Chapter 8

Evaluation

8.1 Introduction

Our research objective has been to design, build, and test a linguistically-based SLG system that generates speech with pitch accents and intonational boundaries that are most acceptable to experts. The SLG system and the linguistic theory on which it is based were described in the previous chapters. The present chapter reports on an experimental evaluation in which the accents and boundaries generated by the system were compared with those assigned by a number of human experts, in order to determine to what extent our objective has been realized.

Some parts of the underlying linguistic theory were already evaluated. Phonologi1cal phrasing, an essential part of our approach to accent placement, was investigated in Chapter 3, and the newly proposed φ-algorithm was tested using a corpus of flower descriptions. However, the complete algorithm for pitch accent placement as described in the previous chapter, although supported by exemplary data, has not yet been formally evaluated. The $\mathcal{NEW}$ intonational phrasing algorithm proposed in Chapter 4 was also evaluated with respect to flower descriptions. Although this evaluation showed that it performed significantly better than at least two of the other algorithms, this result may to some extent be ascribed to domain-specificity. Thus, although parts of the underlying linguistic theory were tested, the output of the SLG system itself has not been formally evaluated. In addition, previous tests took place in the domain of flower descriptions, and it remains to be shown that the algorithms perform equally well when applied to other domains.

A point that has often been made in the literature is that SLG has an advantage over TTS, because it can exploit the detailed and reliable linguistic information generated by the NLG component for prosody assignment (Section 5.2.3). In order to get an impression of the performance difference between an SLG system and a general-purpose TTS system, the evaluation also included a comparison of output of our SLG system with that of KUNTTS, a representative TTS system for Dutch.

This chapter is organised as follows. We will first describe the method and material of the evaluation in Section 8.2. Next, we will present the results for pitch accent placement in Section 8.3, and those for intonational phrasing in Section 8.4. These results will then be discussed in Section 8.5, and compared to those reported in other work. We will conclude with a summary in Section 8.6.

8.2 Method and material

The purpose of this evaluation is to establish to what extent the pitch accents and intonational boundaries generated by the SLG system are in accordance with those preferred by human experts. One method is therefore to present the spoken output to a panel of (expert) listeners, and ask them to judge the acceptability of the accents and boundaries. As argued in Section 1.1.4, however, the disadvantage of a perceptual evaluation is that other aspects of the wave form, such as segmental errors, inadequate phonetic realization, or speaker characteristics, may interfere with judgement of the intonation. We therefore decided to rely on a symbolic assessment of the acceptability judgements of experts, asking them to prescribe the preferred accent placement and intonational phrasing for a given text, and comparing these with the predictions of our algorithms.

Below, we will describe the text material that was used (Section 8.2.1), and the assignment of pitch accents and intonational boundaries to these texts by the human experts (Section 8.2.2), our SLG system (Section 8.2.3), and the KUNTTS system (Section 8.2.4).

8.2.1 Text material

Evaluation at the symbolic level was applied earlier to investigate phonological phrasing and accent placement in Chapter 3, and intonational phrasing Chapter 4. We used a text corpus with flower descriptions that was syntactically analysed by hand, including the argument structure and the focus distribution. For this evaluation, we used *another* type of text material, in order to investigate whether the algorithms perform equally well when applied to other domains: travel descriptions and numbers (or expressions containing numbers) generated by our SLG system.[1] In contrast to the former evaluations, the text material did not have to be parsed manually, because all semantic and syntactic information was generated by the SLG system.

The travel descriptions were generated by PTDGen, the SLG system for the public transport domain (PTD). PTDGen can generate three types of descriptions: *concise*, *normal*, and *elaborate* (cf. Section 5.3.4). For each type of description, two texts were generated: a travel description involving only one

[1]For this reason, flower descriptions generated by WFDGen were not included in this evaluation.

transfer, and another involving three transfers. This resulted in six texts ranging from short (2 sentences) to long (14 sentences). In addition, the two output examples in Appendix E were included. Thus, 8 travel descriptions with a total of 56 sentences, and an average of 11.5 words per sentence, were used. Two example texts are given in Appendix C. In the remainder of this chapter, we will refer to this text material as TD (Travel Descriptions).

The second type of text material consisted of numerals, and phrases or sentences containing numerals, which were generated by our SLG system on the basis on manually supplied APS's. These are examples that occurred in Section 7.3.5, or variants with identical syntactic structures but different words. In total, there were 60 instances with an average of 4.6 words per instance. This text material will henceforth be referred to as NR (Numbers); a part of it is given in Appendix D. With the exception of a full stop at the end of sentences, the text material did not contain any punctuation symbols, digits, or abbreviations, as these are not generated by the system.

Some compounds, like the numeral in (1), can receive two or more accents.

(1) $\overset{*}{\text{e}}\text{enentw}\overset{*}{\text{i}}\text{ntig}$
one-and-twenty
'21'

For these phrasal compounds (see the paragraph Compounds in Section 7.3.1.1), the parts (i.e. *eenen* and *twintig*) were separated by a blank, and counted as a separate words. Accordingly, the total number of words in TD is 643, which represents the total number of accent decisions, while the total number of accent decisions in NR is 274.

8.2.2 Intonation by the human experts

The text material was presented to four human experts. All of them are native speakers of Dutch and have been involved in research on intonation; none of them participated in the previous experiment with the corpus of flower descriptions. After being informed of the purpose of the evaluation, they were asked to indicate the locations of the pitch accents and intonational boundaries to be used in an ideal reading of the text. To accommodate the possibility that alternative accentuations are equally acceptable, a distinction was made between obligatory accents (indicated by '*') and optional accents (indicated by '+'). Analogously, they could assign obligatory ('#') and optional boundaries ('|'). The result will be referred to as the *reference* or *prescription* (see Appendix C and Appendix D). In accordance with our research objective, tune was not taken into consideration.

The initial prescriptions were collected, formatted in a convenient way, and then presented again to the experts. They were asked to compare their own prescription with those of the three others, and to reconsider their choices in case of differences. Not only did this allow them to correct inconsistencies in their own judgements, it also served to increase the consensus. We postpone a presentation of the number of accents and boundaries assigned by the experts, and the agreement among them, to Section 8.3.1.

8.2.3 Intonation by the SLG system

The texts were generated by our SLG system, and accents and intonational boundaries were assigned by the prosodic realization module ProsR. Pitch accent placement was carried out by the rule-based algorithm described in Chapter 7 and summarised in Table 7.1. Intonational phrasing was done by the $\mathcal{NEW}$ algorithm presented in Section 4.8.1 and summarised in Table 4.14. In contrast to the human experts, ProsR did not assign optional accents or boundaries.

8.2.4 Intonation by the TTS system

As already mentioned, an additional point of interest in the evaluation was a comparison between SLG and TTS. Therefore, the text material was also processed by KUNTTS (Section 5.6), which is representative for the type of prosody assignment achieved by current TTS systems for Dutch (Kerkhoff and Rietveld 1994; Kerkhoff and Rietveld 1995; Kerkhoff, Rietveld, and van Bergem 1998). The accents and boundaries assigned by KUNTTS were extracted from the trace files, allowing us to compare TTS to SLG. KUNTTS did not assign optional accents or boundaries.

The *Syntax-To-Prosody module* (STP) of KUNTTS was developed as part of the multilingual TTS system developed in the ESPRIT-project POLYGLOT (Dirksen 1992; Dirksen and Quené 1993), and is based on metrical theory and the principles of focus-accent theory (Baart 1987). Its input consists of a syntactic parse by a robust wild card parser. The syntactic tree is translated to a binary-branching metrical tree, the nodes of which are labelled weak or strong according the syntactic relations between nodes. For example, the phrase structure rule in (2) produces a left-branching metrical tree in which the NP is strong (\) with respect to the V, and the DET is weak (/) with respect to the N. It is assumed that by default every major phrase receives a pitch accent, where location of the accent is determined by following the path of strong nodes until a terminal node is reached. The rules in (2) represent the 'normal' situation in which an accent within the NP assigns focus to the entire VP, as shown in in (3).

(2) $VP \longrightarrow NP \setminus V$
$NP \longrightarrow DET \ / \ N$

(3) hij heeft ($_{VP}$ ($_{NP}$ een bo*ek) ($_V$ gelezen))
he has a book read
'he read a book'

The lexicon specifies which words are typically *not* accented; by and large these are function words. If such a word appears in a position that is strong according to the phrase structure rules, a deaccenting rule is invoked which reverses the weak-strong labelling. As a result, an accent will shift to a sister node, as shown in (4).

(4) hij heeft ($_{VP}$ ($_{NP}$ het) ($_{V}$ gele*zen))
he has it read
'he read it'

In addition, there are rules that reverse the weak-strong labelling in stress shift environments, akin to our RR.

As far as prosodic phrasing in concerned, STP places a major boundary between NP and VP, and a minor boundary between NP and PP, provided that the created prosodic phrase contains at least one accent. Some additional post-processing takes place to avoid phrases which are too short.

8.3 Results for pitch accent placement

We will first present the quantitative agreement with respect to pitch accent placement for the humans, ProsR and KUNTTS. Next, we will analyse differences between the experts and ProsR to see if these can be traced back to particular deaccentuation rules.

8.3.1 Agreement in accentuation

Table 8.1 shows the number of accents in TD and NR assigned by the four human experts and the two algorithms. As already mentioned, neither KUNTTS nor ProsR assigned optional accents. It can be seen that the experts assigned about the same number of obligatory and optional accents. The average number of obligatory pitch accents in NR was substantially higher than in TD, because most of the text material consisted of short expressions (numbers), containing few function words and lacking any unfocused material.

In order to estimate the agreement among the four human experts, Fleiss kappas were calculated using three categories (obligatory, optional, and no accent). For TD, kappa was 0.79 (643 objects, $p < 0.000$); for NR, kappa was 0.89 (274 objects, $p < 0.000$). Kappa values of 0.8 and higher are generally considered to indicate a high level of agreement (Rietveld and van Hout 1993).

The agreement between two accent assignments can also be expressed in the form of a 3×3 confusion matrix. For instance, Table 8.2 expresses the agreement in accent placement in TD between human expert CO and the ProsR algorithm. From this, we can calculate an agreement index (between zero and one) by taking the sum of the numbers on the diagonal, and dividing this by the total number of accent decisions:

$$agreement = \frac{233 + 0 + 322}{643} = 0.86$$

Table 8.1: Pitch accents in TD and NR assigned by the human experts and the algorithms, in both absolute numbers, and as a percentage of the total number of accent decisions in TD (n=643) and NR (n=274).

	TD:				NR:			
	Obligatory:		Optional:		Obligatory:		Optional:	
Humans:	N:	%:	N :	%:	N:	%:	N:	%:
CO:	260	40.44	37	5.75	172	62.77	3	1.09
JC:	284	44.17	40	6.22	169	61.68	7	2.55
JH:	246	38.26	45	7.00	164	59.85	10	3.65
MS:	250	38.88	55	8.55	167	60.95	7	2.55
$\overline{x}$:	260	40.44	44.25	6.88	168	61.31	6.75	2.46
sd:	17.05	2.65	7.89	1.23	3.37	1.23	2.87	1.05

	TD:				NR:			
	Obligatory:		Optional:		Obligatory:		Optional:	
Algorithms:	N:	%:	N:	%:	N:	%:	N:	%:
KUNTTS:	249	38.72	0	0	148	54.01	0	0
ProsR:	266	41.37	0	0	163	59.49	0	0

However, this method creates an unreasonable disadvantage for KUNTTS and ProsR, because neither of these assigned optional accents, which implies that all optional accents are interpreted as instances of disagreement. In fact, *optional* means that the accent is allowed, but may equally be omitted. Thus, whether the 37 optional accents assigned by CO are obligatory (9 cases) or none (28 cases) according to ProsR is actually not important, and we therefore chose to ignore optional accents (and optional boundaries for that matter). The same approach was taken in Chapter 3 and Chapter 4, where only clear cases (words that must be accented or cannot be accented) were taken into consideration, and variable cases (words that may be accented) were ignored. As a result, the agreement index for Table 8.2 was calculated in the following way:

$$agreement = \frac{233 + 322}{233 + 24 + 27 + 322} = 0.91$$

Table 8.3 shows the agreement indices for pitch accent assignment. It can be observed that the average agreement of ProsR is comparable to that of the human experts, whereas that of KUNTTS is substantially lower. This is clearly visible in the graphical representation of the same data in Figure 8.1, which

Table 8.2: Confusion matrix expressing the agreement in accent placement in TD between ProsR and human expert CO.

		CO: Obligatory:	Optional:	None:	Total:
ProsR:	Obligatory:	233	9	24	266
	Optional:	0	0	0	0
	None:	27	28	322	377
	Total:	260	37	346	643

suggests that the accentuation of ProsR was close to that of the human experts, but there is still room for improvement.

Figure 8.2 shows the dendrograms resulting from an hierarchical cluster analysis (using complete linkage) on the agreement indices for pitch accent placement. As expected, the human experts were clustered first, closely followed by ProsR, whereas KUNTTS was at a large distance.

8.3.2 Disagreement in accentuation

Unlike the approach used in the evaluations reported in Chapter 3 and Chapter 4, the human experts were not asked to form a consensus accentuation, which may serve as a reference. Instead, there were four individual prescriptions, which showed high agreement. In order to facilitate a qualitative comparison between ProsR and the human experts, we made reference to a majority criterion. It was assumed that ProsR generated *too few accents* in either of the following situations:

1. all four experts assigned an obligatory accent, but ProsR failed to assign one;

2. three experts assigned an obligatory accent and one assigned an optional boundary, but ProsR failed to assign one.

Similarly, we assumed that ProsR generated *too many accents* if either:

1. none of the experts assigned an accent, but ProsR assigned one;

2. three experts assigned no accent and one assigned an optional accent, but ProsR assigned one.

Below, we will analyse both types of disagreement.

Table 8.3: Agreement indices in TD and NR among the human experts and the algorithms with respect to pitch accent placement.

TD:						
	CO:	JC:	JH:	MS:	KUNTTS:	ProsR:
CO:		0.97	0.96	0.97	0.76	0.92
JC:	0.97		0.98	0.98	0.75	0.92
JH:	0.96	0.98		0.96	0.76	0.95
MS:	0.97	0.98	0.96		0.76	0.91
$\overline{x}$:	0.97	0.98	0.97	0.97	0.76	0.93
sd:	0.01	0.01	0.01	0.01	0.00	0.02

NR:						
	CO:	JC:	JH:	MS:	KUNTTS:	ProsR:
CO:		0.97	0.96	0.96	0.76	0.92
JC:	0.97		1.00	0.99	0.79	0.95
JH:	0.96	1.00		0.99	0.79	0.96
MS:	0.96	0.99	0.99		0.79	0.96
$\overline{x}$:	0.96	0.99	0.98	0.98	0.78	0.95
sd:	0.01	0.02	0.02	0.02	0.02	0.02

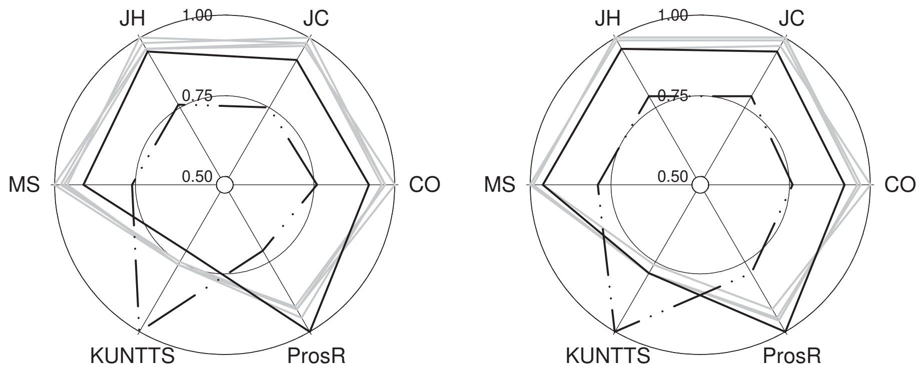

Figure 8.1: *A graphical representation of the data in Table 8.3, showing the agreement indices for TD (left) and NR (right) with respect to pitch accent placement. The centre of the circle represents an agreement of 0.5 and its outside represents an agreement of 1. ProsR is represented by the black line, KUNTTS by the dashed black line, and the four human experts by grey lines.*

```
                  Rescaled Distance Cluster Combine

                 0         5        10        15        20        25
                 +---------+---------+---------+---------+---------+

        JC       --+
                   |
        MS       --+-+
                   | |
        CO       --+ +---------+
                     |         |
        JH       ----+         +-----------------------------------+
                               |                                   |
        ProsR    --------------+                                   |-
                                                                   |
        KUNTTS   --------------------------------------------------+
```

```
                  Rescaled Distance Cluster Combine

                 0         5        10        15        20        25
                 +---------+---------+---------+---------+---------+

        JC       --+
                   +-+
        JH       --+ |
                     +-----+
                     |     |
        MS       ----+     +-------+
                           |       |
        CO       ----------+       +-------------------------------+
                                   |                               |
        ProsR    ------------------+                               |-
                                                                   |
        KUNTTS   --------------------------------------------------+
```

Figure 8.2: Dendrograms resulting from an hierarchical cluster analysis on the agreement indices with respect to pitch accents (cf. Table 8.3) for TD (top) and NR (bottom).

Table 8.4: Differences involving ProsR assigning too few accents.

Rule:	TD:	NR:	Total:
Focus	15	0	15
URR	0	8	8
Total:	15	8	23

8.3.2.1 Too few accents

There were 23 cases in which ProsR assigned too few accents (see Table 8.4). These were caused either by lack of focus or overapplication of the Unit Rhythm Rule (Section 7.3.5); overapplication of SAAR, RR or any of the other deaccentuation rules did not occur.

As far as TD is concerned, 13 out of 15 differences involved focusing of repeated placenames. PTDGen backgrounded all placenames that were mentioned before in the current paragraph (i.e. the description of the current section of the journey). Apparently, this focus strategy was too restrictive, resulting in too few focused placenames and consequently too few accents. In contrast, the human experts appeared to accent every placename, unless it was mentioned in the directly preceding sentence. We will return to this issue in Section 8.5.1.

With respect to NR, there were four cases where an accent on *honderd* 'hundred' was incorrectly deleted by URR. Two of these are given in (5).

(5) a. ?h$\overset{ø}{o}$nderd en z$\overset{*}{e}$ven
hundred and seven
107
b. ?h$\overset{ø}{o}$nderd en tw$\overset{*}{e}$e
hundred and two
102

In contrast, there were many similar examples where *honderd* was correctly deaccented by URR, two of which occur in (6).

(6) a. h$\overset{ø}{o}$nderd v$\overset{*}{i}$eren d$\overset{*}{e}$rtig
hundred four-and thirty
134
b. h$\overset{ø}{o}$nderd t$\overset{*}{i}$en h$\overset{*}{o}$ofdstukken
hundred ten chapters

This might be related to the fact that initially there are three accents in (6), but only two in (5), and that human speakers try to avoid a series of three or more accents.

In the two other cases (7), *honderd* was erroneously assumed to be a unit, due to shortcomings of the current implementation, causing it to be subject to URR deaccentuation.

(7) a. ?h$\overset{\emptyset}{\text{o}}$nderd d$\overset{*}{\text{u}}$izend
hundred thousand

b. ?h$\overset{\emptyset}{\text{o}}$nderd k$\overset{*}{\text{a}}$sten
hundred cupboards

There were three cases in which another unit, *duizend* 'thousand', was deaccented. All three are structurally identical to the one in (8). As explained at the end of Section 7.3.5, there appear to be two ways to accent these long numbers. The first (8-a) is without an accent on *duizend* and with no internal AD-boundary, which was the option taken by ProsR, whereas the second (8-b) is with *duizend* accented and followed by an AD-boundary, which was chosen by the experts. This analysis is supported by the fact that all experts (with the exception of MS, who assigned an optional boundary) preferred an obligatory boundary after accented *duizend*. If ProsR would have inserted this AD-boundary, this would have blocked deaccentuation of *duizend* by URR. This disagreement can thus be attributed to a phrasing decision, which we will further discuss in Section 8.4.2.

(8) a. [$_{\text{AD}}$ h$\overset{\emptyset}{\text{o}}$nderd v$\overset{*}{\text{i}}$eren v$\overset{\emptyset}{\text{i}}$jftig d$\overset{\emptyset}{\text{u}}$izend z$\overset{*}{\text{e}}$ven h$\overset{\emptyset}{\text{o}}$nderd twe$\overset{*}{\text{e}}$en
hundred four-and fifty thousand seven hundred two-and
d$\overset{\emptyset}{\text{e}}$rtig k$\overset{*}{\text{a}}$sten]
thirty cupboards
'154,732 cupboards'

b. [$_{\text{AD}}$ h$\overset{\emptyset}{\text{o}}$nderd v$\overset{*}{\text{i}}$eren v$\overset{\emptyset}{\text{i}}$jftig d$\overset{*}{\text{u}}$izend] [$_{\text{AD}}$ z$\overset{*}{\text{e}}$ven h$\overset{\emptyset}{\text{o}}$nderd twe$\overset{*}{\text{e}}$en
d$\overset{\emptyset}{\text{e}}$rtig k$\overset{*}{\text{a}}$sten]

8.3.2.2 Too many accents

The majority of the ten differences involving too many accents occurred in TD (see Table 8.5). Eight of these involved a predicate that should not have been accented, suggesting SAAR did not work satisfactorily. However, as many as five of these were due to the fact that an adjacent argument was a placename which was erroneously backgrounded, a type of error that was discussed in the previous section. In (9), for example, the argument *Amsterdam* is backgrounded, which prevents SAAR from deaccenting (the verbal particle of) the predicate *aankomen*. This is thus an error in focus distribution, rather than an error of SAAR.

(9) ?Dan komt u in Amsterdam $\overset{*}{\text{a}}$an om ...
then arrive you in Amsterdam `prt` at ...

Table 8.5: Differences involving ProsR assigning too many accents.

Context:	TD:	NR:	Total:
predicate	7	1	8
unit	1	1	2
Total:	8	2	10

One of the remaining three errors in relation to SAAR is given in (10). Since the time of departure is definitely not an argument, SAAR cannot apply here. A possible explanation is that *gaan*, which also serves as an auxiliary of future tense in Dutch, is a semantically light verb, as most of meaning of 'going to' is expressed by the preposition of its argument (*naar* 'to') rather than by the verb itself (see also Section 7.3.6.2). Alternatively, it may be related to the observation by Baart (1987) that sometimes initial verbs do not appear to behave according to SAAR (Gussenhoven 1992b, Appendix).

(10) ?Gå* daarna om å*cht uur ze*ven twi*ntig naar Bre*da
go next at eight hour seven twenty to Breda
'Next, go to Breda at 8.27'

Finally, (11-a) is one of the two errors involving a unit term (cf. Section 7.3.5. Application of URR to *minuten* 'minutes' was blocked because of the following boundary inserted by ProsR. However, none of the experts assigned this boundary. Given the lack of a boundary, ProsR would not have accented *minuten* either. In other words, this apparent accentuation error can again be traced back to a disagreement with respect to phrasing. The second error of this type, occurring on *uur* in (11-b), cannot be explained in this way.

(11) a. ?u hebt twe*e minu*ten # om o*ver te stappen op ...
you have two minutes to `prt` to transfer onto ...
'you have two minutes to transfer to ...'

b. ?die tre*in gaat om twe*e u*ur # via ...
that train goes at two hour via

To sum up, our survey of differences in pitch accent placement between ProsR and the human experts employing a majority criterion showed that the number of cases involving too few accents (23) was more than twice of that involving too many accents (10), indicating that ProsR tended to assign too few rather than too many accents. About two-thirds of the missing accents involved placenames, and resulted from the focus strategy. The remainder was due to overapplication of URR, part of which could be explained by intonational phrasing differences between ProsR and the experts. The majority (8) of the superfluous accents could be attributed to overapplication of SAAR. However, five of these were actually caused by unaccented placenames, indicating interaction between accentuation errors.

8.4 Results for intonational phrasing

We will first present the quantitative agreement with respect to intonational phrasing for the human experts, ProsR and KUNTTS. Next, we will analyse differences between the experts and ProsR to see if these can be traced back to particular phrasing rules.

8.4.1 Agreement in phrasing

Table 8.6 shows the number of intonational boundaries in TD and NR assigned by the human experts and the algorithms. Assuming that intonational boundaries could only occur between accentable words (cf. Section 8.2.1), the total number of boundary decisions was 587 (= 643−56) and 214 (= 274−60), respectively. While the previous section showed that the experts assigned about the same number of accents, the difference with respect to boundaries turned out to be larger. One of the experts (MS) assigned relatively few obligatory boundaries in both domains, and another (OC) relatively few optional boundaries in TD. Both algorithms, and KUNTTS in particular, predicted substantially fewer boundaries than the experts (with the exception of MS).[2]

Fleiss kappas were again calculated in order to estimate the agreement among the four human experts, using three categories (obligatory, optional, and no boundary). For TD, kappa was 0.58 (587 objects, $p < 0.000$); for NR, kappa was 0.65 (214 objects, $p < 0.000$). Kappas between 0.6 and 0.8 are generally considered to indicate an acceptable level of agreement (Rietveld and van Hout 1993).

Table 8.7 presents the agreement with respect to intonational phrasing, where the agreement indices were calculated in the same way as those for accentuation, that is, without taking optional boundaries into account. A graphical representation of the same data is shown in Figure 8.1. It can be observed that the agreement between the algorithms and MS was relatively high. Part of the explanation is that both assigned relatively few boundaries. Furthermore, the differences between the human experts and the algorithms were smaller than for accentuation, in particular for KUNTTS. As will be discussed in Section 8.5.3, this may be be attributed to the fact that boundaries occurred less frequently. Still, the agreement among the human experts was higher. With respect to TD, ProsR appeared to be slightly better than KUNTTS.

The result of a hierarchical cluster analysis (using complete linkage) on the agreement indices is shown in Figure 8.4. As expected, the results were less

[2] This is in accordance with the strategy employed by KUNTTS that a wrong boundary is worse than a missing boundary.

Table 8.6: Intonational boundaries in TD and NR assigned by the human experts and the algorithms, in both absolute numbers, and as a percentage of the total number of boundary decisions in TD (n=587) and NR (n=214).

	TD:				NR:			
	Obligatory:		Optional:		Obligatory:		Optional:	
Humans:	N:	%:	N:	%:	N:	%:	N:	%:
CO:	55	9.37	14	2.39	13	6.07	20	9.35
JC:	48	8.18	57	9.71	9	4.21	25	11.68
JH:	45	7.67	43	7.33	15	7.01	20	9.35
MS:	2	0.34	46	7.84	2	0.93	26	12.15
$\overline{x}$:	37.5	6.39	40	6.81	9.75	4.56	22.75	10.63
sd:	24.03	4.09	18.35	3.13	5.74	2.68	3.2	1.5

	TD:				NR:			
	Obligatory:		Optional:		Obligatory:		Optional:	
Algorithms:	N:	%:	N:	%:	N:	%:	N:	%:
KUNTTS:	6	1.02	0	0	2	0.93	0	0
ProsR:	29	4.94	0	0	4	1.87	0	0

clear cut than for the accentuation data. As far as TD was concerned, three experts (JC, JH, and CO) were clustered, whereas MS was different and was clustered with KUNTTS. ProsR appeared to be in between the three human experts on the one hand and MS and KUNTTS on the other. In the NR data, the same three experts were clustered again, whereas MS, ProsR and KUNTTS were clustered in another group. This clustering may be explained by the fact that JC, JH and CO assigned substantially more boundaries than MS, ProsR, and KUNTTS (cf. Table 8.6).

Table 8.7: Agreement in TD and NR with respect to intonational boundaries.

TD:						
	CO:	JC:	JH:	MS:	KUNTTS:	ProsR:
CO:		1.00	0.99	0.97	0.90	0.94
JC:	1.00		1.00	0.98	0.91	0.95
JH:	0.99	1.00		0.98	0.91	0.96
MS:	0.97	0.98	0.98		0.99	0.98
$\overline{x}$:	0.99	0.99	0.99	0.98	0.93	0.96
sd:	0.02	0.01	0.01	0.01	0.04	0.02

NR:						
	CO:	JC:	JH:	MS:	KUNTTS:	ProsR:
CO:		1.00	0.99	0.99	0.94	0.95
JC:	1.00		1.00	0.99	0.96	0.97
JH:	0.99	1.00		0.99	0.93	0.94
MS:	0.99	0.99	0.99		0.99	0.99
$\overline{x}$:	0.99	1.00	0.99	0.99	0.96	0.96
sd:	0.01	0.01	0.01	0.00	0.03	0.02

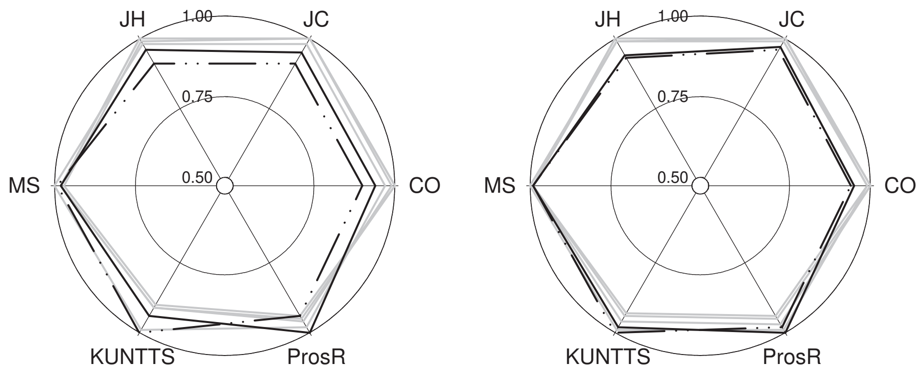

Figure 8.3: *A graphical representation of the data in Table 8.7, showing the agreement indices for TD (left) and NR (right) with respect to intonational phrasing. The centre of the circle represents an agreement of 0.5 and its outside represents an agreement of 1. ProsR is represented by the black line, KUNTTS by the dashed black line, and the four human experts by grey lines.*

```
                   Rescaled Distance Cluster Combine

         0         5        10        15        20        25
         +---------+---------+---------+---------+---------+

JC       --+
           +---+
JH       --+   |
               +-----------------------+
               |                       |
CO       ------+                       +-------------------+
                                       |                   |
ProsR    ------------------------------+                   |
                                                           |-
                                                           |
KUNTTS   ------+                                           |
               +-------------------------------------------+
MS       ------+
```

```
                   Rescaled Distance Cluster Combine

         0         5        10        15        20        25
         +---------+---------+---------+---------+---------+

JC       --+
           +-----+
JH       --+     |
                 +-----------------------------------------+
                 |                                         |
CO       --------+                                         |
                                                           |
MS       --------+                                         |-
                 +-------+                                 |
ProsR    --------+       +                                 |
                         +---------------------------------+
                         +
KUNTTS   ----------------+
```

Figure 8.4: Dendrograms resulting from an hierarchical cluster analysis on the agreement indices with respect to intonational boundaries in TD (top) and NR (bottom).

Table 8.8: Differences involving ProsR assigning too few boundaries.

Context:	TD:	NR:	Total:
in a coordination	0	5	5
before an extraposed PP	12	0	12
in a VP	2	0	2
after a topicalised AdvP	1	0	1
Total:	15	5	20

8.4.2 Disagreement in phrasing

For a qualitative comparison between ProsR and the human experts, we used the same type of majority criterion as we did for accentuation. That is, we assumed ProsR generated *too few boundaries* if either:

1. all four experts assigned an obligatory boundary, but ProsR failed to assign one;

2. three experts assigned an obligatory boundary and one assigned an optional boundary, but ProsR failed to assign one.

Similarly, ProsR generated *too many boundaries* if either:

1. none of the experts assigned a boundary, but ProsR assigned one;

2. three experts assigned no boundary and one assigned an optional boundary, but ProsR assigned one.

8.4.2.1 Too few boundaries

The differences involving too few boundaries are summarised in Table 8.8. Starting with TD, there were 12 differences due to the absence of a boundary before an extraposed PP. In these cases, the rule AD-in-S-2 (Section 4.8.1.2), which is responsible for inserting a boundary before an extraposed constituent, failed to apply. In (12), for example, there is only one accent in the first part of the sentence, which is not sufficient to allow AD-in-S-2 to insert the preferred boundary before the extraposed PP *om negen uur zesen veertig*.[3]

[3] In at least nine cases, this may seem to be caused by the non-accentuation of a placename, which is in turn due to focus error. For example, the placename *Amersfoort* was backgrounded and thus not accented, leaving only one accent in the first part of the sentence. However, if *Amersfoort* had been accented, then the verbal particle *aan* would have been deaccented by SAAR, and there would still be only one accent left.

(12) ?Die komt in Amersfoort $\overset{*}{\text{a}}$an om n$\overset{*}{\text{e}}$gen uur z$\overset{*}{\text{e}}$sen v$\overset{*}{\text{e}}$ertig
which arrives in Amersfoort `prt` at nine hour six-and forty
'which arrives in Amersfoort at 9:46'

Two more differences were found within VP's. In (13-a), the experts preferred a boundary after *naar Delft.* However, AD-in-VP (Section 4.8.1.3) could not account for this boundary, because this VP daughter had only one accent. For the same reason, the rule could not provide a boundary after *Sneek* either in (13-b).

(13) a. ?u kunt van $\overset{*}{\text{A}}$peldoorn naar D$\overset{*}{\text{e}}$lft met $\overset{*}{\text{e}}$en $\overset{*}{\text{o}}$verstap
you can from Apeldoorn to Delft with one transfer

b. ?u kunt van L$\overset{*}{\text{e}}$lystad C$\overset{*}{\text{e}}$ntrum # naar Sn$\overset{*}{\text{e}}$ek met dri$\overset{*}{\text{e}}$
you can from Lelystad Centre to Sneek with three
$\overset{*}{\text{o}}$verstappen
transfers

The single case of a missing boundary after a topicalised constituent is given in (14). The topicalised AdvP *acht minuten later* received only two accents, which was not sufficient to trigger a following boundary by AD-in-S-1 (Section 4.8.1.2).

(14) ?$\overset{*}{\text{a}}$cht minuten l$\overset{*}{\text{a}}$ter gaat u met de interc$\overset{*}{\text{i}}$ty van Amsterdam
eight minutes later go you with the intercity from Amsterdam
naar $\overset{*}{\text{A}}$mersfoort
to Amersfoort

Turning to NR, missing boundaries occurred in the context of coordinated structures. For instance, the syntactic structure of the number 154,732 in (15) is a coordinated structure with *honderd vieren vijftig duizend* as the first conjunct and *zeven honderd tweeen dertig* as the second.

(15) ?h$\overset{\o}{\text{o}}$nderd v$\overset{*}{\text{i}}$eren v$\overset{\o}{\text{i}}$jftig d$\overset{\o}{\text{u}}$izend z$\overset{*}{\text{e}}$ven h$\overset{\o}{\text{o}}$nderd tw$\overset{*}{\text{e}}$een d$\overset{\o}{\text{e}}$rtig
hundred four-and fifty thousand seven hundred two-and thirty
k$\overset{*}{\text{a}}$sten
cupboards
'154,732 cupboards'

Since all but one of the initial accents in the first conjunct were deleted, the phrasing rule AD-in-coordination (Section 4.8.1.5) could not insert a boundary between the conjuncts. However, all four experts assigned an obligatory boundary in this position. We may explain this in terms of the complexity of the first conjunct: the first conjunct lacks the required number of accents, but it is sufficiently complex to trigger a boundary. There were two more cases that were very similar to (15). Notice also that this boundary would have been correctly inserted if the accent on *duizend* had not been deleted by URR, and vice versa, that deletion of the accent on *duizend* would have been blocked if this boundary had been present. This type of interaction between accentuation and phrasing was discussed in Section 7.3.6.4.

8.4.2.2 Too many boundaries

There was just one case with a superfluous boundary (16). It was due to AD-in-S-2, which inserted a boundary before extraposed S, as both AD's contained at least 2 accents.

			*	*			*					*
(16)	?u	hebt	twee	minuten	#	om	over	te	stappen	op	de	intercity
	you	have	two	minutes		for	`prt`	to	transfer	to	the	intercity
		*										
	naar	Leeuwarden										
	to	Leeuwarden										

To sum up, relative to the preferences of the human experts, ProsR tended to generate too few boundaries, but inserted almost no superfluous boundaries. For NR, an undershoot of AD-in-Coordination was observed, which could partly be explained by interaction with accentuation differences. For TD, the majority of the errors was due to undershoot of AD-in-S-2, which failed to insert a boundary before an extraposed PP.

8.5 Discussion

We will start with some general remarks regarding the evaluation. Next, we will compare our results for pitch accent placement with those of other approaches (Section 8.5.2), and similarly, compare our results for intonational phrasing with those of other approaches (Section 8.5.3).

8.5.1 General remarks

Comparison with earlier results In short, we have found that the accentuation by ProsR agreed with that by the human experts in 93% of the cases in TD and 95% in NR, and that its intonational phrasing agreed with that of the experts in 96% of the cases in both TD and NR (ignoring optional accents and boundaries, and taking the average agreement over all experts). In Chapter 3, we found that the simple version of the pitch accent placement algorithm (focus distribution, SAAR, and RR) resulted in an agreement of 94% with the consensus reference for the corpus of wild flower descriptions. This is comparable to the figures found here with the complete algorithm for pitch accent placement. It should be noted, however, that backgrounding was manually indicated in WFD, which prevented many accentuation errors related to the focus distribution. In addition, complex numbers did not occur in WFD either.

In Chapter 4, it was found that the $\mathcal{NEW}$ intonational phrasing agreed with the reference in 97.66% of the boundary decisions in the flower descriptions. This is just slightly higher than the 96% percent found here, suggesting that the possible effect of tuning on the corpus was relatively small.

Ideally, similar evaluations should be carried out in many different domains, or against a very large corpus that is representative of written text in general. Given the fact that large-scale prosodically-annotated corpora are not yet available for Dutch, such extensive evaluations are beyond the scope of the current work. Nevertheless, since the proposed algorithms for pitch accent placement and intonational phrasing show a similar performance when applied to three different text domains (flower descriptions, travel descriptions, and numbers), they appear to be fairly domain-independent.

Errors in focus distribution A number of accentuation differences (13 out of 23 involving too few accents) were caused by focus errors. PTDGen backgrounded, and therefore did not accent, placenames that were mentioned in the current paragraph (i.e. the description of the current section of the journey), whereas the experts preferred to accent every placename, unless it was mentioned in the directly preceding sentence. However, our objective has not been to predict focus distributions, but to account for their intonational implications (Section 2.5.1), which makes this type of error less relevant for our purposes.

Interaction For several differences, it was found that they were dependent on other differences. For instance, deaccenting of placenames caused failure of SAAR. Similar interactions were observed between URR and AD-in-Coordination. This suggests that decreasing one type of difference could simultaneously reduce other types of differences.

SLG versus TTS Comparing ProsR with KUNTTS, we found that the accentuation algorithm of KUNTTS performed considerably worse. This is to be expected given the advantages of SLG over TTS. We have not looked in detail at these errors, as evaluating KUNTTS was not the goal of this evaluation, but the majority of these can be attributed to the lack of discourse information (focus) in the input, and errors in word class assignment (POS tagging) and syntactic parsing. In addition, the TTS system is not tuned to accent long numerals in the form of words rather than digits.

As far as intonational phrasing is concerned, however, the differences turned out to be far less pronounced. A possible explanation is that the agreement index tends to favour an economical boundary insertion, as will be explained in Section 8.5.3, and virtually all the phrasing errors of KUNTTS involved too few boundaries. An additional disadvantage for the TTS system was that its intonational phrasing normally relies on punctuation, commas in particular, which was not present in TD.

Symbolic versus perceptual evaluation We chose to use a symbolic assessment of the acceptability judgements of our subjects because of the methodological problems associated with a perceptual evaluation (Section 1.1.4). This type of evaluation is similar to a production experiment: we compared the accents and boundaries generated by ProsR with those produced by the human experts. We assumed that there is a preferred intonational form, but allowed optional accents and boundaries in order to describe variants that are equally good. We also cited van Bezooijen (1989), who found a high correlation between a symbolic evaluation of the type conducted here and an additional auditory evaluation, which suggests a perception experiment would yield similar results.

Yet, from a more applied point of view, an additional perception experiment, in which the adequacy of the produced intonation in real speech is judged by listeners, would be desirable. After all, it may be that the lack (or presence) of an accent (or boundary) in one location is far more disturbing than in another. Consider the following pair repeated from Section 8.3.2.1: while the missing accent on *honderd* appears to be an acceptable variant in (17-a), it is a real error in (17-b), where it gives rise to a contrastive interpretation.

(17) a. ?h$\overset{ø}{o}$nderd en z$\overset{*}{e}$ven
hundred and seven
107
b. ?h$\overset{ø}{o}$nderd k$\overset{*}{a}$sten
hundred cupboards

Perhaps this difference is related to the reason for omitting the accent. In (17-a), *honderd* is deaccented for rhythmic reasons, whereas in (17-b), a deaccentuation can only be interpreted as lack of focus. A similar suggestion was made by Kruyt (1985, p. 81), where 'the effect of syntactic grouping' is roughly equivalent to the effect of RR:

> "These two kinds of leaving accentable words unaccented seem to be essentially different. The effect of syntactic grouping probably depends on such factors as speech rate and style of speech. In this case, leaving a word unaccented means the omission of an accent that could have been realized, and it does not necessarily make much difference to the listener whether the word is or is not accented. In contrast, the second kind of leaving accentable words unaccented, signalling 'givenness', comes closer to a conscious act of the speaker who has interpreted the text and is aware of the relations between the sentences. Probably, this kind of leaving words unaccented has greater relevance for communication."

In this connection, we make a final comment on the observation that ProsR produced too few accents, and too few boundaries in particular, rather than too many. As a trade-off exists between the two types of errors, we may choose to adapt the rules in such a way that more accents and boundaries are predicted, decreasing undershoot at the expense of overgeneration, but keeping the same

Table 8.9: Confusion matrices comparing the PROS accentuation with the accentuations by the subjects and the newsreader (Bezooijen 1989).

		Subjects:	
		+Accent:	–Accent:
PROS:	+Accent:	187	87
	–Accent:	11	303

		Newsreader:	
		+Accent:	–Accent:
PROS:	+Accent:	233	41
	–Accent:	28	286

overall agreement with the human experts. From a purely linguistic point of view, it is hard to decide which is better, but this may again be different from an applied point of view. For instance, there is experimental evidence that less accents are produced in fast speech (Caspers 1994, ch. 2). Therefore, the current model might fit a speech synthesiser that speaks fast, whereas the more generous model might be better suited to a synthesiser with a low speech rate.

8.5.2 Comparison with other approaches to accentuation

Van Bezooijen (1989) reports on a extensive evaluation of PROS (Kager and Quené 1987; Kager and Quené 1989; Quené and Kager 1989), a pitch accent placement algorithm for a Dutch TTS system that relies on heuristic rules. The test set consisted of 8 newspaper text (24 sentences) and 8 single sentences, with a total of 588 words. This material was presented in written form to 12 phonetically experienced subjects, who were asked to assign pitch accents according to their personal preferences. A reference accentuation was constructed by including accents assigned by at least 7 subjects. In addition, the material was read aloud by a professional newsreader, whose accents were transcribed by three intonational experts. A comparison of these two reference accentuations with the accentuation by PROS is given in Table 8.9. The agreement index between PROS and the subjects is 83.3%, whereas that between PROS and the newsreader is 88.3%. PROS assigned too many rather than too few accents, in particular in comparison to the subjects' accentuation. This behaviour is contrary to that of ProsR.

In addition, van Bezooijen conducted an auditory evaluation. For each sentence, three versions were synthesised according to the accentuations of PROS, the subjects and the newsreader, respectively. The same 12 subjects were asked to rate the adequacy of each utterance on a 10-point scale. The PROS accentuation received an average score of 6.0, whereas those of the subjects and the newsreader received scores of 7.7 and 7.4 respectively.[4] The difference between PROS and the two others turned out to be statistically significant at the 1% level. Interestingly, this suggests that even if an algorithm agrees with a human accentuation for 83-88% of the words, it can still be perceived as less adequate.

Nachtegaal (1997) describes a small-scale evaluation of accentuation in GoalGetter, an SLG system for Dutch that generates soccer reports (Theune 1997b; Theune 1997a; Theune, Klabbers, Odijk, and de Pijper 1997). The pitch accent placement algorithm of GoalGetter consists of a set of hand-crafted rules adapted from PROS-3 (Dirksen 1992; Dirksen and Quené 1993), which is in turn based on the linguistic theory developed by Baart (1987). In addition, GoalGetter includes a discourse-based approach to deaccentuation (van Deemter 1994; van Deemter and Odijk 1997). The system was used to generate three texts, comprising 22 sentences, or 211 words. These texts were read aloud by four non-professional speakers. Listening to the recordings, nine experts transcribed the accent locations. Next, the transcribed accents were compared with those predicted by the system. If 8 out of 9 expert listeners transcribed an accent, but the system failed to assign one, this was considered an error. Likewise, if 8 out of 9 listeners did *not* transcribe an accent, but the system assigned one anyway, this was counted as an error as well. This comparison resulted in 7 errors (an agreement of 96.68%) for the speaker that was best according to the listeners, and 10 (95.26%), 8 (96.21%) and 14 (93.36%) errors for the other speakers. The average agreement over all speakers is therefore 95.28%.[5] Most of the errors involved too few accents, indicating that GoalGetter is sparing with its accents. The fact that the agreement scores are considerably higher than those reported in van Bezooijen (1989) may be attributed to the benefits of a more advanced algorithm, and most importantly, to an SLG rather than a TTS context.

Hirschberg (1993) is an extensive study of pitch accent placement in American English using hand-crafted rules and automatic classification techniques. The input consisted of POS tags, the position of primary stress as indicated by a rule-based complex nominal analyser, and informational features resulting from modelling of the local and global focus structure and of contrastive contexts. The hand-crafted rules reached an accuracy of 82.4% on 2 radio news stories comprising a total of 9 minutes of speech from 9 different speakers. The

[4] Contrary to what might be expected, the scores for subjects and newsreader were not close to 10. Van Bezooijen attributed these relatively low values to, among other things, the poor quality of the synthesised speech.

[5] Although a kappa-coefficient of 0.827 is mentioned (Nachtegaal 1997, p. 30), indicating a rather high agreement among the transcribers, the actual agreement is not reported. That is, it is not known for how many of the 211 words 8 out of 9 transcribers agreed on either an accent or no accent. It may be the case, therefore, that many words were not taken into consideration during the comparison. In fact, comparing the data in Appendix 5.C shows that lowering the agreement criterion to 6 out of 9 listeners results in 14 errors (93.36%) for the best speaker, and 19 (91.00%), 14 (93,26%), and 24 (88.63%) errors for the others.

same rules scored a near-perfect 98.3% when applied to 302 short citation-form sentences/phrases (with an average length of 5.5 words, and a total of 1756 words), read by a single speaker under laboratory circumstances. This extremely high performance was attributed to the virtual lack of context for the sentences in this corpus.

In subsequent experiments, Classification and Regression Trees (CART) were used for automatic induction of accentuation models. When applied to the Audix corpus, which consists of 10 news stories read by a female professional newscaster (about 30 minutes of speech), the highest accuracy achieved was 80% (cross-validated). This assumed that intonational phrase boundaries were also available in the input. The same method was applied to a part of the ATIS corpus (about 24 minutes of spontaneous speech by 26 speakers, 298 sentences with an average of 11.5 words per sentence), which consists of speech of users who engaged in a task-oriented dialogue with a (simulated) travel agent. Again assuming the availability of intonational phrase boundaries, the cross-validated accuracy was 85.1%.

Pan and McKeown (1999) describe pitch accent assignment algorithms that employ Information Content (IC), an information theoretic measure for the relative informativeness of a word, which can be automatically derived from a large collection of texts. Their corpus contained 30 minutes of multi-speaker, spontaneous speech of doctors reporting on the post-operative status of a patient. The corpus was orthographically transcribed by a medical expert and intonationally labelled with pitch accents by a ToBI expert. Two methods were used to build pitch accent models on the basis of IC values and POS tags: a rule induction system called RIPPER, and Hidden Markov Models (HMM). Results were obtained using 10-fold cross-validation. RIPPER reached an accuracy of 73.71%, while the HMM accomplished 74.06%. It should be noted, however, that pitch accent placement was a difficult task here, as it concerned spontaneous speech. Pan and McKeown report that the Bell Labs TTS system, which normally scores around 80%, accomplished an accuracy of only 71.75% for this task.

Similar work is described in Pan and Hirschberg (2000), using uni-gram and bi-gram word predictability as a measure for word collocation, which is shown to be a useful predictor of pitch accent placement. The corpora used were in the same medical domain described above: a 50-minute, multi-speaker corpus of spontaneous speech and an 11- minute, single-speaker corpus of read speech. Again, RIPPER was used for automatic rule induction, and results were obtained using 5-fold cross-validation. The results, however, were limited to the prediction of accents on non-trivial base NP's, i.e. non-recursive noun phrases containing at least two elements. The best results were 85.03% accuracy for read speech, and 77.43% for spontaneous speech.

Finally, McKeown and Pan (to appear) describe results obtained with memory-based learning, another machine learning technique. The speech corpus consisted of recordings of a doctor reading aloud five reports that were generated by MAGIC (Medical Abstract Generation for Intensive Care), which is an

SLG system that automatically generates a post-operative status report for a patient who has undergone bypass surgery (McKeown, S.Pan, J.Shaw, Jordan, and Allen 1997).[6] Although at 60.41% the accuracy seemed to be rather low, McKeown and Pan argued that this was due to an unfair bias in their evaluation method. In fact, a subjective evaluation experiment showed that the memory-based system was judged significantly better than both the Bell Labs TTS system and the IC-based model described earlier.

To sum up, the accuracy of recent pitch accent placement algorithms ranges from about 85% for TTS, around 95% in an SLG context, up to 98% for citation-form sentences. Real comparison is difficult because of differences in language, text/speech material, and evaluation method. However, with an average agreement of 93% for TD and 95% for NR, the performance of ProsR is what may be expected for a moderately complex task in an SLG environment.

8.5.3 Comparison with other approaches to phrasing

With respect to intonational phrasing algorithms for Dutch, there appear to be no quantitative results available in the literature to compare the current results with. For English, some results on prosodic phrasing algorithms are available, but a comparison is not straightforward, because different assumptions are made regarding the type of prosodic breaks and the performance criteria.

Taylor and Black (1998) present algorithms for automatically assigning phrase breaks to unrestricted text, intended for use in TTS. First, the input text is converted into a sequence of part-of-speech tags. Next, a Markov-model is used for predicting the most likely sequence of phrase breaks for the the input, also taking the length of phrases into account. Read speech (news stories, weather reports, etc.) from the Spoken English Corpus were used for training. For testing, 10 separate stories were used, containing 7762 words and 1404 breaks. This material was manually labelled with two types of breaks: major and minor breaks. The former corresponds to our intonational phrase boundaries[7], while the latter may be comparable to our phonological phrase boundaries. Most of their results, however, concern the correct prediction of the *location* of prosodic breaks regardless of their *type* (i.e. major or minor), which precludes a comparison with our results.

In one case, however, Taylor and Black report results for a model that predicts both major and minor boundaries. The following measure was used to express its performance:

$$Junctures\ correct = \frac{N - D - S - I}{N} \times 100\%$$

[6] As a result, there is no need for manual labelling of the corpus, because all linguistic information is automatically generated, and the pitch accent model can be trained on detailed and reliable syntactic and semantic features. This shows how an NLG system can be put to use in generating a corpus for training inductive language models.

[7] "Major phrase boundaries delimit intonation phrases and are the only position where boundary tones can occur." (Taylor and Black 1998, Introduction)

N is the total number of junctures (word pairs) in the test set. D is the total number of *deletion errors*: a break that was marked in the reference sentence but not in the test sentence. I is the total number of *insertion errors*: a break that was marked in the test sentence but not in the reference. S is the total number of *substitution errors*: a break that occurred at the right place but was of the wrong type (i.e. major instead of minor or vice versa). The major-minor algorithm reached a junctures-correct score of 90.211%. This score, however, gives no further information about how many *major* breaks were correctly predicted, and can therefore not be directly compared with the agreement indices in Table 8.7.

Taylor & Black introduce two other performance measures called *Breaks-correct* and *Junctures-insertion* (or false insertions w.r.t. junctures), where B is the total number of junctures which are breaks:

$$Breaks\ correct = \frac{B - D - S}{B} \times 100\%$$

$$Junctures\ insertion = \frac{I}{N} \times 100\%$$

These measures are simililar to those of *precision* and *recall* as used in Information Retrieval. Taylor & Black argue that these are better performance indicators than junctures-correct.

> "The difference between breaks-correct and junctures-correct lies in whether non-breaks are included in the calculation. The junctures-correct score gives credit when both the test and reference sentences have a non-break at the same juncture, while the breaks-correct score only looks at junctures with breaks. In our data the number of non-breaks outnumbers the number of breaks by a ratio of about 4:1, and hence an algorithm which marks everything as non-breaks will score about 80% junctures-correct, but 0% breaks correct. [...] Because the breaks-correct score is not dependent on the relative distributions of breaks and non-breaks, we regard this as a better indicator of algorithm performance." (Taylor and Black 1998)

Admittedly, this argumentation make sense. In our case, where as few as 6% of the junctures in TD and 5% of those in in TD were obligatory boundaries (cf. Table 8.6), an algorithm that assigns no boundaries at all would still score about 95% junctures-correct.[8]

[8] Although in principle a similar argument can be made with respect to accentuation, it is far less appropriate, because the number of accented words is usually much closer to the number of non-accented words. Compare for instance Table 8.1, showing that about 40% of the words in TD were accented, and about 61% of those in NR.

Table 8.10: Confusion matrix expressing the agreement in boundary placement in TD between ProsR and human expert CO.

		CO: Obligatory:	Optional:	None:	Total:
	Obligatory:	23	1	5	29
ProsR:	Optional:	0	0	0	0
	None:	32	13	513	558
	Total:	55	14	518	587

As our algorithm predicts only major breaks (intonational boundaries) and no minor breaks, substitution errors (S) did not occur.[9] For example, breaks-correct for Table 8.10 was calculated with a zero substitution error:

$$Breaks\ correct = \frac{B - D - S}{B} \times 100\% = \frac{55 - 32 - 0}{55} \times 100\% = 41.82\%$$

Junctures-insertion was calculated in the following way:

$$Junctures\ insertion = \frac{I}{N} \times 100\% = \frac{5}{587} \times 100\% = 0.85\%$$

Table 8.11 presents the breaks-correct and junctures-insertion scores when ProsR is compared with the human experts in this way. Table 8.12 repeats the results reported in Taylor and Black (1998). At first sight, our results appear to be substantially worse. However, this comparison is rather misleading. Recall from the start of Section 8.4.1 that we only took intonational boundaries *within* sentences into account. That is, we did not count the trivial intonational boundary at the end of a sentence (or phrase in NR) as correctly predicted. In contrast, Taylor & Black tested whole texts:

[9] It should be noted that Taylor and Black's definition of a substitution error cannot be applied to our results. Major and minor breaks refer to boundaries of a different *type*, i.e. with a different acoustic realization and perceptual salience. Our distinction between obligatory and optional boundaries refers to the *occurrence* of a boundary, not to boundaries of a different nature. For instance, substitution of a major for a minor break is different from the substitution of an obligatory for an optional boundary. While the former is clearly an error, the latter is just one of two acceptable realizations of an optional boundary. This also explains why a junctures-correct calculation cannot be applied to the results in Table 8.7.

Table 8.11: Breaks-correct and junctures-insertion when comparing ProsR with human experts, excluding trivial sentence-final breaks.

	Breaks-correct (%):		Junctures-insertion (%):	
	TD:	NR:	TD:	NR:
CO:	41.82	30.77	0.85	0
JC:	43.75	33.33	0.17	0
JH:	48.89	20	0.17	0
MS:	100	50	1.70	0
$\overline{x}$:	58.62	33.53	0.72	0

Table 8.12: Minor and major phrase breaks for a part of the Spoken English Corpus as predicted by the HMM-based prosodic phrasing algorithm described in Taylor and Black (1998).

Model:	Minor cor.:	Major cor.:	Junct. cor.:	Junct. insertion:
1-gram	60.125	56.415	90.128	2.035
6-gram	68.703	59.502	90.211	3.923

Table 8.13: Breaks-correct and junctures-insertion when comparing ProsR with human experts, including trivial sentence-final breaks.

	Breaks-correct (%):		Junctures-insertion (%):	
	TD:	NR:	TD:	NR:
CO:	71.17	87.67	0.78	0
JC:	74.04	91.30	0.16	0
JH:	77.23	84.00	0.16	0
MS:	100	98.39	1.56	0
$\overline{x}$:	80.61	90.34	0.67	0

> "[...] our system is designed and optimised solely for use in a unconstrained text-to-speech system. To that end we test the system on the input format it can expect in real use: a continuous stream of ascii characters with no division into words, sentences or other units. The division of the text into such units is performed automatically within the initial text processing modules of the text-to-speech system." (Taylor and Black 1998)

In the reported experiments, a POS tagger was used that has an overall performance of 97% and includes a tag for final punctuation symbols (period, question mark, and exclamation mark), for which performance is probably near perfect. This means that the end of a sentence was virtually always correctly identified, and thus the reported results include many trivial sentence-final boundaries.[10]

If we include the trivial sentence-final boundaries as well (56 for TD and 60 for NR), we obtain the results in Table 8.13, which clearly looks much better than those in Table 8.11. The average breaks-correct scores compare favourably with the major correct scores in Table 8.12. It should be noted, however, that the latter also include substitution errors (minor instead of major break), which do not occur in our results. Assuming that major break insertions (instead of no break) make up about half of the 2 to 4% junctures-insertion in Table 8.12, junctures-insertion by ProsR is still lower as well.

Ostendorf and Veilleux (1994) present another stochastic approach to prosodic phrasing, using a dynamic programming algorithm for finding the most likely prosodic parse. Their test data consisted of 5 versions of an FM radio news story, containing 23 sentences and 386 words, read by 4 different announcers (2 male and 2 female). Prosodic break levels from 0 to 6 were manually

[10] As a matter of fact, Taylor and Black (1998) describe a deterministic phrasing algorithm which places a phrase break after every punctuation mark in the text (abstracting away from the major-minor distinction). This algorithm has a breaks-correct score of 54.274% and a junctures-insertion score of 0.852%, indicating that punctuation is a very reliable indicator of phrase breaks, and that punctuation nearly always implies a phrase break.

Table 8.14: Confusion matrix for predicted breaks in Ostendorf and Veilleux (1994) (using the hierarchical system, without syntax, and without minor phrase length ratio test).

		Actual:		
		Major:	Minor:	No-break:
	Major:	49	6	5
Predicted:	Minor:	5	12	6
	No-break:	7	10	286
	Total:	61	28	297

assigned by three listeners, and from this minor and major breaks were derived. Their major breaks are equal to intonational phrase boundaries (Ostendorf and Veilleux 1994, p. 28).

The best result achieved by the hierarchical system in Ostendorf and Veilleux (1994) is repeated in Table 8.14. It was calculated in the following way.

> "A predicted sentence is compared to each of the five spoken versions, and the closest spoken version is used as a reference for that sentence (The closeness of parses is measured using a Euclidian distance with 0 for no break, 1 for minor break and 2 for major break.) Then the correspondence between predicted and observed breaks is tabulated in a confusion matrix." (Ostendorf and Veilleux 1994, p. 40)

Using the performance measures proposed by Taylor and Black (1998), this gives a junctures-correct of 89.90%, major-breaks-correct of 80.33% and junctures-insertion of 2.85%. On the one hand, these figures include 5 substitution errors and 6 insertion errors due to *minor* break insertion. A version of the algorithm that assigns only major breaks is likely to perform even better. On the other hand, the figures in Table 8.14 resulted from comparing each sentence with the *best* match of the five available versions. According to Ostendorf and Veilleux (1994, p. 40), the correct prediction and false prediction rates that correspond to the average over the five separate test versions (as the $\overline{x}$ in Table 8.13) are roughly 10% lower. In addition, these results also included sentence-final major breaks[11], which means that 23 of the 49 correctly predicted major breaks were in fact trivial. Taking these reservations into account, we

[11] "Sentence breaks are deterministically assigned at periods, but these are included in the performance results reported here (as major breaks) to be consistent with results reported elsewhere." (Ostendorf and Veilleux 1994, p. 40)

Table 8.15: Confusion matrix for predicted breaks in Ostendorf and Veilleux (1994) (using simple classification trees, with syntax, and without minor phrase length ratio test).

		Actual: Major:	Minor:	No-break:
Predicted:	Major:	57	3	4
	Minor:	0	0	0
	No-break:	5	18	298
	Total:	62	21	302

can tentatively conclude that the major-break-correct score is comparable to ours in Table 8.11, but the junctures-insertion rate is almost twice as high.

Ostendorf and Veilleux also trained classification trees for prosodic phrasing. The results for the best tree are repeated in Table 8.15. This model predicted only major breaks, which makes a comparison easier. We can assume that the distinction between minor and no-break is irrelevant, which means that we collapse the second and third column of the confusion matrix. This gives a junctures-correct score of 91.97%, a major-breaks-correct score of 91.94%, and junctures-insertion score of 1.81%. Excluding the trivial sentence-final major-breaks decreases the major-breaks-correct score to 87.18%. We probably have to subtract another 10% to compensate for the effects of the comparison with the *best* version only, but a breaks-correct of about 77% is still considerably better than our results in Table 8.11.

Finally, Ostendorf and Veilleux also report results for the rule-based phrasing algorithm by Bachenko and Fitzpatrick (1990), which predicts both major and minor breaks.[12] This algorithm produced a junctures-correct score of 86.79%, a major-break-correct score of 77.78%, and a junctures-insertion score of 4.92%. These results stem from a comparison with the best matching sentence of the five versions available, and include sentence-final breaks. The value for junctures-insertion is particularly high in comparison with the other algorithms.

To sum up, comparing our intonational phrasing results with other results

[12] However, their interpretation of prosodic breaks appears to be different from that in Ostendorf and Veilleux (1994): "In the transcriptions, we both distinguished three types of prosodic event: a primary phrase boundary, a secondary phrase boundary, and the absence of a boundary. The most salient characteristic of the primary phrase boundary was a pause, while that of the secondary boundary involved a change in pitch." Bachenko and Fitzpatrick (1990, p. 155). This suggests that the Ostendorf and Veilleux's major breaks (and our intonational boundaries) should actually be compared to Bachenko and Fitzpatrick's *secondary* boundaries. For this reason, we have not attempted a direct comparison of our results with those in Bachenko and Fitzpatrick (1990).

in the literature is not without difficulties. Both Taylor and Black (1998) and Ostendorf and Veilleux (1994) describe algorithms that predict major as well as minor breaks, and the results they report often concern an average over both types, precluding a direct comparison with our results, which concern only major breaks (intonational boundaries). In addition, different evaluation methods and measures are used. For instance, counting the trivial sentence-final major breaks can produce flattered results. Likewise, comparison with the nearest version as in Ostendorf and Veilleux (1994) is certainly defendable, but flatters the results as well. Still, abstracting away from details, our results appear to be somewhat better than those of Taylor and Black (1998) and Bachenko and Fitzpatrick (1990), comparable to those of the hierarchical model in Ostendorf and Veilleux (1994), and somewhat worse than those of the classification tree model in Ostendorf and Veilleux (1994) when breaks-correct scores are considered. In all cases, however, our insertion of erroneous boundaries (junctures-insertion) is considerably lower, indicating that ProsR hardly overgenerates. As a consequence, ProsR has the best performance of all when agreement or junctures-correct scores are considered (again, abstracting away from some details). Since there is a trade-off between breaks-correct and junctures-insertion, this suggests that the performance of ProsR can brought close to that of the classification tree model by inserting more boundaries.

8.6 Summary

This chapter reported on an experimental evaluation of the output of our SLG system CONPAS, more in particular the algorithms for pitch accent placement and intonational phrasing of its prosodic realization module ProsR. The evaluation involved a comparison at the symbolic level of accents and boundaries generated by the system with those assigned by four human experts. The material consisted of eight travel descriptions (TD, 643 words) and 60 expressions containing numerals (NR, 274 words). It was found that ProsR's accentuation had an average agreement of 93% for TD and 95% for NR, while its intonational phrasing showed an agreement of 96% for both domains. Comparison to other approaches suggests that its accentuation is better than that of most TTS systems, and comparable to other SLG systems, while its phrasing is superior to other approaches in terms of agreement, and competitive in terms of other performance measures. Finally, the fact that agreement among the human experts is still higher indicates that there is still room for improvement of accentuation and phrasing models.