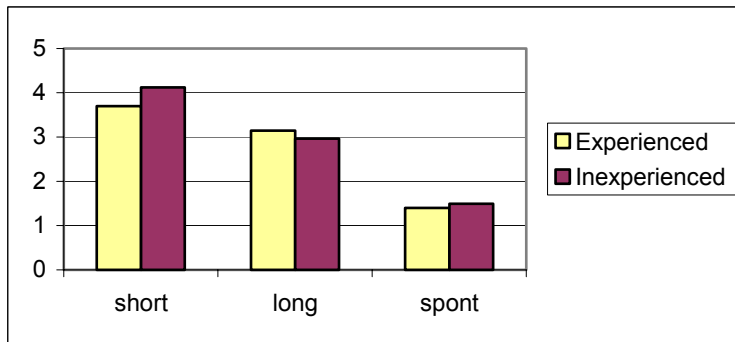


Figure 4.3. Difference between the experienced and inexperienced test subjects for sentences synthesised with the delimiter rules only.

5. Discussion and future research

This chapter discusses the results, what could have been done differently and suggestions about how the performance of the prosodic phrase boundary predictor might be improved.

Training and validation data

To get more reliable statistics, it is necessary with a lot more linguistically and perceptually annotated training data. Since the statistics is calculated from only 2232 words, many probabilities are based on few occurrences, and could not be used for prediction. Black and Lenzo (1999) points out that to get satisfactory reliable statistics, the training database must be big enough. For their system, they used the MARSEC database *roach93*, which consists of about 37,000 words. Also the validation material needs to be of bigger size; currently a change of 0.1 percentage units in the accuracy corresponds to just one more correctly predicted boundary. However, the test results showed that even a small training data set improves the intelligibility of the synthetic speech, also when the statistics is derived from news data and applied on spontaneous speech data.

Another problem is that the statistics and rules are based on speech from different newsreaders. Modelling the prosody of one single speaker would, as Hirschberg (2002) points out, be better, since the speech will be more homogenous and the statistics more reliable. She also means that using perceptual annotation as the golden standard may not be the optimal way for speech synthesis. The recordings contain hesitations and other disfluencies that are not appropriate to model in synthetic speech. She suggests syntactic or other cues as a base for prediction of prosodic phrase boundaries, rather than fluent human performance. Perhaps an intermediate approach could be a solution for this; the boundaries are annotated as 'real' boundaries only if the perception transcriber thinks they are syntactically motivated.

Furthermore, I do not believe that news reading is the optimal training material. The news text is read under time pressure, which leads to less prosodic boundaries than perhaps is desirable for speech synthesis, where it is even more important with a clear structure of the message. Prose read by one single professional reader would probably work as the most uniform and consistent training material.

Different communicative situations

Using the statistics and rules elaborated for news material for prosodic boundary prediction in interview did improve the synthetic speech to some extent. However, the mean grade for these two utterances was only 1.85 while using the statistics and rules, and 1.4 without any phrase breaks at all. This is mainly due to the fact that the interview is different from the news material. It consists of longer chunks of words since it lacks punctuation symbols marking the end of the sentence, as well as commas to indicate weaker boundaries. This shows that it is probably more advantageous to use a training material of the same type as you want to synthesise. This can be done using the same scripts as for the news material on the annotated speech from the radio interview.

Weights and limit for number of occurrences

Currently, the weights used with the statistics have the function to increase the number of occurrences of the weak and strong boundary. Another type of weight that decides which features that are generally more important than others could be tried out. All probabilities should then be multiplied with this weight, not only the probabilities for boundaries.

Linguistic analyses

When the training material is small, it is better to decrease the number of feature values. For this study, the part-of-speech tags trigrams and the parse feature involved many values with less than five occurrences. Using only the part of the parse tree that is closest to the word and the highest node of the same words (phrase first_last) did not seem to be of any significant use. Furthermore, syntactic parsing by rules is both space and time consuming and not reliable enough for speech synthesis (Sanders and Taylor, 1995). However, Taylor and Black (1998) thinks that syntactic parsing is helpful when predicting prosodic boundaries, and this might be done with a statistical parser.

Changing β_2 to favour the precision

For this purpose, it is more important that the precision is high, since we do not want to over-generate; it is better to insert too few boundaries than boundaries at wrong locations, especially since the news material includes major delimiters, which always involves a strong boundary. The f-scores could be forced to favour the precision by setting β_2 to a value less than 1. However, since the over-generations mainly occur at minor delimiters and the news are read under time pressure, I did not choose to weight the precision higher than the recall. Perhaps this could be tried out on other material in the future.

Acoustic realisations and the listening test

For the moment, only two levels of silent pauses are used, but it is possible to add more levels, and the realisations of the levels can easily be changed. Currently, a short pause is realised as a silence interval of 100 msec, and a long pause as 600 msec. The longer pauses are used only sentence final, i.e. at major delimiters. Furthermore, prosodic phrase boundaries do not always involve a silence interval (e.g. Hansson, 1998). At least one additional realisation of a prosodic phrase boundary, where only the pre- and post-boundary information is present is needed.

Listening test

The listening test showed that subjects thought that the statistics and rules based only on the news training material improved the synthetic speech, but some of the test subjects pointed out that the shorter pause was too short in some of the realisations, which shows that more levels of silence length are preferable. Furthermore, all subjects were more or less confused by the fact that the intonation and the word transitions were incorrect in many utterances. All subjects thought that they had graded the utterances with some inconsistency; that they were more indulgent when they got used to the synthetic speech. This was true for all subjects except one. The average grade for the first 16 utterances was 3.12 and 3.49 for the last 16 utterances. However, since the utterances were randomised in two different ways, it is possible that this can be disregarded.

It would have been better to synthesise the test utterances in a different way. Inserting commas at the weak boundaries and full stops at the strong boundaries and running the input sentences all the way through the synthesiser without stopping it to produce the .pho-file where the silent intervals and durations were inserted, would produce much better acoustic realisations of the boundaries and thus make the test easier for the subjects. However, the script is open to further improvements of the final lengthenings, and to some extent prepared to handle f_0 -movements when this data is available. The acoustic analysis of the training material could be used to investigate how the prosodic phrase boundaries should be realised in different contexts. Also f_0 -variations are part of the acoustic analyses, which would further improve the realisations of the boundaries. The GROG-project is currently investigating how these changes are realised, and they could be included in boundary prediction later.

6. References

- Black, Alan W. & Lenzo, Kevin A. (1999). Building Voices in the Festival Speech Synthesis System. Carnegie Mellon University.
- Bruce, Gösta; Granström, Björn & House, David (1990). Prosodic phrasing in Swedish speech synthesis. *Proceedings of the workshop on speech synthesis*. AuTRANS, France.
- Bruce, Gösta; Granström, Björn; Gustafson, Kjell & House, David (1992). Interaction of F0 and duration in the perception of prosodic phrasing in Swedish. *Nordic Prosody VI*.
- Carlson, Rolf; Granström, Björn; Heldner, Mattias; House, David; Megyesi, Beata; Strangert, Eva & Swerts, Marc (2002). Boundaries and groupings – the structuring of speech in different communicative situations: a description of the GROG project. *TMH-QPSR Vol. 43. Fonetik 2002*.
- Fant, Gunnar, Kruckenberg, Anita & Ferreira, Joana Barbosa (2003). Individual variations in pausing. A study of read speech. *PHONUM 9 (2003), 193-196*. University of Umeå.
- Filipsson, Marcus & Bruce, Gösta (1997). LUKAS – a preliminary report on a new Swedish speech synthesis. *Working Papers 46 (1997), 45-56*.
- Gee, James Paus & Grosjean, François (1983): Performance Structures: A Psycholinguistic and Linguistic Appraisal. *Cognitive Psychology 15 (1983), 411-458*.
- Hagström, Bo (2001). Utvärdering och utvidgning av en parser(grammatik) för svenska. C-level thesis, Dpt. of Speech, Music and Hearing, KTH, Sweden.
- Hansson, Petra (1998). Pausing in Spontaneous Speech. Lund University.
- Heldner, Mattias & Megyesi, Beata (2003). Exploring the Prosody-Syntax interface in Conversations. *ICPhS 15*. Barcelona, Spain.
- Heldner, Mattias & Strangert, Eva (2001). Temporal effects of focus in Swedish. *Journal of Phonetics (2001) 29, 329-361*.
- Heldner, Mattias (2001). On the non-linear lengthening of focally accented Swedish words. *Paper V in Focal accent – f₀ movements and beyond*. PHONUM 8. Umeå University.
- Hirschberg, Julia (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication 36 (2002) 31-43*.
- Horne, Merle, Strangert, Eva & Heldner, Mattias (1995). Prosodic Boundary Strength in Swedish: Final Lengthening and Silence interval Duration. *ICPhS 9, 170-173*. Stockholm, Sweden.
- Lindström, Anders, Bretan, Ivan & Ljungqvist, Mats (1996). Prosody Generation in Text-to-Speech Conversion Using Dependency Graphs. *ICSLP 96*.

Megyesi, Beata (2002). Data-Driven Syntactic Analysis. Methods and Applications for Swedish. Dpt. of Speech, Music and Hearing, KTH. Sweden.

Ostendorf, M. (1997): Prosodic Boundary Detection. In Horne, M. (ed.): *Prosody: Theory and Experiment. Studies Presented to Gösta Bruce*. Kluwer Academic Publishers.

Sanders, Eric & Taylor, Paul (1995). Using statistical models to predict phrase boundaries for speech synthesis. *Proceedings of EUROSPEECH '95*. Madrid, Spain.

Silverman, K. (1987): The structure and processing of fundamental frequency contours. University of Cambridge.

Steinhauer, Karsten & Friederici, Angela D. (2001). Prosodic phrase boundaries, Comma Rules, and Brain Responses: The Closure Positive Shift in ERPs as a Universal Marker for Prosodic Phrasing in Listeners and Readers. *Journal of Psycholinguistics Research*, Vol. 30, No. 3, 2001.

Strangert, Eva (1990). Perceived pauses, silence intervals and syntactic boundaries. *Reports from the Department of Phonetics*. University of Umeå.

Strangert, Eva (1992). Prosodic cues to the perception of syntactic boundaries. ICSLP 92.

Taylor, Paul & Black, Alan (1998): Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language* 12, 99-117.

Wang, Michelle & Hirschberg, Julia (1992). Automatic classification of intonational phrase boundaries. *computer Speech and Language* 6, 175-196.

7. Appendix

A. Mapping table of the linguistic features

1	content/function	content or function word, minor or major delimiter
2	pos-L	part-of-speech tag of the left context word
3	pos	part-of-speech tag of current word
4	pos-R	part-of-speech tag of the right context word
5	pos-L0	part-of-speech bigram of left word and current word
6	pos-0R	part-of-speech bigram of current word and right word
7	pos-L0R	part-of-speech trigram of left, current and right word
8	parse	parse hierarchy
9	phrase_depth	phrase depth of current word
10	phrase_depth-R	phrase depth of the right context word
11	phrase_depth-0R	phrase depth of current word and right word
12	phrase_depth-diff	phrase depth difference between current and right word
13	last_any_boundary	number of words since last boundary (weak or strong)
14	last_strong_boundary	number of words since last strong boundary
15	phrase_firstlast	first and last parse of the parse hierarchy

B. Examples from the statistics

12 Probabilities phrase depth difference (Value-Strong-Weak-No-Sum-Occurrences):

-3	0.0588	1	0.0588	1	0.8824	15	1	17
-2	0.0968	12	0.0565	7	0.8468	105	1	124
-1	0.0932	58	0.0627	39	0.8441	525	1	622
0	0.0610	64	0.0362	38	0.9029	948	1	1050
1	0.0047	2	0.0800	34	0.9153	389	1	425
2	0.0490	5	0.1275	13	0.8235	84	1	102
3	0.0233	1	0.1628	7	0.8140	35	1	43
4	0.0909	2	0.1364	3	0.7727	17	1	22
5	0.0000	0	0.3077	4	0.6923	9	1	13
6	0.2500	1	0.0000	0	0.7500	3	1	4
7	0.0000	0	0.0000	0	1.0000	1	1	1
9	0.0000	0	1.0000	1	0.0000	0	1	1

3 Probabilities pos optimal (Value-Strong-Weak-No-Sum-Occurrences):

	0.0000	0	0.0000	0	1.0000	1	1	1
AF	0.0000	0	0.1333	4	0.8667	26	1	30
AQ	0.0065	1	0.0581	9	0.9355	145	1	155
CC	0.0000	0	0.0088	1	0.9912	113	1	114
CIS	0.0000	0	0.0000	0	1.0000	27	1	27
CS	0.0000	0	0.0233	1	0.9767	42	1	43
DT	0.0000	0	0.0084	1	0.9916	118	1	119
DTR	0.0000	0	0.0000	0	1.0000	1	1	1
FE	0.8199	132	0.1118	18	0.0683	11	1	161
FI	0.0645	2	0.6452	20	0.2903	9	1	31
I	0.0000	0	0.0000	0	1.0000	6	1	6
MC	0.0000	0	0.1000	3	0.9000	27	1	30
MO	0.0000	0	0.0000	0	1.0000	4	1	4
N	0.0000	0	0.0000	0	1.0000	2	1	2
NC	0.0174	9	0.0853	44	0.8973	463	1	516
NP	0.0106	2	0.1217	23	0.8677	164	1	189
PF	0.0000	0	0.0109	1	0.9891	91	1	92
PH	0.0000	0	0.0000	0	1.0000	41	1	41
PS	0.0000	0	0.0000	0	1.0000	9	1	9
Q	0.0000	0	0.1818	4	0.8182	18	1	22
RG	0.0000	0	0.0491	8	0.9509	155	1	163
RH	0.0000	0	0.0000	0	1.0000	13	1	13
SP	0.0000	0	0.0174	5	0.9826	282	1	287
V	0.0000	0	0.0136	5	0.9864	363	1	368

C. Test form

Lyssna på hur orden är grupperade i dessa meningar, d.v.s. hur pauserna är distribuerade. Bedöm detta på en skala 1-5. Förekommer de på rätt ställen? Är det för många eller för få pauser? Underlättar de för förståelsen av yttrandet?

Ringa in en siffra per yttrande (1 = mycket dåligt, 5 = mycket bra).

Det är inte *hur* pausen låter som ska bedömas, utan *var* pausen finns.

Försök att bortse från att syntesen ibland uttalar ord fel, och att satsmelodin kan låta lite konstig ibland. Övergångarna mellan orden kan dessutom låta lite onaturliga.

Övningsfiler:

A01. 1 2 3 4 5

A02. 1 2 3 4 5

A03. 1 2 3 4 5

A04. 1 2 3 4 5

A05. 1 2 3 4 5

B01. 1 2 3 4 5

Kommentar:

B02. 1 2 3 4 5

Kommentar:

...

D. Synthesised sentences

5 sentences for practising:

1. För den icke språkliga människan representerade ett speciellt ljud inte någon särskild tanke.
2. Om hon viftade med armarna var hon en fågel; om gruppen formade en cirkel och flyttade sig med regelbundna steg kunde de vara månen.
3. Nya bevis för att den så kallade medelhavsdieten är nyttig för hjärtat, presenteras på måndagen vid den pågående hjärtkongressen i Wien.
4. Resultatet visar att den som äter frukt, grönsaker, fisk och olivolja, och dessutom dricker måttliga mängder vin löper mindre risk för hjärtinfarkt och kärlkramp.
5. Finansmannen Johan Björkman misstänks ha undanhållit mångmiljonbelopp i förmögenhetsskatt och är delgiven misstanke om grovt skattebrott, skriver tidningen Dagens Industri.

5 longer sentences from Pisoni:

1. Den allra tidigaste muntliga kommunikationen kommer säkert från vanliga aktiviteter: en grymtning eller ett skrik för att varna de andra bärplockarna för att ett farligt djur närmade sig.
2. I Maskinens Myt, diskuterar Bosse Eriksson utvecklingen av en intressant hypotes att formaliseringen var en process som stimulerades av möjligheten till upprepning i rituella ceremonier.
3. Det är tveksamt om det över huvud taget finns en lins med bländaröppning över fyra komma fem som ger lika god upplösning vare sig vid stor eller liten bländaröppning.
4. I själva verket visste han ganska lite om landet eftersom hans enda möte med landet var en kort tågresa från Frankrike 1906 till staden San Sebastian där han besökte en tjurfäktning.
5. Metoden att datera föremål med hjälp av radioaktivitet kan spåras ända till 1905 då en geolog påpekade att allmänna förekomsten av bly i uranhaltiga stenar och observerade att mängden bly och uran var förvånansvärt konstant i stenprover från samma område.

2 shorter sentences from Pisoni:

6. Dom kan göra nästan vad som helst för ett skratt.
7. Han hade inte släppt in ett enda skott.

5 longer news sentences:

8. Frågan som anses principiellt viktig behandlades av regeringen på torsdagen efter att Slotts- och Domkyrkomuseet i Linköping velat låna två föremål med anknytning till trakten i samband med firandet av den heliga Birgitta i år, men fått nej.

9. Polisens hemliga telefonavlyssning av misstänkta ökar kraftigt, och nu är chefs-JO Claes Eklundh kritisk till hur avlyssningsärenden sköts vid Stockholms tingsrätt.

10. Han säger att det är väldigt, väldigt svårt att kontrollera handläggningen hos tingsrätten och att det behövs en betydligt bättre ordning på papperen.

11. Det finns uppgifter om att American Airlines inom de närmaste dagarna kommer att ansöka om konkursskydd, om inte också flygvärdinnorna går med på lönesänkningar.

12. Försöken att isolera smittbärarna, i kombination med resevarningar har haft positiv effekt, bland annat i Guandongprovinsen i Kina, som är ett av de områden där smittan spritts.

2 shorter news sentences:

13. Det virus som orsakar sjukdomen har konstaterats hos sex av de misstänkta fallen.

14. Det är 28 procent färre än hösten 1994.

2 utterances from spontaneous speech:

15. dom blir ju drabbade i vilken form det än beslutas om det nu var så att man hade kommit fram i säkerhetsrådet till att alla andra åtgärder var uttömda jag menar att det finns ju man kan ju använda diplomati man kan använda politik man kan använda humanitära insatser biståndsinsatser många såna saker som inte är uttömt nej

16. utan får jag säga det utan att vi som är kritiska ska bli pådyvlade åsikter som att vi står på samma sida som saddam hussein eller att vi går bin ladin intressen det är ju ett mentalt tillstånd klimat i debatten som gör att man blir mörkrädd tycker jag när man inte ens få ha en öppen diskussion