Rule-based Prosody Prediction for German Text-to-Speech Synthesis

Stephanie Becker
Saarland University
Saarbrücken, Germany

Marc Schröder
DFKI GmbH
Saarbrücken, Germany

William J. Barry
Saarland University
Saarbrücken, Germany

Abstract

This paper presents two empirical studies that examine the influence of different linguistic aspects on prosody in German. First, we analysed a German corpus with respect to the effect of syntax and information status on prosody. Second, we conducted a listening test which investigated the prosodic realisation of constituents in the German 'Vorfeld' depending on their information status. The results were used to improve the prosody prediction in the German text-to-speech synthesis system MARY.

1. Introduction

The prediction of appropriate prosody is a crucial task for the synthesis of speech. Generating inadequate prosody seriously hampers intelligibility and naturalness. To a certain extent, the problem can be avoided when using corpus-based synthesis, by selecting units from the appropriate parts of a sentence and thus indirectly generating the correct prosody as recorded in the corpus. However, more recent attempts to generate expressive speech including emphasis or focus require the explicit modelling of prosody [9]. As a basis for modelling expressive speech, it is thus necessary to be able to predict unexpressive speech from linguistic features.

The problem of prosody prediction is by no means solved. Too little is known about the multitude of factors and their interactions that influence the prosodic realisation of a sentence. Factors reported to be relevant include part of speech, position of the word in the sentence, sentence type, various aspects of syntactic structure, and information structure. This list clearly is not exhaustive. Given this large number of potentially relevant variables, a statistically based investigation would have been attractive, not least because it could have provided us with an estimate of the relative importance of the various factors. However, we could not follow the statistical approach because our German MARY text-to-speech (TTS) system [17] uses GToBI [8] for representing prosody, and to our knowledge, no large German corpus annotated with GToBI exists. For that reason, we pursue a rule-based approach, which allows for a very controlled prediction of prosody, and which has the advantage that findings can be interpreted (which is often not the case in statistically trained prediction systems).

The paper is structured as follows. The first section formulates a number of concrete assumptions regarding the links between a variety of linguistic factors and prosody, based on the existing literature. The second section describes the analysis of the German corpus MULI, testing some of the assumptions made in the first section, notably regarding the effects of part of speech, syntax and information status on prosodic realisation. The third section presents a listening test which investigates in more depth the effect of a constituent's information status on its preferred prosodic realisation.

Part of this research was funded by the EU Project HUMAINE (IST-507422).

2. Assumptions based on Previous Work

This section formulates assumptions about German prosody based on a review of some previous studies addressing the link between prosody and linguistic factors. The major part of the literature in the field refers to English prosody, and linguistic prosody prediction can certainly not be transplanted directly from one language to another. However, as both English and German belong to the West Germanic languages and therefore show many similarities, we also use findings about English prosody to formulate testable *assumptions* about German prosody.

Part of speech information is highly related to the probability of a word carrying an accent [21]. It is generally assumed that content words usually receive an accent, whereas function words typically do not (e.g. [10]). Nouns or the heads of the arguments within a sentence, which are usually nominal elements, are the most frequently accented words, whereas verbs or predicates carry accents rather rarely [7], [14].

The position of a word within a sentence appears to have an influence on the form of the accent at least in English. Nuclear accents in declarative sentences are typically falling accents, while words in prenuclear position more frequently carry rising accents [7].

In German, finite verbs in verb second position never carry an accent except for the special case in which the speaker wants to emphasise the truth of the sentence [12].

The relation between syntactic structure and accentuation has been the subject of several studies. In the context of prosody prediction for German TTS, [18] proposed to accent the rightmost noun within noun phrases produced by a chunk parser.

The same authors also suggested that the most embedded verb within a sequence of verbs should receive an accent.

For English, it was found that grammatical subjects are often accented, independently of their information status [19].

For German, speakers tend to realise a boundary after the Vorfeld (sentence initial position preceding the finite verb), at least if the Vorfeld contains three or more words [18].

Furthermore, the assumption was made that chunk phrases correspond to prosodic boundaries [1], [18].

Several studies have shown an interaction between information status and prosody. Given and inferable information appears to be frequently deaccented, whereas new information often receives an accent (e.g., [6], [11]). This was recently confirmed for German, at least for direct objects in nuclear position [4]. Furthermore, the investigation of German question-answer pairs revealed that the deaccentuation of given information in answer sentences is preferred [13]. Additionally, the lexical re-

lation (e.g. synonymy) or the bridging relation (e.g. part-whole) between the given information and its antecedent seems to have an influence on the accent type preferred by listeners [3].

Other studies, however, have failed to confirm the assumption that information status and prosody are related (e.g. [22]). An experiment performed by [19] for English found that given information is only deaccented if the anaphor and its antecedent have the same grammatical role or the same surface position.

In West Germanic languages, an interaction between information status and prosodic *phrasing* is usually not assumed. On the other hand, it was observed in a German corpus that new information is often followed by a prosodic boundary [22].

The fact that contrastive elements are accented appears to be uncontroversial. Several authors assume that the (G)ToBI accent L+H* is appropriate for expressing contrast ([11], [20]).

3. Corpus Analysis

In order to verify the validity of the assumptions derived from the literature in the previous section, we carried out an analysis of the MULI corpus.

3.1. The Corpus

We analysed the corpus elicited in the MULI (MUltiLingual Information structure) [2] project, which examined the means with which information structure is realised in English and German. The German part of the corpus contains 250 sentences stemming from the economics section of the German newspaper Frankfurter Rundschau. The text was spoken by one speaker. As the material is also part of the TIGER Treebank [5], the corpus already contained detailed syntactic information. Some special syntactic information was added, mainly word order information like fronting or extraposition. Prosodic annotation followed the GToBI conventions [8]. The annotation of information status is based on the taxonomy of [16], which distinguishes the statuses "brand new" and "unused", representing new information, "evoked", representing given information (in the sense of coreference with an antecedent), and "inferable". In the case of inferable information, the type of bridging relation between anaphor and antecedent (e.g. part-whole) was also annotated. Additionally, information about lexical relations between anaphor and antecedent (e.g. synonymy, hypernymy) was added. Even though the corpus must be considered very small for our purposes, it appears to be the only German corpus available for which both GToBI and information structure are annotated.

3.2. Method

We tested the various assumptions that arose from the literature survey as summarised in the previous section, using the MMAX framework [15]. For each assumption, we carried out frequency counts of the values of the prospective linguistic predictor variables and the predicted prosodic variables.

3.3. Results and discussion

Part of speech was confirmed as an important predictor for accentuation. Content words frequently carry an accent (81%), and function words are mostly not accented (13%). Proper nouns (90%), adjectives (87%), nouns (86%) and numbers (85%) show the highest accentuation rates.

The surface position of a word has an effect on the type of accent realised on it. In prenuclear position, rising accents (L+H*) are frequent (40%). In nuclear position, falling (H+L*) (44%) and low accents (L*) (28%) were realised more frequently. The H* accent appears in both prenuclear (42%) and nuclear (22%) position.

The assumption that finite verbs in verb second position are never accented could not be confirmed in the corpus. The probability for finite verbs in this position to be accented (27%) was only marginally lower than the general probability for finite verbs to be accented (28%).

Following the hypothesis that the rightmost element within a phrase is accented, we investigated the prosodic realisation of the rightmost element in chunk phrases. In fact, the rightmost element carries an accent very frequently (90%), but the part of speech of a word has more influence on its accentuation: the content words in chunk phrases that are not the rightmost ones, also carry an accent frequently (78%).

The assumption that the most deeply embedded *verb* within a verbal sequence always carries an accent could not be confirmed. The probability for embedded infinitives and participles of full verbs to carry an accent (68%) is approximately the same as the general probability for infinitives and participles of full verbs to be accented (64%). The MULI corpus does not contain any auxiliary verbs appearing in embedded position.

As objects carry accents with roughly the same frequency as subjects (objects: 82.1%; subjects: 82.4%), the tendency for subjects to be accented was not confirmed.

An interaction between the German Vorfeld and prosodic phrasing could be observed. In about 53% of the cases in which the Vorfeld contains three words, a prosodic boundary is realised after the Vorfeld. Furthermore, an increasing length of the Vorfeld is accompanied by an increasing likelihood for the realisation of a boundary after the Vorfeld. A similar observation was made for chunk phrases: with the increase in the length of a chunk phrase, the probability that the chunk is followed by a prosodic boundary also becomes higher. If the chunk phrase contains more than four words, the realisation of a boundary is more probable (55%) than the absence of a boundary.

Regarding information status, we observed that nouns representing new information are frequently accented (brand new: 91%, unused: 93%), but the same holds for inferable (89%) and evoked information (91%). Thus the assumed influence of information status on the prosodic realisation of nouns could not be confirmed in the corpus. Note, however, that personal pronouns, which always represent evoked information, are only accented in 11% of the cases.

The effect of lexical and bridging relations was difficult to interpret, because the number of occurrences in the MULI corpus was small and the number of possible relations large. All relations show a similar distribution across accent types, but more data would be needed to consolidate this observation.

The assumption that given information is deaccented if anaphor and antecedent share the same grammatical role could not be confirmed. This type of given information was accented in 92% of the cases.

When examining the number of prosodic boundaries following new (34%) vs. given (43%) or inferable (35%) information in the corpus, the assumption that new information is more often followed by a boundary could not be confirmed.

The examination of contrastive constituents revealed that they are always accented, most frequently with an L+H* (36%) or an H* (31%) accent. Thus the L+H* accent seems to be an appropriate accent for expressing contrast.

In summary, our analyses of the MULI corpus confirmed some of the assumptions deduced from the literature, such as

the accentuation of content words, the form of prenuclear and nuclear accents or the realisation of boundaries after the Vorfeld or chunk phrases of a certain length. However, a considerable number of the hypotheses could not be confirmed. Particularly with respect to the relation between information status and accentuation, the findings of several authors could not be verified in the corpus. One possible interpretation is that in the economics news texts used in the MULI corpus, the distance between inferable or given constituents and their antecedent is often rather large. This may have had the effect that the information, although actually known, was no longer considered to be sufficiently present in the discourse, with the result that the speaker may have preferred to accent it.

4. Listening Test

The apparent conflict between the findings of our corpus analysis and the literature prompted us to gather complementary information regarding the role of information status for the accentuation of constituents, which we investigated by means of a listening test.

The experiment by [4], which confirmed the deaccentuation of given information in German, tested only constituents in sentence final position. However, in the MULI corpus, the major part of given and inferable constituents appeared at the beginning or in the middle of the sentence and were frequently accented. For English, it was found that grammatical subjects at the beginning of a sentence are more often accented than constituents with other grammatical functions, independently of their information status [19]. Still, the same authors found less accents for given subjects when the grammatical role and surface position of the antecedent was the same.

We therefore designed a listening test in order to investigate the preferred prosodic realisation of grammatical subjects at the beginning of a sentence, more specifically in the German Vorfeld, depending on their information status and on the grammatical role and position of the antecedent of given information.

For information status, we distinguished only new vs. given information. The given constituents were always coreferent with their antecedent. Each test stimulus consisted of two sentences, so that the given constituents always referred to an antecedent appearing in the immediately preceding sentence. We formulated sentences from three different text genres (news, literary style, familiar context) to test whether the genre has an effect on the prosodic preferences.

Our hypothesis is as follows. New information carries an accent and is possibly followed by an intermediate boundary; given information whose antecedent does not have the same grammatical role and position within the sentence is also accented; and given information whose antecedent has the same grammatical role and surface position is deaccented.

4.1. Stimuli

In each of the three genres, we designed three target sentences with a Vorfeld consisting of a two-word noun phrase. Using the diphone synthesis system MARY [17], we generated three prosodic versions of each target sentence: an accented version with a following intermediate boundary (H-, break index=3), an accented version without boundary, and a deaccented version. As previous findings consistently suggested the L+H* accent as appropriate in prenuclear position, we used this accent type for the accented versions.

For each target sentence, we created three context sentences. Two context sentences contained the same information as the Vorfeld constituent of the target sentence so that the Vorfeld constituent in the target sentence refers to given information. In one version, anaphor and antecedent had the same grammatical role and surface position, in the other version, they had different roles and positions. In the third version, the Vorfeld constituent of the target sentence was not already mentioned in the context sentence and thus was new. The target sentence needed to be minimally adapted to be plausible as a follow-up to the different context sentences.

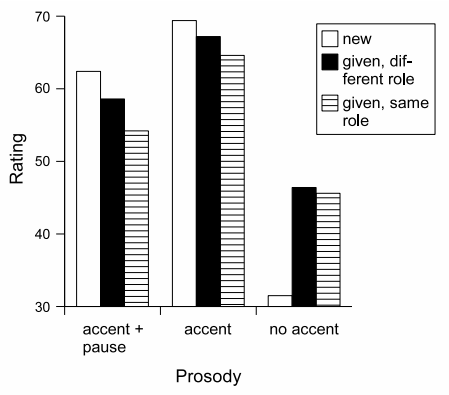

Figure 1: Relation between information status and prosody

4.2. Method

We presented the 27 sentence pairs (3 genres x 3 target sentences x 3 information statuses) in written and in auditory form using the tool 'RatingTest'. For every sentence pair presented in written form on the computer screen, three auditory versions were presented via headphones. Participants were allowed to listen to the versions as often as they wanted to. Just as in training – with two practise trials – they were asked to judge the appropriateness of the sound of the second sentence, especially with respect to the context, i.e. to the content of the first sentence. They were instructed to make their judgements independent of the segmental quality of the speech synthesis. 30 native speakers of German took part in the experiment.

4.3. Results and Discussion

We used SPSS to conduct several analyses of variance.

The analyses showed that the accented version without boundary was always judged to be most appropriate, closely followed by the version with a boundary. This effect was independent of the information status of the constituent in the Vorfeld (see Figure 1). Insofar, the strong formulation of our hypothesis cannot be confirmed.

Nevertheless, a weaker effect in the hypothesised direction was found. It can be seen from Figure 1 that the deaccented version was considered clearly unacceptable for new information while showing medium acceptability for the two types of given information. Beside this, the accented version with boundary, which is the most marked one, received the highest ratings if realised in case of new information. This interaction between information status and prosody is highly significant ($F(4,2421)=21.05$, $p<.001$).

There was no significant interaction between text genre and

any other factors.

In summary, this experiment has shown that grammatical subjects in the Vorfeld are preferably realised with an accent, independently of their information status. One possible reason is that at the beginning of the sentence, speakers often transmit the topic of the sentence. This information appears to be considered so important that it should be realised with an accent. A more subtle effect of information status was observed, however. Deaccentuation is clearly more inappropriate for new than for given information.

Only slight differences were found between the given versions differing in the grammatical role and in the position of their antecedent.

5. Conclusion

We investigated the interaction between different linguistic factors and prosody in German. By analysing a spoken German corpus, we could show that some of the assumptions made in the literature, mainly for English, can also be confirmed for German, but a considerable amount of the assumptions could not. In particular, we could not find a relation between information status and prosody. As the existence of such a relation was experimentally confirmed for sentence final constituents in German, we investigated the preferred prosodic realisation of constituents in sentence initial position, depending on their information status. We found that the accentuation of these constituents is always preferred in German, both for new and for given information, but that the deaccentuation of given information is also acceptable to some degree. By implementing our findings in the TTS system MARY, we obtained better speech synthesis results.

As the analysed corpus is not very large and is spoken by one speaker only, our results cannot claim to be representative of German prosody in general. Further investigations of spoken language with respect to the factors that influence the prosodic realisation in German are needed. New findings in the field could be used to improve the prosody prediction in TTS systems.

6. References

[1] S. Abney. Syntactic Affixation and Performance Structures. In D. Bouchard and K. Leffel, editors, *Views On Phrase Structure*. Kluwer Academic Publisher, 1990.

[2] S. Baumann, C. Brinckmann, S. Hansen-Schirra, G. Kruijff, I. Kruijff-Korbayová, S. Neumann, and E. Teich. Multi-Dimensional Annotation of Linguistic Corpora for Investigating Information Structure. In *Proc. of Frontiers in Corpus Annotation Workshop at HLT-NAACL 2004*.

[3] S. Baumann and M. Grice. Accenting Accessible Information. In *Speech Prosody 2004*, pages 21–24, Nara, Japan.

[4] S. Baumann and K. Hadelich. On the Perception of Intonationally Marked Givenness after Auditory and Visual Priming. In *Proc. of the AAI Workshop "Prosodic Interfaces", Nantes*, pages 21–26, 2003.

[5] S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. The TIGER Treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*, pages 24–41, Sozopol, Bulgaria, 2002.

[6] G. Brown. Prosodic Structure and the Given/New Distinction. In A. Cutler and D. R. Ladd, editors, *Prosody: Models and Measurements*, Berlin, 1983. Springer Verlag.

[7] A. Cruttenden. *Intonation*. Cambridge University Press, Cambridge, 1991.

[8] M. Grice, S. Baumann, and R. Benzmüller. German intonation in autosegmental-metrical phonology. In S.-A. Jun, editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*, pages 55–83. OUP, 2005.

[9] W. Hamza, R. Bakis, E. Eide, M. Picheny, and J. Pitrelli. The IBM Expressive Speech Synthesis System. In *Proc. of the 8th International Conference on Spoken Language Processing*, Jeju, Korea, 2004.

[10] J. Hirschberg. Pitch Accent in Context: Predicting Intonational Prominence from Text. *Artificial Intelligence*, 63(1-2):305–340, 1993.

[11] J. Hirschberg and J. Pierrehumbert. The Meaning of Intonational Contours in the Interpretation of Discourse. In P.R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 271–311, Cambridge, Massachusetts, 1990. MIT Press.

[12] T. Höhle. Über Verum-Fokus im Deutschen. In J. Jacobs, editor, *Informationsstruktur und Grammatik / Linguistische Berichte Sonderheft 4*, pages 112–141, Opladen, 1992. Westdeutscher Verlag.

[13] C. Hruska and K. Alter. How Prosody Can Influence Sentence Perception. In A. Steube, editor, *Information Structure: Theoretical and Empirical Aspects*, pages 211–226, Berlin, 2004. de Gruyter.

[14] D. R. Ladd. *Intonational Phonology*. Cambridge University Press, Cambridge, 1996.

[15] C. Müller and M. Strube. Muli-Level Annotation in MMAX. In *Proc. of the 4th SIGdial Workshop on Discourse and Dialoguey*, pages 198–207, Sapporo, Japan, 2003.

[16] E. Prince. Towards a Taxonomy of Given-New Information. In P. Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, 1981.

[17] M. Schröder and J. Trouvain. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6:365–377, 2003.

[18] A. Schweitzer and M. Haase. Zwei Ansätze zur syntaxgesteuerten Prosodiegenerierung. In *Tagungsband der Konvens 2000 - Sprachkommunikation*, pages 197–202, Berlin. VDE-Verlag.

[19] J. Terken and J. Hirschberg. Deaccentuation of Words Representing Given Information - Effects of Persistence of Grammatical Function and Surface Position. *Language and Speech*, 37(2):125–145, 1994.

[20] U. Toepel and K. Alter. On the Independence of Information Structural Processing from Prosody. In A. Steube, editor, *Information Structure: Theoretical and Empirical Aspects*. De Gruyter, 2004.

[21] C. Widera, T. Portele, and M. Wolters. Prediction of Word Prominence. In *Proc. of EUROSPEECH 1997*, pages 999–1002, Rhodos.

[22] M. Wolters and H.-J. Mixdorff. Evaluating Radio News Intonation: Autosegmental versus Superpositional Modelling. In *Proc. of the International Conference on Spoken Language Processing*, volume 1, pages 581–584, Beijing, China, 2000.