Evaluating Metrical Phonology – A Computational-Empirical Approach
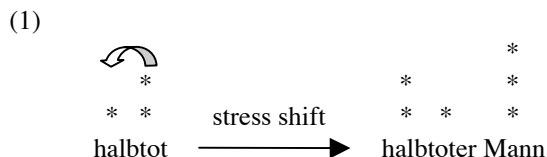
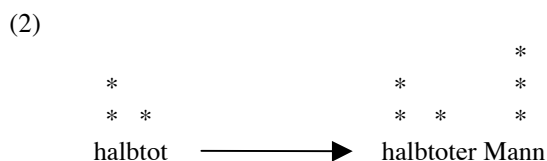**Petra S. Wagner, IKP, Universität Bonn, Germany**

Abstract

This study aims at providing an empirical basis for the evaluation of predictions made by metrical phonology. The predictions are compared to perceptual syllable prominence annotated in a database of German read speech. An evaluation baseline was defined on the correlation between prominence ratings for different speakers. Then, different sets of rules for prominence prediction were tested and evaluated. The results indicated that a correct prediction of word accent is most crucial for good results. Another outcome was, that a prediction based on lexical class is more successful than a purely syntactic approach. Some issues concerning the status of euphony rules in German are discussed.

1. Motivation

Metrical grids are supposed to reflect the relative syllable prominences within a prosodic constituent. The well-formedness of metrical trees and grids usually lacks empirical support other than intuitive judgements. But intuitions concerning stress patterns differ. This circumstance can cause contradictory analyses, as it is the case for stress clash: Kiparsky [9] analyses the prominence pattern in the phrase "halbtoter Mann" on the basis of a stress shift between the syllables "halb" and „tot". In Kiparsky's analysis, the syllable "tot" carries the main stress in the compound "halbtot". If the compound is followed by the noun „Mann" the nuclear stress rule (NSR) assigns the main stress on the phrase final word "Mann", and causes a stress clash with the syllable "tot". So the main stress of the compound shifts from „tot" to „halb" and the secondary stress shifts from "halb" to "tot":

(1)

```
                                            *
         *                      *           *
       *   *     stress shift    *   *   *
     halbtot    ─────────────▶  halbtoter Mann
```

In the leading pronunciation dictionary for German [10], the compound „halbtot" is described as carrying the main stress on the first syllable. Following this assumption, the stress shift rule is no longer necessary for the explanation of the stress pattern in „halbtoter Mann". The pattern follows from the NSR directly:

(2)

```
                                            *
         *                      *           *
       *   *     ─────────────▶  *   *   *
     halbtot                    halbtoter Mann
```

Besides the problem of building theories on contradictory intuitions, the poor empirical basis prevented phonologists from defining detailed rules and constraints relating to secondary and tertiary German lexical stress [6].

The study presented here aims at providing an evaluation model for phonological rule building by comparing predictions of prominence patterns to those actually perceived by native speakers. To follow this aim, several steps need to be undertaken. First, a perceptual measure needs to be defined and data needs to be annotated according to it. Then, it needs to be shown that the perceptual measure is indeed comparable to the one predicted by metrical phonology. Since a theory cannot be expected to predict all the speaker-specific variances, a baseline has to be defined on the inter-speaker correlation of syllable prominence.

2. The Database

Based on the method introduced by [3], each syllable in a large prosodic database of German read speech [7] was labelled by three subjects for its perceived prominence. The database contains 227 sentences and three stories and was read by three speakers (10661 syllables in total). The prominence was annotated by three listeners on a grid which was later transformed into a scale between 0 and 31. High correlations between listeners were found [8]. In order to even out listener-specific effects in prominence ratings, subsequent analyses were based on the median prominence ratings of all three listeners.

Besides, the database contains annotations for locations and strength of prosodic phrase boundaries, lexical class, lexical stress, a variety of linguistic annotations, and different segmental and suprasegmental ones.

3. Comparison of Predicted and Perceived Prominence Patterns

3.1 Comparability

Before building a model of comparison of phonological and phonetic prominence patterns, it needs to be ensured that both measures are indeed related. The phonetic prominence measure is interval scaled but can be interpreted on a ordinal scale as well. We believe that any metrical grid is ordinal scaled as well since it is supposed to reflect the prominence of each syllable of an utterance relative to the surrounding syllables. Thus, a comparison can be based on an ordinal scale. Unless the phonological predictions are far from correct, the hypothesis should hold, that predicted and perceived prominence ranks correlate substantially.

3.2 Baseline Definition

A theory of metrical phonology cannot be expected to make predictions which are correct for any speaker in any situation. Speaker specific variations need to be taken into account. Studies on German nominal compound stress have dealt with speaker variability and detected significant differences between both speakers [1] and dialects [2]. Therefore, the inter-speaker prominence correlation for identical sentences was used as a baseline for predictive accuracy. The average inter-speaker correlation in our database was high (ρ=0.78, p<0.0001). Figure 1 illustrates the of prominence ratings of the sentence "Er ist weggelaufen" for three different speakers.

3.3 A First Rule Set

A phonological algorithm for prominence prediction was taken from Uhmann [12]. The rules appeared to be a good starting point for a test of our initial hypothesis that perceived and predicted prominences correlate, because they predict prominence patterns up to the intonational phrase level and our phonetic data does not contain any words spoken in isolation. Also, there appears to be wide agreement among phonologists concerning most of the rules. No sophisticated theory of German lexical stress is included in the algorithm. The original rules were somewhat simplified concerning the detection of focus exponents, because such an annotation would have needed a deep semantic analysis, which is extremely difficult to undertake for a large database. The initial rule set looked like this:

(i)   Every syllable receives one beat.

(ii)   Every syllable whose nucleus consists not only of an unreduced vowel or a syllabic consonant (ə, ɐ, n̩, m̩, ŋ̩), receives a beat.

(iii)   The first unreduced syllable of each native word receives a beat, the last unreduced syllable of each loan word receives a beat. (lexical stress assignment)

(iv)   Each syllable carrying lexical stress in a content word receives a beat. (simplified stress assignment to potential focus exponents)

(v)   The last syllable in the intonation phrase carrying lexical stress and being a potential focus exponent receives a beat. (NSR)

Table 1: First rule set for predicting prominence
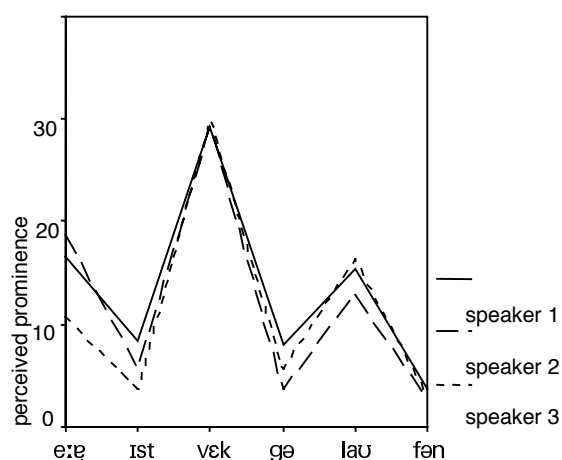
Figure 1: Median prominence ratings for three different speakers of the sentence „Er ist weggelaufen."

The rules were implemented and used to generate prominence patterns for all utterances contained in the database. No euphony rules (rules which make sure that a language-specific rhythmical pattern is obeyed) were implemented since no reliable conditions concerning their proper application could be detected. To determine the predictive power of the rule set, a correlation analysis was calculated based on the ranks of perceived and predicted syllable prominence. Both measures correlate substantially (Spearman, ρ=0.64, p<0.01). This result supported our initial hypothesis that both measures are indeed related to each other. This confirms our approach to an empirical evaluation.

3.4 Qualitative Evaluation

In order to detect weaknesses in the rule set, phrases with a correlation ρ<0.6 between predicted and perceived prominence, were examined further. An overview of this

evaluation is presented in Table 2.

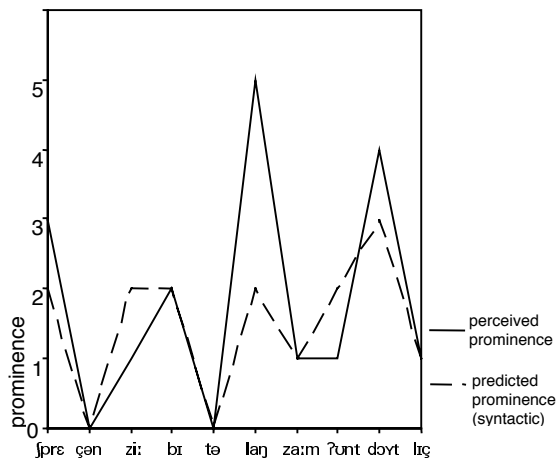| wrong lexical stress	35%
function words perceived with higher than predicted prominence	25%
unpredicted prosodic focus	17%
last heavy syllable (no carrier of lexical stress) more prominent than predicted	5%
missing euphony rules to avoid rhythmical gaps	10%
unclear	8%

Table 2: Reasons for wrong predictions made by the initial rule set

It turned out to be the case that most low correlations between predicted and perceived prominence pattern were explicable (92%). The majority of the cases were caused by wrongly predicted lexical stress (35%) or the wrong word prominence (25%). Since no sophisticated rules for predicting lexical stress had been included, this was not surprising. Very often, specific function words were perceived much more prominent than predicted. This indicates that the simple differentiation into function and content words might not be adequate. Apparently, specific function words are inherently more prominent than others. The latter appears to be the case for affirmative particles, demonstrative and interrogative pronouns). In some cases, the speakers interpreted the sentence in a way which resulted in a prosodic focus on a word other than the last content word in the utterance (which would be the typical location for it according to a „wide focus" interpretation.) Another 10% of the low correlations are the obvious result of missing euphony rules, mostly of those preventing rhythmical gaps. Clear cases of an avoidance of stress clash were not detectable. Interestingly, in 5% of the badly predicted cases, the last heavy syllable was perceived much more prominent than expected, even if it was not carrier of lexical stress. In a few instances, even a schwa-syllable was perceived as quite prominent in utterance-final position. This indicates that the domain of the NSR is not only the last lexical stress on a content word but ought to be extended to at least all unreduced utterance-final syllables.

3.5 Comparison of a syntactic and a semantic approach to sentence stress

The analysis using a first set of rules lead to the conclusion, that there are still a number of uncertainties related to a precise prediction. A complementary approach to the one by Uhmann (regarding content words as potential focus exponents) is the prediction based on the syntactic structure. One algorithm predicting intonation phrase level stress without the need of relating to semantic content is described by Féry [4]. The predictive accuracy of both approaches was compared on a subset of the database (one speaker). This time, lexical stress was assigned on the basis of the annotations in the database, because the tested rules were not designed for lexical stress.

3.5.1 A semantic approach to sentence stress

The rules predicting semantically based utterance level stress were identical to the rules in Table 1, apart from the lexical stress assignment, which was based on the annotations in the database.

3.5.2 A syntactic approach to sentence stress

The implemented rules are identical to the ones in 3.5.1 up to lexical stress level. Then, the stress is assigned in the following way:

Rules for a syntax-based assignment of stress:

(i)   the rightmost constituent of two or more syntactic sister nodes receives (at least) one more beat than its co-constituent(s).

(ii)   calculate step (i) for two syntactic levels, beginning two levels below the start-node.

Figure 2: Syntax-based prominence assignment

The original rule took into account all syntactic levels but was slightly changed to simplify the (manual) annotations. Since most syntax trees did not consist of more than 3 levels, this is not regarded as a significant problem. The syntactical analyses were based on the results of the HPSG-based syntax parser described in [11]. In those cases where several syntactic analyses were found, the one corresponding best to the interpretation of the annotater was chosen. An example for sentence stress assignment using the rules above is illustrated in Figure 2.

Figure 3: Predicted and perceived prominence using a syntax-based prominence prediction.
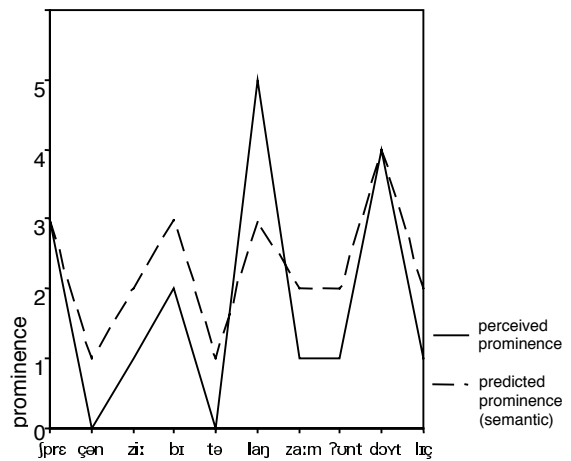

Figure 4: Predicted and perceived prominence using a semantics-based prominence prediction.

3.5.3 Results

Both rule sets were compared to a prediction based purely on the lexical phonological structure of each syllable (syllable weight and lexical stress). Both approaches showed an improvement over the phonological approach. However, the rules based on lexical semantics were clearly superior. One possible reason for this might be the fact that Uhmann's rule system contains a NSR which is also syntactically motivated. But since the rightmost constituent of a syntactical phrase often coincides with a content word, the syntactic approach also partially overlaps with the semantic one. It is concluded, that syntactic structure appears to be less important for an appropriate prediction of prominence as long as the NSR is obeyed. Probably, as previous studies have shown, syntax interacts more directly with location and strength of prosodic boundaries. In Table 3, the predictive accuracy of the different rule sets are compared. Figures 3 and 4 show predicted and perceived prominences of both approaches for a sample sentence. For reasons of illustration, the phonetic prominence was mapped onto a scale of 6 prominence levels.

Rule Set based on:	Correlation between predicted and perceived prominence (speaker 1)
phonology	$\rho=0.64$, $p<0.001$
phonology + syntax	$\rho=0.72$, $p<0.0001$
phonology + semantics	$\rho=0.76$, $p<0.0001$

Table 3: Predictive power of different approaches to prominence prediction

Given the perceived prominences of speaker 1, the semantically motivated rule set is already very close to the baseline correlation of $\rho=0.78$. However, the correlation is lower ($\rho=0.73$) taking into account the entire database. Therefore, a further improvement of the rules remains necessary.

4.   Revised Rule Set

From the previous analyses we learned that an assignment of prominence based upon a distinction into function and content words is not sufficient for a reliable prediction of word prominence, since specific function words are more prominent than others. But we could also show, that the semantic approach was more successful than a syntactic one. Research on speech synthesis [14] has already dealt with lexical class specific prominence assignment, also taking into account simple syntactical constraints (word class of contiguous words, position in utterance). Based on these results, an alternative approach towards prominence prediction is built and tested.

The rules are built on the assumption that each word class can be assigned an inherent prominence value. The syllable carrying lexical stress receives prominence value. Words with a high inherent prominence tend to be affected less by syntactic context. Also, words with a high inherent prominence tend to be less prominent when followed by another word of the same prominence, whereas words which are low in prominence tend to be more prominent in such contexts. For a detailed description of the rule system, cf. [12].

4.1 Results and Discussion

A correlation analysis calculated on the entire database resulted in a $\rho=0.75$ ($p<0.0001$) between predicted and perceived prominence. Figure 5 shows a comparison of predicted and perceived prominence with the revised rule set.

The outcome is better than the one only using a rough differentiation into function and content words. The correlation is close to the baseline. An informal evaluation made evident that most cases where predicted and perceived prominence still deviate significantly, are explicable by either unexpected prosodic focus or missing euphony rules. The former are difficult to predict without any contextual or deep semantic analysis. The latter may be captured through further studies. As before, stress shift did not play any role, but there is a clear dominance of binary rhythmic patterns apparent in the data.
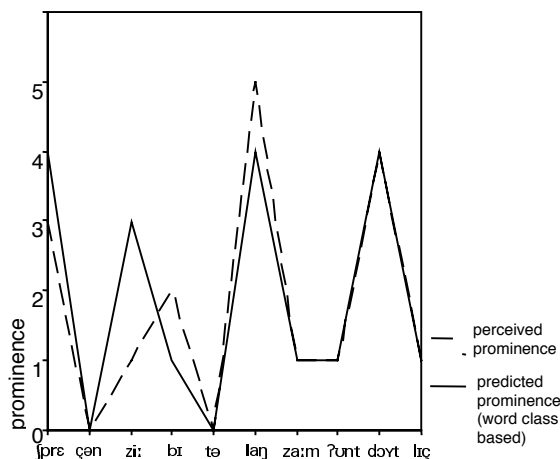
Figure 5: Predicted and perceived prominence using a revised prominence prediction

5. Conclusions

The presented study ought to have shown the following: There is a strong link between perceived prominence and the metrical patterns predicted by phonological theories. A very simple approach to the prediction of perceived prominence is quite successful when it takes into account syllable weight, lexical stress and class. The influence of syntax on perceived prominence appears to be less important, as long as well-known concepts such as the NSR are obeyed. Results can be improved, if a mere differentiation into function vs. content words is exchanged by a more sophisticated approach employing lexical class specific prominence values. The presented approach can already make very precise predictions without referring to for sentence semantic or contextual analyses.

Considering the poorly understood status of euphony rules in German [5,15], it was found out that stress shift obviously plays only a minor role in German, whereas a clear preference for alternating stress patterns is evident.

6. Further Work

With a working tool of evaluation on hand, further analyses are planned. So far, our approach was mainly concerned with the prediction of prominence within intonational phrases. The difficult issue of predicting lexical stress was barely touched. Here, a lot of work remains to be done.

A study of semantic and pragmatic factors resulting in a deaccentuation of „given" words or highlighting of focus exponents may prove valuable, but a different kind of data would be necessary for a detailed examination, since the utterances would have to be placed into contexts which are designed for this purpose. A significant amount of the wrong predictions appear to be due to the lack of euphony rules. These are currently studied and specified [13].

7. Acknowledgements

8. Literature

[1]   Wilbur Benware. Accent variation in German nominal compounds of the type (A(BC)). *Linguistische Berichte* 108, 1987:102-127.

[2]   Ursula Doleschal. Zum deutschen Kompositionsakzent. Tema con variationi. *Wiener Linguistische Gazette* 40/41, 1988:3-28.

[3]   Gunnar Fant and Anita Kruckenberg. Preliminaries to the study of Swedish Prose Reading and Reading style. *STL-QPSR 2/1989*, 1989:1-68.

[4]   Caroline Féry. Metrische Phonologie und Wortakzent im Deutschen. *Studium Linguistik* 20, 1986:16-43.

[5]   Caroline Féry. Rhythmische und tonale Struktur der Intonationsphrase. In: H.Altmann (ed.). *Intonationsforschung*. Tübingen:Niemeyer, 1988:41-64.

[6]   Carsten Günther. *Prosodie und Sprachproduktion*. Tübingen, Niemeyer:1999

[7]   Barbara Heuft, Thomas Portele, Florian Höfer, Jürgen Krämer, Horst Meyer, Monika Rauth and Gerit Sonntag. Parametric description of F0-contours in a prosodic database. *Proceedings of ICSLP '95*, Stockholm, Sweden, vol. 2, 1995:378-381.

[8]   Barbara Heuft. *Eine prominenzbasierte Methode zur Prosodieanalyse und –synthese*. Computer Studies in Language and Speech 2. Peter Lang, Frankfurt: 1999.

[9]   Paul Kiparsky. Über den deutschen Akzent.

Untersuchungen über Akzent und Intonation im Deutschen. Studia Grammatica VII. Akademie-Verlag, Berlin, 1966: 69-98.

[10]   Max Mangold. *Duden Aussprachewörterbuch. Ein Wörterbuch der deutschen Standardaussprache.* Mannheim, Wien, Zürich, Dudenverlag:1990.

[11]   Stefan Müller. The BABEL system – An HPSG Prolog implementation, Overview. *Proceedings of the Fourth International Conference on the Practical Application of PROLOG.* London, UK:1996.

[12]   Susanne Uhmann. *Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie.* Linguistische Arbeiten 252. Niemeyer, Tübingen:1991.

[13]   Petra Wagner. Rhythmical Alternations in German Read Speech. *Special Session on Prosodic Modelling and Transcription at Prosody 2000*, Kracow, Poland:2000.

[14]   Christina Widera, Thomas Portele and Maria Wolters. Prediction of Word Prominence. *Proceedings of EUROSPEECH ´97.* Rhodes, Greece, 1997:999-1003.

[15]   Richard Wiese. *The Phonology of German.* Clarendon Press, Oxford:1996.