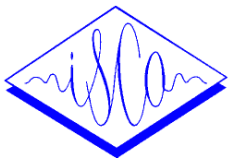

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

A COMPARISON OF NATURAL, THEORETICAL AND AUTOMATICALLY DERIVED ACCENTUATIONS OF DUTCH TEXTS

Hugo Quené and Arthur Dirksen

Research Institute for Language and Speech, Rijksuniversiteit Utrecht
Trans 10, 3512 JK Utrecht, the Netherlands

ABSTRACT

A phrase-level comparison of three transcriptions shows 8% deviation between naturally produced and theoretically predicted accentuations, which is mainly due to variation in pragmatic factors. Deviations between these two templates and the output of an accentuation algorithm are only slightly larger (9% and 11%). Hence, part of the latter deviations may fall within the limits of freedom for pragmatic factors. Some remaining errors could be solved by improving the 'given'/'new' distinction.

1. INTRODUCTION

Adequate prosodic cues are known to contribute considerably to the perception and comprehension of connected speech [e.g. Collier & 't Hart 1975; Cutler & Clifton 1984]. For example, prosody may help the listener to extract the linguistic structure of an utterance, to demarcate words and word groups, and to find the most important information in an utterance. These perceptual functions of sentence prosody become even more important when listening to synthetic speech, which provides fewer segmental cues to the intended speech sounds, words, and meaning. In order to compensate for this reduced speech quality, text-to-speech systems should aim at producing an adequate sentence prosody.

Accentuation is an important factor for sentence prosody. It is determined in part by pragmatics [Bolinger 1972], in part by syntactic and rhythmic considerations. At the pragmatic level, notions such as 'given', 'new', 'prominence' and 'focus' are crucial. Syntactic and rhythmic factors determine how, for example, 'focus' translates into the binary word attribute 'accent' [Gussenhoven 1984; Baart 1987]. While these syntactic and rhythmic rules are deterministic, there is great freedom at the pragmatic level, allowing the resulting accentuation to vary considerably:

(1a) een aantal ONDERZOEKERS # meent overigens ## dat de VRAAG ## of passief meeroken # SCHADELIJK is ## al LANG positief # kan worden BEANTWOORD

(1b) een AANTAL onderzoekers # MEENT overigens ## dat de vraag # of PASSIEF MEEROKEN # schadelijk is ## al lang POSITIEF # kan worden beantwoord
"a number [of] researchers # thinks indeed # that the question # whether passive smoking # injurious is # already [a] long [time ago] positively # can be answered"

This variation is one of the main problems in text-to-speech applications: it is a priori unknown which accentuation an algorithm should produce. Since there is no single 'reference' or 'template' accentuation to imitate, a 'terminal analog' approach to accentuation is impossible.

The algorithm PROS [Quené & Kager 1989] derives accentuation from prosodic domains [Nespor & Vogel 1986], using rules from accentuation theory. Usually, it produces a so-called 'neutral' accentuation (viz. without presuppositions regarding pragmatic factors), unless it finds obvious references to 'given' information (e.g. **zulke** "such"). Although the resulting accentuation may be correct from a theoretical perspective, it may also deviate significantly from the natural accentuation as produced by human speakers.

The primary aim of this research is to investigate the variation between two possible template accentuations which PROS might attempt to produce: natural accentuation as produced by a human speaker, and accentuation as predicted by theory (when correct assumptions regarding pragmatic factors are based on context). In addition, the accentuation automatically derived by our algorithm is compared with each of these two templates.

2. COMPARISONS

All comparisons were based on transcriptions of a corpus of radio news bulletins (meant to be read aloud), comprising 10766 words in 600 sentences, grouped into 43 texts. For the NATURAL accentuation, these texts were read aloud by a professional male speaker, and recorded on audio tape. The recordings were transcribed by the first author and one other transcriber. In addition, major prosodic breaks were transcribed, for reasons explained below.

For the theoretical accentuation (THEORY), the same transcribers followed the combined theoretical rules [Gussenhoven 1984; Baart 1987], marking those words which were predicted to be accented. If both transcribers agreed, then a word was considered to be accented according to theory, otherwise unaccented. In addition, major phrase breaks were predicted analogously, following rules for I-phrase construction [Nespor & Vogel 1986].

The automatically derived accentuation (PROS) was produced by a modified version of the accentuation program. Usually, this program performs three sub-routines for each sentence, viz. (i) syntactic word labeling, (ii) prosodic domain construction, and (iii) accentuation. The modified program can read/write its sentence representation from/to an external file before/after each sub-routine. For the present purpose, sub-routines (i+ii) were performed, after which the resulting analyses were written to file. Subsequently, these syntactic labels and prosodic domain boundaries were verified and (where necessary) corrected by the first author (this is termed the WDT representation). Finally, this representation was fed into the (iii) accentuation sub-routine. Hence, errors during previous sub-routines could not deteriorate the output accentuation.

In the accentuation sub-routine, the major decision the algorithm has to make is whether a prosodic phrase (Phi domain) should be accented. The exact location of a *phrase-accent* on the proper word is determined (rather independently) by syntactic and rhythmic rules. Because the most important variation in accentuation is caused at the phrase level (by pragmatic factors), a phrase-by-phrase comparison was deemed more appropriate. In example (2), the three word-level accent differences can be better described as two (related) differences in phrase-accents:

(2a) de INBREKERS # hebben # de MEESTE DOCUMENTEN # meegenomen
(2b) de INBREKERS # hebben # de meeste documenten # MEEGENOMEN
"the burglars # have # the most [of the] documents # taken along"

In order to allow for such a comparison, prosodic phrase (Phi domain) boundaries were copied from the WDT representation into each of the three accentuation transcriptions (any major prosodic boundaries already transcribed were left intact).

The actual phrase-accent comparisons between two transcriptions were performed by a computer program. If an accented word is found in one transcription, then the program searches for an accented word in the same phrase in the other transcription; if the latter search fails, then a difference is reported and counted. Subsequently, these differences in phrase-accents were inspected and classified by the first author. Five 'error' categories were discriminated: [1] semantics, focus assignment, [2] 'given'/'new' distinction, [3] idiomatic expressions, [4] syntax, [5] other (double classifications were allowed). In some cases, multiple phrase-accent differences are related to a single pragmatic difference, as in example (2). Such related differences were treated as a single 'error', yielding a lower number of *un*related (independent) 'errors'.

3. RESULTS AND DISCUSSION

Results of the comparison between NATURAL and THEORY transcriptions, the two possible templates, are summarised in Table I below. Note that the sum of classified differences exceeds the number of unrelated phrase-accent differences, because differences may be classified as having multiple causes.

Table I: Number of phrases with different accentuation in NATURAL vs. THEORY transcriptions (relative to total number of phrases), and classification of these differences (relative to number of unrelated differences).

transcriptions compared:	NATURAL / THEORY
phrase-accent differences	584 (13%)
unrelated differences	370 (8%)
[1] semantics, focus	96 (26%)
[2] given/new	131 (35%)
[3] idioms	45 (12%)
[4] syntax	89 (24%)
[5] other	21 (6%)

These results show considerable variation between the two template accentuations. In other words, accentuation theory yields an 'error' rate of 8% in phrase-accent predictions. This deviation is mainly due to the variation allowed at the pragmatic level, as shown by the classification results in Table I. In theory, focus assignments and semantic references are fixed, for a particular sentence in a particular context. Nevertheless, the results sub [1] show that a speaker may choose to set the pragmatic parameters differently. For example, a speaker may choose to accent information which is already 'given' to the listener, or to refrain from accenting 'new' information (sub [2]) [Fuchs 1984; Nooteboom & Kruyt 1987]. The latter can be seen in example (3) below:

(3) het MACHTSMONOPOLIE # van de Communistische Partij
"the power-monopoly # of the Communist Party"

In this case, a speaker may assume the second phrase to be implicitly 'given' by the first phrase (at least in the spring of 1990), although both phrases would be 'new' (and hence accented) in a neutral reading.

Table II: Number of phrases with different accentuation in PROS vs. NATURAL and THEORY transcriptions (relative to total number of phrases), and classification of these differences (relative to number of unrelated differences).

transcriptions compared:	NATURAL / PROS	THEORY / PROS
phrase-accent differences	806 (18%)	773 (17%)
unrelated differences	421 (9%)	496 (11%)
[1] semantics, focus	192 (46%)	195 (39%)
[2] given/new	113 (27%)	175 (35%)
[3] idioms	52 (12%)	47 (9%)
[4] syntax	25 (6%)	90 (18%)
[5] other	57 (14%)	26 (5%)

Results of the comparisons between PROS output and the two templates are summarised in Table II above. The number of unrelated differences is only slightly higher than that in Table I. Yet, any incorrect decision of the algorithm affects more phrases, yielding a considerably higher absolute number of differences, as compared to the results in Table I. Again, the deviations are mainly due to incorrect decisions regarding focus assignment, semantic references and 'given'/ 'new' distinctions, as shown by the classification results. The resulting PROS accentuation deviates in an approximately equal amount from both templates (9% and 11% resp.). More importantly, however, this deviation roughly equals the amount of deviation *between* these two tem-

plate accentuations (8%). Hence, this deviation between PROS and its templates might be partially within the limits of freedom allowed for pragmatic parameters. Consequently, these deviations do not necessarily imply that the PROS output is incorrect: it may be perceptually equivalent to the NATURAL and/or THEORY accentuations. Since our comparisons did not involve all three transcriptions simultaneously, however, any conclusions regarding the proportion of differing but equivalent accentuations are unwarranted.

The erroneous (i.e., non-equivalent) accentuations produced by our algorithm are very difficult to correct. The necessary pragmatic knowledge is far beyond the scope of such an algorithm, as exemplified by (3) above. However, we feel that deviations due to incorrect 'given'/'new' distinctions might be partly resolved, by improving our means to detect 'given' information ('new' or <+Focus> is assumed by default). Currently, only explicit references to 'given' information are used [Quené & Kager 1989]. In the future, literal repetitions, and perhaps also synonyms and even hyponyms might be recognised by the program. This would take care of 52 (13%) and 91 (18%) deviations in the two comparisons of Table II sub [2], respectively.

4. CONCLUSIONS

The accentuations compared above were based on different values for pragmatic parameters. This is the primary source of variation between the accentuations, even when these parameters were (1) set by the speaker at the moment of reading (NATURAL), and (2) derived from the context by transcribers (THEORY). If the pragmatic parameters are (3) set by an algorithm (PROS), then their values deviate from both (1) and (2). Yet, part of these 'pragmatic errors' of our algorithm may fall within the limits of freedom for pragmatic parameters. Some remaining 'pragmatic errors' could be solved by improving the 'given'/'new' distinction. However, any person or computer can only set pragmatic parameters perfectly, if he/she/it has a perfect understanding of the full pragmatic context.

ACKNOWLEDGEMENT

This research was supported by the research programme for *Analysis and Synthesis of Speech*, which is funded by the Stimuleringsprojectteam Informatica-onderzoek Nederland (SPIN).

REFERENCES

Baart, J.L.G. (1987) *Focus, Syntax, and Accent Placement: towards a rule system for the derivation of pitch accent patterns in Dutch as spoken by humans and machines.* dissertation, Leiden.

Bolinger, D.L. (1972) Accent is predictable (if you're a mind reader), *Language* 48(3): 633-44.

Collier, R. & 't Hart, H. (1975) The role of intonation in speech perception. In: A. Cohen & S.G. Nooteboom (eds.) *Structure and Process in Speech Perception.* Springer, Berlin. 107-21.

Cutler, A. & Clifton, C. (1984) The use of prosodic information in word recognition. In: H. Bouma & D.G. Bouwhuis (eds.) *Attention and Performance X: Control of Language Processes.* Lawrence Erlbaum, London. 183-96.

Fuchs, A. (1984) 'Deaccenting' and 'default accent'. In: D. Gibbon & H. Richter (eds.) *Intonation, Accent and Rhythm.* De Gruyter, Berlin.

Gussenhoven, C. (1984) *On the Grammar and Semantics of Sentence Accents.* Foris, Dordrecht.

Nespor, M. & Vogel, I. (1986) *Prosodic Phonology.* Foris, Dordrecht.

Nooteboom, S.G. & Kruyt, J.G. (1987) Accents, focus distribution, and the perceived distribution of given and new information: an experiment, *J. Acoust. Soc. Am.* 82:1512-24.

Quené, H., & Kager, R. (1989) Automatic accentuation and prosodic phrasing for Dutch text-to-speech conversion. In: J.P. Tubach & J.J. Mariani (eds.) *Proceedings European Conference on Speech Communication and Technology (EuroSpeech 1989).* Edinburgh, CEP Consultants. volume I: 214-17.