What makes a word prominent? Predicting untrained German listeners' perceptual judgments

Baumann, Stefan; Winter, Bodo

DOI:
10.1016/j.wocn.2018.05.004

Document Version
Peer reviewed version

Citation for published version (Harvard):
Baumann, S & Winter, B 2018, 'What makes a word prominent? Predicting untrained German listeners' perceptual judgments', *Journal of Phonetics*, vol. 70, pp. 20-38. https://doi.org/10.1016/j.wocn.2018.05.004

What makes a word prominent?

Predicting untrained German listeners' perceptual judgments

Stefan Baumann[1] & Bodo Winter[2]

[1] I*f*L Phonetik, University of Cologne, Germany (stefan.baumann@uni-koeln.de)

[2] Department of English Language and Applied Linguistics, University of Birmingham, United Kingdom (bodo@bodowinter.com)

Corresponding author: Stefan Baumann
IfL Phonetik, University of Cologne
Herbert-Lewin-Str. 6,
50931 Köln
Germany
Tel: 0049-221-470-4259, Fax: 0049-221-470-5938
stefan.baumann@uni-koeln.de

Highlights

- Both prosodic and non-prosodic cues determine prominence perception
- All 17 variables tested were shown to affect prominence
- Pitch accent position and type are strongest determinants of prominence
- Listeners fall into two groups: pitch-guided listeners and lexical-syntactic listeners
- Prominence is a "redundant" system, being signaled by multiple cues

Abstract

One important feature of linguistic communication is that some parts of utterances are more prominent than others. Prominence as a perceptual feature of spoken language is influenced by many different linguistic variables, but it is not clear how these variables interact in perception and what variables are most important for determining prominence. We report results from a prosody transcription task which assessed how untrained German listeners are simultaneously affected by gradient signal-based factors such as pitch, intensity and duration, as well as discrete prosodic factors (pitch accent type and placement) and non-prosodic factors (semantic-syntactic, lexical). All 17 linguistic variables tested were reliably associated with listeners' prominence judgments. We used random forests, a data mining algorithm, to uncover which variables are most important in determining the prominence judgments. This analysis showed that discrete prosodic variables relating to intonational phonology, specifically the type of pitch accent and its position, were most predictive of prominence. However, how much these factors matter differed between listeners, with prominence judgments being characterized by large individual differences. An exploratory cluster analysis suggests that some listeners pay more attention to prosodic variables (but less to semantic-syntactic and lexical variables), while others do the reverse. Our results paint a complex picture of prominence perception that is highly variable across listeners.

Keywords: Prominence, intonation, perception, random forests, individual differences, German

1. Introduction

When speakers communicate with each other, not all information is equally important. Some parts of an utterance are intrinsically more informative, such as novel discourse topics and uncommon words, while some parts are actively highlighted by speakers as being important through prosodic and syntactic means. As a result of both semantic-pragmatic *importance* and prosodic and syntactic *highlighting* (Streefkerk, 2002), listeners perceive certain utterance parts as more or less prominent. Loosely defined, "perceptual prominence" refers to any aspect of speech that somehow "stands out" to the listener.

As an analogy for prominence in speech, we may consider a tree standing alone on an empty field. This tree is more prominent than a tree in a forest, since it differs in height, shape and color from its environment. What determines perceptual prominence in speech is much less well understood. Already at the level of language structure, there is a host of potential cues for prominence, including the speaker's choice of words, syntactic constructions, and pitch accents. Then, within the more circumscribed domain of prosody, many phonetic variables are associated with prominence, including pitch movement, loudness, duration and voice quality. These different cues may interact in complex ways, and they may have different effects on different listeners (Cole, Mo, & Baek, 2010a; Cole, Mo, & Hasegawa-Johnson, 2010b). Our paper investigates this multi-layered network of prominence cues at the level of the individual word within a sentence.

It is currently still unclear which linguistic variables have the strongest impact on the perception of prominence (but see Wagner, Tamburini, & Windmann, 2012; Arnold, Wagner, & Baayen, 2013; Wagner et al., 2015). For the sake of the present discussion, we distinguish between (1) continuous-valued prosodic parameters, (2) contrastive prosodic

categories and (3) non-prosodic factors. Of course, we have to recognize that not all linguistic variables can neatly be categorized into one of these three groups.

By continuous-valued prosodic factors we understand those acoustic parameters that are signal-driven, such as intensity, fundamental frequency (F0) and duration. As contrastive prosodic factors, we classify those discrete and/or symbolic variables that relate to intonational phonology, such as the particular pitch accent types described in the German Tones and Break Indices system (GToBI; Grice, Baumann, & Benzmüller, 2005). As with other ToBI systems, GToBI characterizes pitch accents as discrete and abstract phonological elements that mediate between the actual phonetic elements they are composed of and their associated linguistic meanings (cf. Cole & Shattuck-Hufnagel, 2016; Cangemi & Grice, 2016). Finally, non-prosodic factors include semantic, syntactic and lexical variables. These relate to word choice or choice of syntactic structure (e.g., focus particles, part-of-speech differences, and word frequency).

Our goal in this study is to investigate the impact of these three classes of linguistic variables on prominence perception. We are furthermore assessing interrelations between the linguistic variables (i.e., which variable has the strongest influence on perceived prominence?) and potential differences in the perception strategies of listener groups (such as 'pitch listeners' versus 'spectral listeners'; Schneider & Wengenroth, 2009). Our study aims to contribute to the study of prominence both in terms of theory (descriptive and theoretical generalizations of prominence cues in German) and in terms of methodology (showing how multiple analytical techniques can be synthesized to get a more comprehensive picture of prominence perception).

2. Background

2.1. Prominence cues

The domain of investigation for most of our non-prosodic variables is the word. However, much of the past literature on prominence has focused on the level of the syllable within the word, in particular the large number of studies on lexical stress[1]. In fact, one of the aims of the present study is to examine whether what is known about syllable-level prominence also holds for word-level prominence within an utterance.

In 'stress-accent languages' (Beckman, 1986) such as English, syllables within a word are either strong (stressed) or weak (unstressed). Several correlates of stress in English and other Germanic languages have been identified. First, vowel quality and other segmental features in unstressed syllables are reduced compared to stressed syllables. The segments of stressed syllables and words generally tend to be hyperarticulated in order to enhance their perceptual clarity, at least in contrast to unstressed syllables (De Jong, 1995). Second, stressed syllables have more local pitch movement (Fry, 1958; Sluijter & van Heuven, 1995). Third, stressed syllables are longer in duration (Fry, 1955; Turk & Sawusch, 1996). Fourth, stressed syllables have overall higher intensity (Fry, 1955; Lea, 1977; Rietveld, 1984; Kochanski, Grabe, Coleman, & Rosner, 2005), which results in the perception of increased loudness. Fifth, stressed syllables have shifted spectral balance, with higher intensity in high-frequency components (Sluijter & van Heuven, 1996).

[1] Strictly speaking, we have to talk of 'post-lexical stress', because we are dealing with concrete prominence at the utterance level. The term 'lexical stress' is often used when addressing abstract strength relations of syllables in words (see e.g., Ladd, 2008). Thus, the studies discussed here investigate post-lexical (acoustic) cues for the detection of lexical stress.

These studies on lexical stress show that several cues help the listener determine which syllable within a word is more prominent than another. In terms of perception, several of these cues have been shown to play a role, but to differing degrees. In particular, loudness and vowel quality have been shown to be relatively weak cues in earlier work on English (Fry, 1955, 1958, 1965), although Sluijter, van Heuven and Pacilly (1997) find that in Dutch, it is not overall intensity (uniform across the frequency spectrum) but particularly intensity in high frequency components that matters for the perception of lexical stress. Beckman (1986) claims that duration and intensity do not act independently as correlates of prosodic prominence, both in production and perception. In a perception experiment on American English, she found that the most dominant cue for stress recognition was what she coined 'total amplitude', a factor that combines duration and intensity into a single acoustic category (also in line with the results of Kochanski et al., 2005).

With respect to phonological factors, we are particularly interested in pitch accent types as classified by the German Tones and Break Indices system (GToBI). This annotation scheme aims at describing 'Standard German' and, like other ToBI systems, has its roots in autosegmental-metrical phonology (see Beckman & Hirschberg, 1994; Beckman et al., 2005; Ladd 2008; online guidelines for American English ToBI: Veilleux, Shattuck-Hufnagel, & Brugos, 2006; for GToBI: Grice, Baumann, Ritter, & Röhr, 2017). Within the ToBI framework, a major distinction is made between pitch accents and boundary tones. These are classified according to two communicative functions: Whereas pitch accents, which are associated with stressed (metrically strong) syllables, serve to highlight relevant constituents, boundary tones, which are associated with phrase-final syllables, serve to delimit prosodic phrases. The tonal inventory of GToBI makes use of

two basic levels, H(igh) and L(ow) tones – marked by a star (*) to show the association with a stressed syllable and by a minus (-) or percent sign (%) to indicate the association with a (minor or major) boundary. The actual tone values are relative to the pitch range that a speaker exploits.

Both the position and the type of pitch accent have been shown to influence prominence perception. With respect to accent position, it has long been claimed — both in the British tradition of intonation analysis (e.g., Halliday, 1967) and in the American generative tradition (e.g., Chomsky & Halle, 1968) — that the last stress or accent in an utterance is most prominent, i.e., the nuclear stress or accent. This structural or positional prominence is usually regarded as most important for an appropriate interpretation of the whole utterance, which is in turn the basis of many studies investigating the relation between accentuation and focus (e.g., Gussenhoven, 1984; Selkirk, 1984; Uhmann, 1991). Prenuclear accents have often been regarded as less important (see Büring's 2007 notion of 'ornamental' accents), while postnuclear accents have been widely disregarded in the literature. In fact, many frameworks simply do not allow for prosodic prominences after the nucleus. An exception is the 'phrase accent', in the sense of Grice, Ladd and Arvaniti (2000) that has been proposed for a number of languages, including English and German. In these cases, a tone is not only associated with the edge of a constituent but may also be associated with a lexically stressed syllable, constituting a secondary prominence in postnuclear position. This postnuclear prominence is marked by increased duration and intensity but not necessarily by tonal movement (especially in the case of L phrase accents). In the present study, we deal with phrase accents constituting a secondary prominence, such as in second occurrence focus contexts. In these, a textually given element is marked

morpho-syntactically by a focus particle (such as *only* or *even*) (see Baumann, Mücke, & Becker, 2010).

Ayers (1996) showed for American English that nuclear accents are perceived as more prominent than prenuclear accents and postnuclear elements. The same study showed the relevance of the type of (nuclear) accents: non-downstepped accents were perceived as more prominent than downstepped accents by American English listeners. This result was empirically validated by Baumann and Röhr (2015) for German, testing seven different nuclear accent types. The accent types were found to vary in their degree of perceived prominence, which was attributed to differences along three tonal dimensions: direction of pitch movement (rises are more prominent than falls), degree of pitch excursion (steep rises and falls are more prominent than shallow rises and falls) and height of the starred tone (high accents are more prominent than downstepped and low accents) (as to the relation between accent shape and prominence see also Knight, 2008, for English and Niebuhr, 2009, for German). Which pitch accent is used depends in part on the particular focus context. In German, contrastive focus and narrow focus accents usually display a rising onglide to the accented syllable (L+H* in GToBI), which is also perceived as most prominent, while nuclear accents in broad focus contexts often show a falling onglide (e.g., H+!H*) (Mücke & Grice, 2014).

So far, we have discussed the role of continuous-valued prosodic parameters and contrastive prosodic categories in the perception of prominence. The final set of parameters that needs to be considered are non-prosodic factors, in particular structural and expectation-based factors which relate to the lexical, semantic and syntactic dimensions of language. These factors also affect prominence, both in production and perception

(Wagner, 2005; Arnold & Wagner, 2008; Cole, Mo, & Baek, 2010, henceforth Cole et al. 2010a; Cole, Mo, & Hasegawa-Johnson, 2010, henceforth Cole et al. 2010b). More generally, the wider discourse context has an influence on "the location, degree and tonal melody of prosodic prominence at the level of the word, phrase, utterance and discourse unit" (Cole 2015: 20).

Many of the non-prosodic factors generate expectations in the listener. For example, listeners *expect* words following a German focus particle such as *sogar* 'even', *nur* 'only' or *auch* 'also' to be more important, which also leads to an increase in their perceived prominence. Bishop (2012) provides an illustrative example of context-induced expectations in American English, showing that the word *motorcycle* in a sentence such as *I bought a motorcycle* received a higher average prominence rating when preceded by the question *What did you buy?* (object focus) as opposed to *What happened?* (sentence focus), even though the target stimulus was lexically, syntactically and acoustically identical in all conditions. Thus, the judgments depended to a considerable extent on the prediction of the contextually appropriate information structure of the target sentence — irrespective of the utterance's actual prosodic form.

Several approaches have looked at prominence with respect to the degree to which a linguistic unit is predictable. A word may be predictable for paradigmatic or syntagmatic reasons. Highly frequent or repeated words, as well as words that are likely to occur in combination with their neighboring words, are often acoustically weak and/or phonologically reduced (e.g., Jurafsky, Bell, Gregory, & Raymond, 2001; Aylett & Turk, 2004; Lam & Watson, 2010). Aylett and Turk (2004) integrate this interplay into their 'Smooth Signal Redundancy Hypothesis', according to which efficient information

transfer is achieved via maintaining an inverse relationship between 'language redundancy' and 'acoustic redundancy'. The former corresponds to semantic-syntactic and lexical variables expressing the degree of predictability or (un-)importance of a word (the inverse of 'surprisal' in Information Theory; Shannon, 1948), while the latter corresponds to gradient prosodic variables serving to highlight a word phonetically (cf. Turk, 2010: 228f.). In recent studies on the automatic detection of prominence, Kakouros and Räsänen (2016) showed that syntagmatically unpredictable words can be found simply via the location of low-probability prosodic events (most predictive cues: duration, energy, F0). However, the syntagmatic predictability of words cannot be based on acoustic prosodic features alone but also depends on higher-level constraints such as the rhythmic structure of utterances in a given language (cf. Arvaniti, 2009).

2.2. Prominence rating studies

Many previous studies that investigate perceptual prominence rely on judgments from a small group of human annotators. For example, Arnold, Wagner and Baayen (2013) used seven different linguistic variables to predict prominence ratings collected from three annotators (Bonner Prosodische Datenbank; Heuft, 1996). They used a data mining algorithm, i.e., random forests (Breiman, 2001), to assess which linguistic variables were most predictive of prominence ratings, finding that the amplitude of the portion of the signal that contained the main pitch excursion was the strongest predictor. However, in their study, "prominence" is based solely on the data from three listeners that annotated the Bonner Prosodische Sprachbank. In a similar study, Kochanski et al. (2005) use the

Intonational Variation in English corpus (Grabe, Post, & Nolan, 2001), where prominence marks have been placed by two phoneticians.

Using prominence judgments by a small group of annotators limits the sample size of listeners, which prevents researchers investigating listener variation in a systematic fashion (see discussion in Cole et al., 2010a, and Cole & Shattuck-Hufnagel, 2016). Moreover, trained participants, such as phoneticians, may behave differently from untrained participants in perceptual tasks (for empirical evidence, see Lancia & Winter, 2012). The judgments of expert annotators in particular may furthermore be biased by their theoretical views. This invites potentially circular generalizations made in prosody research, since they are based on annotations by researchers who were aware of the intonational categories to be labelled. Finally, the annotators in the studies mentioned so far had a lot of time at their disposal and were able to re-listen to particular sentences they annotated for prominence. This luxury is not available to listeners in more realistic communication scenarios, which require rapid detection of perceptual prominence. Because of all of these reasons, we think it is important to also conduct research on naïve listeners' prominence judgments, and, moreover, that it is insightful to compare what experts judge to be prominent to what naïve listeners judge to be prominent. One aim of the present study is thus to verify – if at all – that the expert annotations meaningfully correspond to the behavior of participants with less theoretical knowledge and phonetic listening skills.

Previous prominence judgment studies using a larger number of participants – including untrained listeners – are already available. Shport (2015) recruited 20 native listeners of Japanese and 20 native listeners of English for a categorization task in which

participants had to decide which of the first two syllables in a Japanese nonsense word is more prominent. The location of the F0 peak and the slope of the F0 fall after the peak were acoustically manipulated. Results showed that Japanese listeners used the F0 slope cue to a larger extent; English listeners mainly used peak alignment for their judgments. Importantly, there were also within-group individual differences, especially among the native Japanese listeners, who varied in their perceptual strategies for the pitch accent contrasts tested, with some listeners basing their judgments more on the slope of the pitch fall, and others less. The relatively large number of listeners was essential for the detection of these systematic patterns of listener variation.

In other studies, untrained listeners were asked to rate the prominence of *all* syllables or words in a given sample of utterances. Using a corpus of spoken Dutch, Streefkerk, Pols and ten Bosch (1999) tested how well acoustic-phonetic cues predicted perceived syllable prominence, as measured by the proportion of participants labelling a syllable as stressed. The study confirmed the relevance of some of the above-mentioned variables, such as F0 height and range, duration, relative loudness of the vowel as well as spectral slope. In Eriksson, Thunberg and Traunmüller (2001), listeners used sliders to indicate the degree of prominence for each syllable of 20 versions of the same utterance (see Arnold, Wagner and Möbius, 2011 for an overview of different scales for prominence judgments). The stimuli were produced at different levels of vocal effort, which resulted from varying the distance between speaker and addressee (see Eriksson & Traunmüller, 2000). In a set of linear regression analyses, the mean prominence ratings for each syllable were correlated with signal-based variables, which were categorized as belonging to the domains of "vocal effort", "pitch" and "duration". Together, these three factors described

48% of the variance in mean prominence ratings. Further regression analyses showed that incorporating other factors, such as words being used contrastively or being accented, allowed to describe 57% of the variance in prominence ratings. This provides quantitative evidence for the notion that prominence perception depends on a multitude of different variables.

Cole et al. (2010a, 2010b, see also Cole & Shattuck-Hufnagel, 2016) developed a *Rapid Prosody Transcription* (RPT) task for collecting coarse-grained prosodic judgments (quick binary decisions on prominences and phrase boundaries for each word) from untrained listeners. The method has the advantage that a large set of untrained listeners can be asked to perform a task approximating "in the wild" prominence judgments, much more so than deliberate expert judgments. Using this task, Cole et al. (2010b) have shown that prominence perception is both signal-driven (the longer and louder a word, the more prominent it is) and expectation-driven (the less predictable a word, the more prominent it is). Unpredictability was operationalized in terms of word frequency and discourse givenness (i.e., whether a word was repeated or not). Results revealed that word frequency affects the perception of prominence (less frequent words are more prominent), independent of the fact that less frequent words are generally also longer in duration (e.g., Jurafsky et al., 2001). This suggests that there are at least two different levels, or types, of prominence present in this study, one is based on the features of the speech signal and one is based on a listener's expectations derived from her linguistic knowledge.

2.3. The present study

Up to now, rapid prosody transcription studies only made use of a small set of cues, which allowed researchers to demonstrate the usefulness of the method. To allow stronger generalizations about the factors that influence prominence, a larger set of cues needs to be investigated, and the cues need to be related to each other to look for interactions as well as to investigate which cues are the strongest. This is exactly what the present study sets out to do.

We build on previous RPT tasks (especially Cole et al., 2010b) and extend them in several ways. First, we use a much larger set of linguistic variables as predictors of prominence. Besides testing signal-based factors such as average F0, duration and spectral slope, we also test phonological factors such as pitch accent type in the GToBI system. This allows us to see to what extent slow deliberate judgments by experts in terms of particular intonational phonological categories correspond to rapid prominence judgments by untrained listeners without any knowledge of intonational phonology. We additionally incorporate semantic-syntactic and lexical factors (part-of-speech, presence/absence of syntactic cues for prominence, word frequency). This is the first time such a diverse range of variables has been investigated for the same set of prominence judgments. Second, following the methodological lead of Arnold et al. (2013), we use random forests to disentangle the relative contribution of these linguistic variables in predicting prominence judgments, asking the question: Which factors contribute most to perceived prominence? Third, we use the random effects of linear mixed effects models to explore listener differences (cf. Drager & Hay, 2012). Fourth and finally, we extend the RPT task to a new language, namely German.

We are particularly interested in comparing categories from the GToBI system to naïve listener judgments. The GToBI categories (comprising our "discrete prosodic" variables) were chosen in part because GToBI is an established system of annotating German intonation (e.g., Grice et al., 2005; Grice & Baumann, 2016; Grice, Baumann, & Jagdfeld, 2009; Ritter & Grice, 2015; Baumann & Röhr, 2015), and it is important to show that this system corresponds meaningfully to the behavior of listeners who are not trained GToBI annotators. In fact, the relevance of using these phonological categories goes beyond merely verifying an existing annotation system: As GToBI categories are generalizations over pitch contours, they correspond to particular continuous shapes which signal phonological contrasts (e.g., to express differences in information structure). Of course, pitch accents are ultimately composed of gradient phonetic parameters, but their classification within a phonological system, such as GToBI, can be used to approximate these gradient phonetic parameters. We compare (a) the discrete phonological-prosodic factors (pitch accent type and pitch accent position, based on GToBI) both to (b) continuous-valued phonetic-prosodic parameters (such as pitch, intensity and duration of each word) and a couple of (c) non-prosodic factors (semantic-syntactic and lexical variables). The set of our (non-GToBI) variables was chosen based on prior research on prominence in other domains, such as syllable-level prominence within the research on lexical stress. Thus, for these variables, we investigate the extent to which syllable-level cues to prominence generalize to the level of the word.

3. Methods

3.1. Participants

Twenty-eight prosodically untrained native speakers of German (18 women, 10 men), mostly from Hesse and the Rhineland, participated as listeners in the experiment. They were aged between 18 and 58 years (with a mean of 24.8 years) and did not report to have any auditory impairments.

3.2. Stimuli

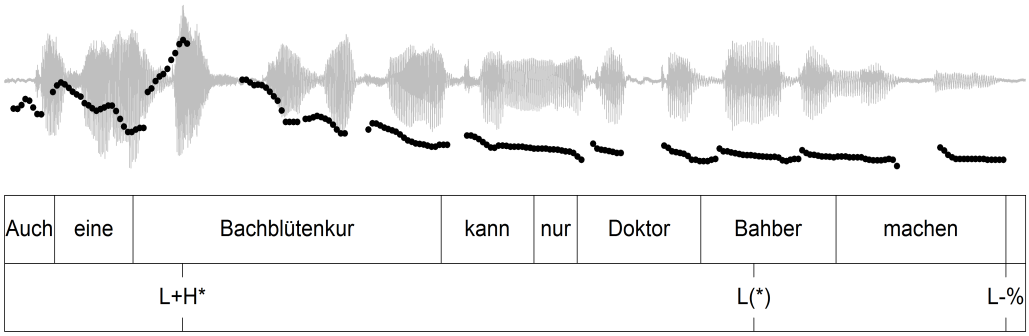
Sixty German sentences (between 5 and 18 words, 562 words in total; see Appendix A) were read by 14 different native speakers of German (11 female and 3 male, aged between 22 and 38 years). These sentences were selected from various small databases of read German that served as the basis for other published work (Röhr & Baumann, 2010; Baumann et al., 2010; Turco, Dimroth, & Braun, 2013; Mücke & Grice, 2014). Textually identical sentences were never produced with the same intonation. An important criterion for selecting sentences from existing research was to include as many different accent types in different accent positions as possible, based on a GToBI analysis of the sentences (see Table 1 and section 3.4.2.) (consensus annotation by three annotation experts).

In the original production studies, the sentences were uttered in various contexts (broad focus, narrow focus, verum focus, second occurrence focus etc.). These contexts were absent in the task we presented. This is why a large portion of the sentences in our stimulus set (38%) did not display a "default" (broad focus) intonation, i.e., a realization we would expect in an out-of-the-blue utterance. In a naturally occurring broad focus utterance, prosodic and non-prosodic factors that affect prominence perception are often co-varying with each other: e.g., the nuclear accent (prosodic) usually falls on the last argument (non-prosodic), whose head generally is a noun (non-prosodic), which is often

not only an infrequent word (non-prosodic) but also phonetically the longest and loudest word in the utterance (prosodic). However, to disentangle the influence of prosodic and non-prosodic factors on prominence perception, we need a diverse set of stimuli that includes a wide variety of non-canonical prosodic realizations. Our focus on isolated sentences also means that our study is not designed to also assess the role of expectation-based factors that stem from considerations of the discourse context (as e.g. in Bishop, 2012). Nevertheless, we do look at expectations generated from a more local context (within sentence, such as focus operators). We acknowledge that our choice of stimuli may affect the results. Future research needs to investigate different text types, including spontaneous utterances, as well as utterances in context.

Figure 1 shows two example utterances. The first utterance is characterized by a rising nuclear accent early in the intonation phrase (on the noun *Bachblütenkur* 'a cure with Bach flowers') and a low phrase accent (see Grice et al., 2000) on the proper noun *Bahber*, which in this example represents a 'second occurrence focus' (i.e., contextually given but focused information). The second example utterance contains a nuclear accent of a (smaller) intermediate phrase (on the noun *Bekannten* 'friend'), followed by a nuclear accent of a (larger) intonation phrase (on the noun *Empfehlung* 'recommendation'), with a prenuclear accent on the preceding adjective *gute* 'good' (the terms 'intermediate phrase' (ip) and 'intonational phrase' (IP) are explained in more detail in section 3.4.2. below). The first two accents in the phrase are rising (L+H* in GToBI), the final accent is low (L*).

(a)



(b)

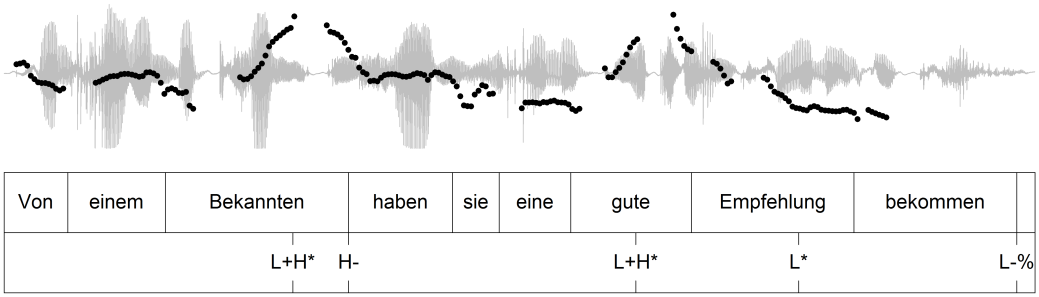


Figure 1: Waveforms with superimposed F0 contours for the utterances (a) *Auch eine Bachblütenkur kann nur Dr. Bahber machen* ('Also a cure with Bach flowers can only be done by Dr. Bahber') and (b) *Von einem Bekannten haben sie eine gute Empfehlung bekommen* ('From a friend they got a good recommendation') with annotated accents using the GToBI system.

3.3. Experimental procedure

We used the Rapid Prosody Transcription task following Cole et al. (2010a, 2010b) and Cole and Shattuck-Hufnagel (2016) (see section 2.2). Participants were asked to underline

the words they deemed to be prominent on a printed transcript while or immediately after listening to a speech sample. The instructions were:

> "Ihre Aufgabe besteht nun darin, sämtliche Wörter, die Sie in einer Äußerung als betont / hervorgehoben / wichtig wahrnehmen, auf dem Transkript zu unterstreichen."
>
> '*Your task is now to underline all the words on the transcript which you perceive as stressed / highlighted / important*.'

We deliberately chose a selection of potentially equivalent terms for the notion of prominence. The range of terms given is compatible with different notions of prominence, including signal-based, structure-based and meaning-based prominence. This means that different listeners may interpret the task differently, a point to which we return below. As in other RPT studies, capitalization and punctuation marks were removed from all written stimulus sentences in order to avoid orthographic influences on listeners' judgments (see Appendix A).

Data were collected at the University of Cologne and at the Goethe University Frankfurt, with listeners being seated in a silent room. For the presentation of the sound stimuli, we used PowerPoint slides on a MacBook Pro. Listeners had the option of hearing a particular stimulus twice, but they did not have the option of playing specific portions of the sample while doing their transcriptions. In the course of the experiment, the 60 stimuli were played over headphones and were divided into three blocks of 20 utterances each,

with an optional short break between each block. The stimulus order was constant within each block. The order of the blocks was pseudo-randomized.

3.4. Overview of linguistic variables

As discussed above, we can divide the linguistic variables studied into three groups: (1) continuous-valued prosodic, (2) contrastive (discrete) prosodic and (3) non-prosodic variables. Each variable is described in detail below. All acoustic variables were measured with Praat (Boersma & Weenink, 2013).

3.4.1. Continuous-valued prosodic variables

a. Pitch/F0:

Since at least the early empirical studies of Fry (1958), pitch has been considered an important phonetic correlate of perceived prominence (at word level) in West Germanic languages (but see Kochanski et al., 2005). For the present dataset, we measured the MEAN F0 and the MAXIMUM F0 of each word (with MAXIMUM F0 corresponding to the phonological H target in the case of accented words). In order to reduce errors due to microprosody, pitch halving and doubling, and irregular phonation, all extracted F0 values were checked and manually corrected whenever necessary.

Although dynamic properties, which indicate pitch *movement* – such as pitch slope and range –, rather than static properties, are considered as being particularly important for prominence perception (e.g., Rietveld & Gussenhoven, 1985; see also the notion of 'prominence-lending pitch movement' by 't Hart, Collier and Cohen 1990: 96ff.), we refrained from including these measures in our analyses, for several reasons: First, relevant

pitch movement often spans several words, so that its interpretation may be fundamentally different depending on which context it occurs in. As a consequence, measuring pitch slope and range *on every word*, especially on function words, would not be very informative. Moreover, the slope of a pitch rise or fall is most adequately defined for *accented* syllables or words, but a restriction to accented words would mean to exclude a large part of the dataset. Finally, we already investigate another variable elsewhere that adequately approximates pitch shape characteristics, namely the GToBI accent type categories (see below). Nevertheless, for the reader interested in dynamic properties, it should be noted that PITCH RANGE (F0 excursion in semitones) and PITCH SLOPE (F0 excursion in semitones per second) are in fact associated with prominence judgments in a statistically reliable way (under consideration of the issues just mentioned)[2].

b. Length/Duration:

Another well-established cue to prominence is the acoustic duration of a constituent (Fry, 1955; Sluijter et al., 1997; Kochanski et al., 2005), which has also been explored in the RPT tasks by Cole et al. (2010a, 2010b) for American English. In the present study, we investigate the influence of both VOWEL DURATION and SYLLABLE DURATION of the lexically stressed syllable of each word, expecting that longer vowels and syllables are more likely to be judged as prominent.

c. Loudness/Intensity:

[2] We measured F0 RANGE and F0 SLOPE on accented syllables only and computed the absolute value (disregarding sign) of the slope, which puts rising and falling pitch excursions on the same metric. Both absolute F0 RANGE ($\chi^2(1) = 78.0$, $p < 0.0001$) and F0 SLOPE ($\chi^2(1) = 51.4$, $p < 0.0001$) of accented syllables were reliably associated with prominence judgments.

Overall intensity has been shown to be an acoustic correlate of prominence at utterance level (Kochanski et al., 2005; also in accordance with the literature on lexical stress, e.g., Fry, 1955), with increased intensity leading to increased prominence. As a measure of overall intensity, we calculated RMS AMPLITUDE for each word. Besides RMS (measured uniformly across the frequency spectrum), we considered measures that looked at intensity with respect to specific frequency ranges. Intensity measures that take the distribution of energy across the frequency spectrum into account have been claimed to be more reliable correlates of perceived prominence than overall intensity (see Heldner, 2003, for an overview). We investigate one measure for SPECTRAL EMPHASIS, defined as the difference between the overall intensity and the intensity in a low-pass-filtered signal, thus emphasizing the relative contribution of the higher-frequency part of the spectrum (following Traunmüller, 1997, and Traunmüller & Eriksson, 2000). That is, if a syllable is high in spectral emphasis, it has more energy in the high frequency components; if a syllable is low in spectral emphasis, it has relatively more energy in the low frequency components. We additionally considered two measures of SPECTRAL TILT, representing the slope of the frequency spectrum (difference between first harmonic and amplitude peaks in the vicinity of second and third formant, H1-A2 and H1-A3). A flatter tilt indicates more energy in the high frequency components of the spectrum (around F2 and F3), which several studies have found to be associated with prominence (e.g., Sluijter et al., 1995; El Zarka, Schuppler, Lozo, Eibler, & Wurzwallner, 2015). Like these studies, we controlled our measures for speaker gender and vowel identity, i.e., we compared our values with typical values for male and female speakers and for each vowel separately in order to avoid errors in the calculation of formant bandwidth (cf. Iseli, Shue, & Alwan, 2007).

3.4.2. Contrastive (discrete) prosodic variables

a. ACCENT VS. NO ACCENT:

This binary variable is based on the GToBI consensus annotation and codes for whether a word is pitch-accented (ACCENT) or not (NO ACCENT). We expect accented words to be perceived as more prominent than unaccented words. For the ACCENT category, the type of pitch accent is ignored and phrase accents (occurring in postnuclear position) are excluded, since they are not classified as fully-fledged pitch accents (see section 2.1). For the NO ACCENT category, unaccented words and postnuclear prominences are lumped together. Table 1 shows the overall distribution of all relevant prosodic categories in our stimulus set. A total of 187 words (33% of the whole set) are accented, compared to 375 words (67%) which are unaccented.

Type / Position	No accent	Low L*	Falling H+L*, H+!H*	High H*, !H*	Rising L*+H, L+H*	Total
No accent	356	n/a	n/a	n/a	n/a	356
Postnuclear	19	n/a	n/a	n/a	n/a	19
Prenuclear	n/a	10	6	10	54	80
Nuclear ip	n/a	7	6	6	28	47
Nuclear IP	n/a	3	29	15	13	60
Total	375	20	41	31	95	562

Table 1: The distribution of accent types and accent positions in our stimulus set. GToBI accent types are conflated into groups describing the pitch contour in the vicinity of the accented syllable (low, falling, high or rising).

b. ACCENT POSITION:

Within the GToBI system, fully-fledged pitch accents can occur in prenuclear or nuclear position (see section 2.1). Many autosegmental-metrical intonation systems define nuclear accents as the last accent in an 'intermediate phrase' (ip; following Beckman & Pierrehumbert, 1986). A (larger) 'intonation phrase' (IP) is made up of one or more intermediate phrases. For the purpose of the present study, we will call the nuclear accent of the final ip in an IP 'nuclear accent of an IP', and we will call the nuclear accent of a non-final ip 'nuclear accent of an ip' (see Table 1). We also included postnuclear prominences or phrase accents, which are marked by increased duration and intensity, but which lack a local tonal movement (see section 2.1). In line with the order proposed by the prosodic prominence hierarchy (Shattuck-Hufnagel & Turk, 1996), we expect the following decreasing order of perceived prominence for the different accent positions: nuclear accent of IP > nuclear accent of ip > prenuclear accent > postnuclear prominence (= phrase accent).

c. ACCENT TYPE:

Different types of German pitch accents have been shown to differ with respect to their perceived degree of prominence (Baumann & Röhr, 2015). Three dimensions were found to be important: the direction of pitch movement, the degree of pitch excursion and the height of the starred tone. According to these findings, we expect to confirm the following order of perceived prominence: rising accent > high accent > falling accent > low accent (see their distribution in the present dataset in Table 1).

d. RHYTHM-DEPENDENT PROMINENCE

As discussed above, prominence perception is not only due to word-internal factors but also to contextual factors. Studies on lexical and postlexical stress have found that strong and weak syllables tend to be alternating (in particular in West Germanic languages, Liberman & Prince, 1977). Here, we investigate a similar pattern at the level of the sentence, i.e., whether a word is perceived as more prominent in the context of other, non-prominent, words. A simple binary variable was computed which measures the extent to which the *preceding* word was prominent (PRECEDING PROMINENT) or not (PRECEDING NOT PROMINENT). This is the only variable that considers the listener-internal contribution to prominence perception: whereas all other variables were generated based only on the stimuli themselves (including their immediate textual context), this variable considers listener behavior as well, i.e., whether the listener judged words in the immediate vicinity to be prominent or not. This variable is not included in the exploratory random forest analysis because this analysis is item-based and the PRECEDING PROMINENT variable is listener-specific.

3.4.3. Non-prosodic variables

a. PART-OF-SPEECH:

The information status of a word, i.e., whether an item is given, accessible or new in a discourse, can only be meaningfully attributed to content words, in particular to nominal expressions. Since information status has been shown to be related to prominence marking (with new information being more prominent, e.g., Baumann & Riester, 2012), we expect a similar relation between PART-OF-SPEECH as an expectation-based factor and perceived prominence. In fact, previous corpus annotation studies on German (Widera, Portele, &

Wolters, 1997; Baumann, Eckart, & Riester, 2016) suggest a higher degree of perceived prominence of content words in comparison to function words. These studies furthermore suggest that there are differences in the perceived prominence of different content word categories, with nouns, proper names and adjectives being more likely to be judged as prominent than verbs and adverbs. A recent RPT study on American English (Roy, Cole, & Mahrt, 2017) reports similar results, with a particularly high probability of nouns to be marked as prominent.

All words in our dataset were classified according to the *Stuttgart Tübingen TagSet* (STTS; Schiller, Teufel, Stöckert, & Thielen, 1999). Table 2 provides the numbers of occurrences for each word class.

Content words	**300**	**Function words**	**262**
Nouns	121	Pronouns	74
Adverbs	57	Articles	68
Verbs (full)	52	Prepositions	42
Adjectives	36	Auxiliary verbs	26
Proper names	34	Particles	22
		Modal verbs	18
		Conjunctions	12

Table 2: Token distribution of part-of-speech categories in the dataset.

b. FOCUS PARTICLE:

Nine sentences (15%) contained a total of eleven occurrences of the German focus-sensitive particles *nur* 'only', *sogar* 'even' and *auch* 'also' (e.g., in second occurrence

focus constructions; see section 2.1). We expect words that are in the scope of a focus particle to be judged as more prominent than words that are not (see Büring, 2015). We only coded the head of the complex constituent which is in the scope of the particle as 'focused', e.g., the noun in the phrase *auch den Zivildienst* ('also the civilian service').

c. LAST ARGUMENT:

West Germanic languages show a stable pattern in broad focus structures, namely that the (last) verbal argument receives the nuclear pitch accent rather than a predicate or modifier (see e.g., Gussenhoven, 1984). This association with the final accent in a phrase may trigger the expectation that the last argument of a sentence is perceived as particularly prominent. The analysis will show whether this expectation leads to a higher likelihood of prominence marks irrespective of the argument's prosodic realization. Our dataset contains 62 words that are coded as 'LAST ARGUMENT' (11% of the total set of words).

d. NO. OF SYLLABLES per Word:

Longer words, which often are morphologically complex, are also judged to be more semantically complex (Lewis & Frank, 2016). Furthermore, more frequent words have a strong tendency to be shorter (Zipf, 1949), and word frequency is independently associated with prominence judgments, with more frequent words being judged as less prominent (Cole et al., 2010b). Both the semantic complexity (Lewis & Frank, 2016) and the association with frequency (Zipf, 1949) could generate the expectation in language users that longer words are more prominent, which is what we are testing here. The NO. OF SYLLABLES measure is based on the number of syllables of the written word form (not necessarily the phonetically realized number of syllables). The counts for this variable are shown in Table 3.

One-syllable words	292	Four-syllable words	22
Two-syllable words	187	Six-syllable words	2
Three-syllable words	56	Eight-syllable words	3

Table 3: Token distribution of words with different numbers of syllables in the dataset.

e. LOG WORD FREQUENCY:

For each word we determined its frequency by consulting the German version of the SUBTLEX corpus (Brysbaert, Buchmeier, Conrad, Bölte, & Böhl, 2011), which contains more than 25 million words taken from movie subtitles. Word frequencies from movie subtitles have been argued to closely emulate spoken language, and studies on English have shown that the SUBTLEX corpus frequencies in particular are most predictive of cognitive measures such as reaction times (Brysbaert & New, 2009). We expect a negative correlation between word frequency and perceived prominence, i.e., the less frequent a word is, the more perceptually prominent it is. This is in part because infrequent words tend to be more hyperarticulated and longer in duration. However, as Cole et al. (2010b) have shown, there also is an independent effect of word frequency on prominence. In general, infrequent words are informationally more surprising and contribute more new information to a message (Shannon, 1948).

3.5. Data analysis overview

We are dealing with a complex dataset (many linguistic variables that are related to each other, potential listener differences etc.) that has many patterns worthy of investigation. It is important to separate confirmatory analyses (testing established hypotheses) from

exploratory analyses (finding novel patterns in the data). To cope with this complexity, we provide an overarching structure to our analysis that is separated into three stages: First, we briefly test each factor's influence on prominence perception in isolation, using mixed logistic regression. This analysis is confirmatory in that we test predictions coming from previous work (e.g., louder words should be perceived as more prominent) with a new dataset (untrained German listeners) and a new task (the Rapid Prosody Transcription task). To the extent that we demonstrate patterns that are already widely believed to be real, our analyses represent a replication. Moreover, these analyses serve as a 'sanity check' to assess whether each variable does indeed behave the way we expect. In a second analysis stage, we look at relations between the different variables in an exploratory fashion. Here we use a data mining algorithm to assess the relative weighting of the different cues for prominence. In a third and final stage, we investigate whether there are systematic differences between listeners. This analysis, too, is exploratory.

In the first (confirmatory) analysis, we used logistic mixed effects regression to test how specific linguistic variables (such as ACCENT POSITION, MEAN F0 etc.) affect the likelihood of prominence marks. We used separate models for each variable rather than entering all variables simultaneously into the same model. This was done for three reasons: First, many of the variables are correlated with each other (e.g., WORD FREQUENCY and PART-OF-SPEECH), which means that collinearity is a potential concern (see Zuur, Ieno, & Elphick, 2010). Second, we want to be conservative in estimating the impact of each variable, which warrants estimating by-listener varying random slopes for each variable in question (Barr, Levy, Scheepers, & Tily, 2013). This is especially the case because past research suggests that listeners can be expected to vary in how much their prominence

judgments are affected by a particular variable (Cole et al., 2010a, 2010b). Estimating a mixed model with random slopes for a total set of 17 different linguistic continuous and categorical variables is not feasible, which is why setting up separate models for each variable is the preferred option. Third and finally, our exploratory analysis (stage two) considers all variables together in one conjoined analysis and also allows us to investigate the influence of a particular variable in the face of an interaction. For all of these reasons, we build one logistic regression model for each linguistic variable in question. Each model thus independently tests the contribution of a linguistic variable without considering any of the other variables.

The second analysis then disentangles the relative contribution of specific linguistic variables, for which we used random forests (Breiman, 2001). This data mining algorithm takes a set of predictors (in this case, 16 different linguistic variables we investigate – the syntagmatically determined variable RHYTHM-DEPENDENT PROMINENCE was excluded; see section 4.2.) and outcomes (in this case, our prominence ratings). We chose random forests instead of other analysis approaches (such as a logistic regression model with 16 different predictors) for several reasons: The first reason is the above-mentioned collinearity, for which random forests have been argued to be particularly suitable (Strobl, Malley, & Tutz, 2009). A second reason for using random forests is that we have relatively many predictors for relatively few data points, and random forests have been argued to be particularly good for such "low *N* high *p*" data analyses (see Strobl et al., 2009). Finally, random forests can evaluate variable importance while also looking at possible interactions, e.g., a continuous acoustic variable such as duration or pitch may matter more with respect to prominence judgments for one particular pitch accent type as opposed to another pitch accent type.

Random forests have already been used in linguistic applications (e.g., Tagliamonte & Baayen, 2012; Brown, Winter, Idemaru, & Grawunder, 2014; Grice, Savino, Caffo, & Roettger, 2015; Al-Tamimi, 2017; Roettger, 2017), for instance for the prediction of prominence judgments (Arnold et al., 2013).

Our third analysis stage uses the estimated random effects coefficients from the first analysis (confirmatory mixed models) to look at individual differences among our listeners (see Drager & Hay, 2012 for a similar analysis using mixed model random effects). We focused on analyzing whether particular prominence-lending variables were correlated across individuals. As an example, listeners who may base their prominence judgments more on prosodic variables may be less influenced by word frequency. In this analysis, we first tested for specific correlations in a confirmatory fashion (controlling for multiple comparisons). We then performed an exploratory cluster analysis on the prominence judgments of our participants to investigate the presence of any latent listener groups. In other words: Are there specific groups of people that respond in a similar fashion, and if so, what cues do they focus on?

All analyses were conducted with R (R Core Team, 2015) and the packages are listed in Appendix B. The specifics of each analysis will be explained in the respective results section. For more detail and to abide by standards of reproducible research, all analysis scripts and data are made permanently available under the following publically accessible link:

https://github.com/bodowinter/rapid_prosody_transcription_analysis/

4. Results

4.1. Confirmatory mixed model analysis

We model the binary dependent measure "prominence" ("prominent" versus "not prominent") as a function of a particular fixed effect (such as MEAN F0) using a series of mixed logistic regression analyses. Each model includes three types of random intercepts, quantifying variation that is due to listeners, sentences, or speaking voices (see Baayen, Davidson, & Bates, 2008; for a similar analysis see also Brown et al., 2014). Each one of these factors can be perceived as a source of idiosyncratic variation, while furthermore introducing a level of interdependence (multiple responses by the same listener, to the same sentence, to the same voice) that needs to be accounted for statistically. In addition, each model always included by-listener random slopes for the single linguistic variable that was tested (compare Barr, Levy, Scheepers, & Tily, 2013). These listener random slopes are theoretically motivated because listeners can be expected to differ in how particular variables influence their prominence judgments, and because past research on prominence perception has already demonstrated individual differences with respect to particular prominence cues (Cole et al., 2010a, 2010b). As is common in regression models, each continuous variable was *z*-scored to aid interpretation (Schielzeth, 2010), e.g., for SYLLABLE DURATION, the mean duration across all data points was subtracted and the variable was divided by the standard deviation across all data points. This makes the strength of the prominence effect comparable across different variables that have different

metrics, such as between SYLLABLE DURATION and MAXIMUM F0. In other words, we only report standardized slopes which can be compared across models.[3]

We first focus on the continuous variables. Both MAXIMUM F0 ($\chi^2(1) = 112.6$, $p < 0.0001$) and MEAN F0 ($\chi^2(1) = 100.7$, $p < 0.0001$) influenced prominence judgments in a statistically reliable fashion. Increasing the MAXIMUM F0 by one standard deviation (*SD* = 68 Hz) increased the odds of observing a prominent response by 2.77 to 1 (logit estimate: 1.02, *SE* = 0.03). Increasing the average F0 by one standard deviation (*SD* = 51 Hz) increased the odds of observing a prominent response by 2.19 to 1 (logit estimate: 0.79, *SE* = 0.03). Since both variables are *z*-scored, the difference in logit estimates can be interpreted as indicating the strength of the effect. Thus, MAXIMUM F0 had a comparatively larger influence on prominence judgments than MEAN F0. Figure 2a shows the predicted percentage of prominence judgments (model fit) as a function of MAXIMUM F0, showing a clear positive association.

Both SYLLABLE DURATION ($\chi^2(1) = 90.0$, $p < 0.0001$) and VOWEL DURATION ($\chi^2(1) = 84.8$, $p < 0.0001$) influenced prominence judgments in a statistically reliable fashion, with increased duration leading to more prominence judgments in both cases. For

[3] In some cases, there were problems with model convergence, i.e., the estimation of parameters was difficult. These problems were prevented either by switching to another numerical estimation procedure or by simplifying the random effects structure (see online R scripts for details), although we never dropped the by-listener varying random slopes. In the case of MAXIMUM F0 and MEAN F0, convergence was facilitated by additionally *z*-scoring within gender. All *p*-values stem from likelihood ratio tests of the model with the fixed effect in question against the model without the fixed effect in question (see Winter, 2013; Barr et al., 2013). For the likelihood ratio tests of fixed effects, we fitted all models with restricted maximum likelihood. Inflation of the family-wise error rate is a concern because we performed an analysis with 17 separate models that test for the same underlying null hypothesis (i.e., a given variable has no influence on prominence marks). To circumvent this, we Dunn-Šidák corrected all *p*-values for performing 17 tests. Because correcting for multiple comparisons yielded the same substantive conclusions, we decided to report uncorrected *p*-values for simplicity's sake.

SYLLABLE DURATION, an increase in one standard deviation (*SD* = 77 Hz) changed the odds of observing a prominent response by 2.05 to 1 (logit: 0.72, *SE* = 0.03), as shown in Figure 2b. Similar results were obtained for VOWEL DURATION, with one standard deviation increase (*SD* = 37 Hz) leading to a change in odds of 1.9 to 1 (logit: 0.64, *SE* = 0.03). RMS AMPLITUDE also influenced prominence judgments in a statistically reliable fashion ($\chi^2(1) = 126.6$, $p < 0.0001$). For each increase in one standard deviation (*SD* = 4.92 dB) the odds rose by 4.5 to 1 (logit: 1.5, *SE* = 0.4), see Figure 2c. Comparison of the standardized slopes shows that RMS AMPLITUDE has a stronger influence on perceived prominence (logit: 1.5) than SYLLABLE DURATION (logit: 0.72), which in turn had a stronger influence than VOWEL DURATION (logit: 0.64).[4]

Figure 2: Probability of prominence marks as a function of four continuous-valued prosodic variables, (a) MAXIMUM F0, (b) SYLLABLE DURATION, (c) RMS AMPLITUDE and (d) SPECTRAL EMPHASIS. Lines show mixed model predictions from the models reported in the body of the paper, with shaded regions representing 95% confidence bands around those predictions (incorporating random effects). Data points represent the actual prominence marks ("prominent versus not prominent"), with random scatter added for increased visibility.

[4] Including utterance-normalized measures of MEAN and MAX F0, RMS AMPLITUDE and DURATION – instead of using the raw measures – did not change the statistics. The separate logistic regression models for these variables revealed that they were equally statistically reliable. Furthermore, using the normalized values did not change the order of the variables' influence on prominence perception.

Both measures of SPECTRAL TILT were associated with prominence judgments in a statistically reliable fashion, which was the case for both H1-A2 ($\chi^2(1) = 62.5$, $p < 0.0001$; logit: -0.33, $SE = 0.03$) and H1-A3 ($\chi^2(1) = 82.5$, $p < 0.0001$; logit: -0.53, $SE = 0.03$). SPECTRAL EMPHASIS also had a statistically reliable effect on prominence marks ($\chi^2(1) = 76.6$, $p < 0.0001$). With each increase by one standard deviation ($SD = 6.17$), the odds of observing a prominent response increased by 1.5 to 1 (logit: 0.41, $SE = 0.02$). The spectral emphasis measure is shown in Figure 2d. Comparison of standardized slopes reveals that H1-A3 (logit: -0.53) had the strongest influence on perceived prominence, compared to SPECTRAL EMPHASIS (logit: 0.41) and H1-A2 (logit: -0.33).

Another continuous variable, albeit a non-prosodic one, that influenced prominence judgments in a statistically reliable fashion was LOG WORD FREQUENCY ($\chi^2(1) = 65.8$, $p < 0.0001$). For each decrease in LOG WORD FREQUENCY by one standard deviation ($SD = 1.43$), the odds of observing a prominent response increased by 2.3 to 1 (logit: 0.82, $SE = 0.05$), as shown in Figure 3. Another non-prosodic factor that is bound to the word is word length, as measured by the NUMBER OF SYLLABLES. This also reliably influenced prominence marks ($\chi^2(1) = 46.0$, $p < 0.0001$; logit: 0.44, $SE = 0.04$) (in the mixed model, the NUMBER OF SYLLABLES variable was treated as a continuous variable, *z*-scored like the other variables). On average, 9% of the one-syllable words were judged to be prominent, compared to 22% of the two-syllable words, 35% of the three-syllable words, and 32% of the four-syllable words. There were only two instances of a six-syllable word (*Klassenkameradin* 'class mate'), and three instances of an eight-syllable word (*Untersuchungsergebnisse* 'results of an examination'), which were rated to be prominent on average in 13% of the cases. The fact that these very long words appear to be judged as

less prominent than some of the three- and four-syllable words will be picked up in the discussion.

Figure 3: Probability of prominence marks as a function of the (non-prosodic) gradient variable LOG WORD FREQUENCY with superimposed mixed model predictions and 95% confidence bands.

Next we turn to the analysis of the categorical variables, starting with differences in PART-OF-SPEECH. In descriptive terms, proper names were most likely to receive prominence judgments (42%), followed by adjectives (41%), nouns (29%), adverbs (23%) and verbs (17%), as shown in Figure 4 (and perfectly in line with the results of Widera et al., 1997). Particles (*aus, weg, durch* etc.) were rated to be prominent only 12% of the time, followed by modal verbs (*muss, kann, solle* etc.), which were rated to be prominent only 7% of the time. Even lower in the percentage of prominence ratings were conjunctions (5%), pronouns (3%), articles (2%), auxiliary verbs (1%) and prepositions (0%). For the mixed model analysis, we analyzed PART-OF-SPEECH in a binary fashion ("content words" versus "function words"). This was reliably associated with prominence marks ($\chi^2(1) = 81.3, p < 0.0001$), with content words having a predicted higher percentage of prominence marks (~30% on average) than function words (~3% on average) (logit: 2.6, $SE = 0.14$).

Figure 4: Descriptive percentages for prominence marks broken up by lexical category; content words are indicated by grey bars, function words by white bars.

Figure 5 shows mixed model predictions and confidence intervals for all other binary categorical variables. Whether a word was the LAST ARGUMENT in a sentence was associated with prominence judgments in a statistically reliable fashion ($\chi^2(1) = 9.7$, $p = 0.002$). Last arguments were judged to be prominent 23% of the time, all other words were judged to be prominent on average 16% of the time (logit: 0.35, $SE = 0.1$). Morpho-syntactic focus marking also had a statistically reliable influence on prominence judgments ($\chi^2(1) = 13.0$, $p = 0.0003$). Words that followed a FOCUS PARTICLE were judged to be prominent 46% of the time, words that did not only 16% of the time (logit: 3.95, $SE = 1.09$). Finally, the presence or absence of a pitch accent (coded as a binary categorical variable ACCENT VS. NO ACCENT) influenced prominence judgments reliably ($\chi^2(1) = 70.9$, $p < 0.0001$), with pitch-accented words being more likely to be judged as prominent (46%) than non-accented words (2%) (logit: 4.4, $SE = 0.2$).

Figure 5: Predicted probability of prominence marks (from mixed logistic regression models) for all binary categorical variables with 95% confidence intervals.

We furthermore looked at differences in the perceived prominence of different pitch accent types and pitch accent positions. ACCENT POSITION was reliably associated with prominence judgments ($\chi^2(3) = 61.6$, $p < 0.0001$), and so was ACCENT TYPE ($\chi^2(3) = 46.4$, $p < 0.0001$). Figure 6 shows the descriptive averages of the percentage of prominence marks for the different positions and types.

Figure 6: Descriptive percentages of average prominence marks for different ACCENT POSITION and ACCENT TYPE.

Finally, what about the variable RHYTHM-DEPENDENT PROMINENCE, investigating whether a given prominence mark depends on the prominence mark of the previous word? This variable is different from all the others in that it is not entirely dependent on the stimulus itself, but also on the listeners's own prominence judgments. There was indeed a reliable effect of RHYTHM-DEPENDENT PROMINENCE ($\chi^2(1) = 28.99$, $p < 0.0001$). If the preceding word was not marked as prominent, then the percentage of words marked as prominent was 18%. If the preceding word was marked as prominent, this percentage dropped to 9%. Thus, there was a strong preference for prominence marks to *not* follow other words that were marked as prominent. This result supports the claim that West Germanic languages show a tendency for an alternating speech rhythm, at least as a perceptual phenomenon.

4.2. Random forest analysis of prominence cues

In our second analysis, we explored which prosodic or non-prosodic variables are most predictive of our listeners' prominence judgments. For this, we took the "p-score" of each word, which is the proportion of participants who underscored the respective word,[5] and

[5] Note that p-scores represent majority decisions on a binary feature, which are not equivalent to average auditory impressions. Nevertheless, the p-scores do provide an independent (and theory-

conducted a random forests analysis on these. The analysis is items-based, with each word contributing one data point (a total of 562 data points).

We followed the guidelines presented in Strobl, Malley and Tutz (2009) and fitted a random forest with the `ranger` package version 0.8.0 (Wright & Ziegler, 2017) with 2,000 trees and four random variables per tree (the rounded square root of the number of predictors). Variable importance was computed via permutation tests (`permutation = TRUE`), which has been argued to account better for collinearity (Strobl et al., 2009).[6]

The random forest was trained on a random subset of 70% of the data (training set) and its predictions were tested on the remaining 30% (test set). There was a very high correlation between the p-scores predicted by the random forest algorithm and the actual p-scores for the test set ($r = 0.84$, $R^2 = 0.71$, for the central imputed data). This already is an interesting result as it shows that prominence judgments can indeed be predicted very well by looking at the 16 variables we considered for this analysis.[7] Just taking these 16 variables, we are able to describe about 70% of the variation in prominence judgments in a new dataset. Figure 7 shows the "variable importances". These variable importances take interactions and collinearity into account and can only be interpreted relative to each other.

---

unbiased) measure of the perceived similarity (or difference) among words with respect to prominence (see Cole & Shattuck-Hufnagel, 2016:11).

[6] This analysis is only possible if there are no missing values. For some data points, F0 or spectral emphasis could not be computed. In these cases, we ran the random forest either with a reduced dataset (data points with missing values excluded), or two alternative ways of imputing the missing data (K-nearest neighbor imputation or central imputation). The resulting conclusions were the same, and the random forests trained on these different datasets performed similarly with respect to predictive accuracy.

[7] Note that from our set of 17 variables, RHYTHM-DEPENDENT PROMINENCE was excluded here (as well as in the exploratory analysis of individual differences in section 4.3), since it is different in nature: this variable does not rely on the stimulus alone (i.e., is not items-based), but also on each listener's *judgment* for each word in relation to the listener's judgment for each previous word (see section 3.4.2.).

Figure 7: Relative variable importance based on a random forest analysis.

As can be seen, the discrete prosodic variables ACCENT VS. NO ACCENT, ACCENT POSITION, and ACCENT TYPE were by far the most important variables in predicting prominence judgments. One possible reason for this clear result may be that the choice of different (and often quite 'well-pronounced') accent types in different accent positions was a central selection criterion for the stimuli (see 3.2.). Another possible reason is that the phonetic parameters reflected in the discrete prosodic variables may be most important for prominence perception, i.e., the actual shape of the pitch curve. Regardless of these concerns, the random forest variable importances confirm the relevance – and adequacy – of the GToBI categories for prominence perception (cf. Baumann & Röhr, 2015).

Compared to the discrete prosodic variables, semantic-syntactic and lexical factors played a minor role, with WORD FREQUENCY and PART-OF-SPEECH being the most relevant non-prosodic factors. The continuous-valued acoustic variables played a similarly minor role. Among them, however, RMS AMPLITUDE, MEAN F0, MAXIMUM F0 and SPECTRAL TILT (H1-A3) were the most important predictors of prominence judgments. These results to some extent confirm the observation made by Kochanski et al. (2015) that loudness is actually a more important factor in prominence perception than Fry's original studies suggest (at least in German). Moreover, the findings lend some support to the results of Sluijter et al. (1996, 1997), who find measures of spectral slope to be highly predictive of perceived prominence in another Germanic language, Dutch. In contrast to measures of amplitude, pitch and spectral slope, duration was indicated to be relatively less important

in predicting prominence ratings. The two lowest variable importances were obtained for LAST ARGUMENT and FOCUS PARTICLE. Overall, the random forest results demonstrate that prominence is indeed simultaneously cued by multiple linguistic factors. Moreover, the factors specifying pitch contour shape (as reflected in our discrete prosodic variables) appear to be most predictive of listeners' prominence judgments.

4.3. Individual differences in listening behavior

First, how consistent are untrained listeners in their prominence annotation? To measure agreement between listeners, we used Fleiss' kappa κ, which ranges from 0 (no agreement) to 1 (perfect agreement) (Fleiss, 1981). In our case, Fleiss' kappa was 0.53, which shows moderately high agreement in prominence marks, but it is also a clear demonstration of by-listener differences in prominence judgments. The presence of individual differences can formally be established by performing likelihood ratio tests of the by-listener random slope component for all the models reported above (in this case models were fitted with maximum likelihood estimation). These tests were indicated to be statistically reliable in all cases (all $\chi^2 > 4, p < 0.05$), showing that for all of the 16 variables considered here, there are statistically reliable by-listener differences.

In this analysis, we look at individual differences in the random effects estimates (compare Drager & Hay, 2012), in particular the random slopes from the models discussed in section 4.1. Each listener in our logistic mixed effects regression models is associated with a random slope estimate, quantifying the degree to which this listener changes his or her prominence judgments as a function of a specific linguistic variable. For example, listener "KRm" has an RMS AMPLITUDE slope of 1.92, compared to "PBm", who has a

slope of only 0.95. This means that for listener "KRm", increasing amplitude leads to a much bigger increase in the probability of judging a word to be prominent compared to listener "PBm", whose prominence ratings were less affected by amplitude (as indicated by the smaller slope). We can similarly extract the estimates from the other regression models and incorporate them into one big matrix where each row represents a listener and each column represents the listener's slope for a particular linguistic variable (each column was *z*-scored).

We grouped the linguistic variables according to the subdivisions discussed above, i.e., (1) continuous-valued prosodic parameters, (2) contrastive prosodic categories and (3) non-prosodic factors. We then computed the average random slope estimates for each group of variables. For example, the random slopes of the continuous prosodic variables (including RMS AMPLITUDE and MAXIMUM F0) were averaged, yielding one random slope estimate that quantifies the degree to which a listener relied on all acoustic variables together in making prominence judgments.[8] Table 4 shows the correlations between the three groups of variables in our study. As can be derived from the table, continuous prosodic variables such as MEAN F0 and RMS AMPLITUDE are correlated with the discrete prosodic variables (ACCENT TYPE, ACCENT POSITION, ACCENTED) across listeners in a statistically reliable fashion. This means that listeners who strongly based their prominence judgments on acoustic measures such as SYLLABLE DURATION or MAXIMUM F0 were also

[8] Because ACCENT POSITION and ACCENT TYPE are associated with multiple slopes (because they are categorical variables with more than two levels), a different measure was needed for the discrete prosodic factors to capture the extent to which a listener relied on those variables: We calculated the predicted prominence rates (in log odds) for each category and each listener separately. Listener "AKw", e.g., rated prenuclear accents to be prominent in 33% of the cases, compared to 49% for nuclear accents in the ip, 57% for nuclear accents in the IP and about 0% for postnuclear accents. The standard deviation across the corresponding log odd random effect estimates gives an indicator as to how much listener "AKw" changed her ratings as a function of ACCENT POSITION.

strongly influenced by the type or position of the pitch accent. Thus, it seems that variables relevant for prosody (both discrete and continuous-valued) pattern together, which provides independent evidence for the idea that the discrete prosodic variables capture important prosodic properties of the pitch curve. However, neither one of the acoustic or discrete prosodic variables is correlated in listener behavior with the non-prosodic variables, such as lexical and syntactic factors.

	Continuous prosodic	**Discrete prosodic**	**Non-prosodic**
Continuous prosodic	1.0	0.72* ($p < 0.0001$)	-0.2 ($p = 0.31$)
Discrete prosodic	0.72* ($p < 0.0001$)	1.0	-0.34 ($p = 0.076$)
Non-prosodic	-0.2 ($p = 0.31$)	-0.34 ($p = 0.076$)	1.0

Table 4: Correlation coefficients (Pearson's *r*) for correlations among sets of random slope coefficients; uncorrected *p*-values

So far, we have looked at correlations between random slope estimates. In this section, we also look at whether there are any latent listener groups, i.e., whether listeners pattern together in their task behavior. To investigate this, a cluster analysis was performed on the prominence marks. First, a distance matrix was computed from the marks (0 or 1 for each data point for each listener) using Manhattan distance, a measure for discrete-valued

distances. If two listeners gave exactly the same distance marks for all words, their distance is 0. If two listeners gave prominence marks to very different words, their pairwise distance is greater. Second, a hierarchical cluster analysis was performed on this distance matrix (using the Ward D.2 algorithm and silhouette values to determine the optimal cluster solution; see Levshina, 2015, Ch. 15), which suggested a three-cluster grouping as an appropriate solution. One of the resulting groups contained a single listener (who behaved very differently from everybody else in the study) and will not be considered here. The other two groups contained 18 and 9 listeners, respectively.

The fact that there is statistical support for at least two groups is already surprising. But what characterizes these two groups? To assess this question, we computed average random slopes for each cluster. These are shown in Figure 8. Negative values indicate that people in that group relied *less* on that specific linguistic variable when it comes to prominence judgments. Positive values indicate that people in that group relied *more* on that variable. (Since values are *z*-scored, a negative score does not mean that there was a negative relationship between prominence and that variable; it only indicates that the listener group had lower than average values.)

Figure 8: Differences in the random slope estimates between the two listener groups, revealed through the cluster analysis. The variables H1-A2 and SPECTRAL EMPHASIS were excluded because the random slopes proved to be difficult to estimate and because they are very similar to H1-A3; the interpretation of the results does not change if these variables are included.

Figure 8 clearly shows that listeners in group 1 (white, the numerically larger group) paid a lot of attention to pitch-related variables, namely ACCENT TYPE and ACCENT POSITION, both of which are discrete prosodic variables relating to intonational phonology, and to MEAN F0 and MAX F0. To a lesser degree, they attended to spectral energy cues (RMS AMPLITUDE and SPECTRAL TILT). In contrast, these listeners were less affected by the NUMBER OF SYLLABLES, WORD FREQUENCY, LAST ARGUMENT and PART-OF-SPEECH, all of which are either lexical or semantic-syntactic variables, relating to the specific word being used. They were also less affected by VOWEL DURATION and SYLLABLE DURATION (but to a smaller extent). Listeners in group 2 showed the opposite pattern, paying heightened attention to the lexical and semantic-syntactic variables – plus the prosodic cue DURATION – and lowered attention to the type and position of the pitch accent, as well as to the pitch-related variables MEAN F0 and MAX F0. This exploratory analysis thus suggests that some participants may be called "pitch-guided listeners", while others may be termed "lexical-syntactic listeners", who rely more on lexical-grammatical cues and less on prosodic factors. Interestingly, however, the groups do not neatly divide along prosodic and non-prosodic criteria, since duration is not used as an important cue for the "pitch-guided listeners". Thus, from a phonological point of view, listeners may interpret prominence either in terms of stress or metrical structure (in which case durational aspects are primary) or in terms of accentual prominence (in which case pitch is the central acoustic cue).

In fact, the difference between the groups may either be explained by a potentially genuine difference in how listeners perceive prominence (as we just tried to do) or, alternatively, by different interpretations of the task instructions. In section 3.3 we detailed

how we characterized prominence in a deliberately open fashion, which also means that different participants could have latched onto different aspects of the multi-layered definition of prominence that we provided. In particular, it is possible that some participants interpreted the task as looking for *highlighted* words (pitch-guided listeners) while others interpreted the task as looking for *important* words (lexical-syntactic listeners) (see Streefkerk, 2002). Future research needs to establish whether the listener differences we found are based on different interpretations of the task or on genuine differences in the perceptual systems of particular listeners. On the basis of our tentative results, we cannot draw any clear-cut conclusions on this issue. Rather, it should be emphasized that the present groupings are based on an analysis that is decidedly exploratory, which is furthermore based on a relatively small number of listeners. This means that more work needs to be done in order to see whether the same groupings found in the present study can be confirmed in other investigations. At a bare minimum, the present results provide more converging evidence for individual differences in response behavior with respect to prominence tasks (see also Roy et al., 2017; Shport, 2015).

5. Discussion

The present study addressed the question of which factors determine perceived prominence. Our analyses clearly show that a whole swath of linguistic variables matter for prominence perception. The investigation focused on a set of 17 different linguistic variables (not intended to be an exhaustive list), all of which were related to perceived prominence in a statistically reliable fashion. Among these variables, the discrete prosodic ones (ACCENT POSITION, ACCENT TYPE and ACCENT VS. NO ACCENT) were particularly

predictive of perceived prominence and at least in our study, non-prosodic variables, such as lexical-syntactic factors, were comparatively less important.

How are we to interpret the dominance of the three 'discrete' prosodic variables ACCENT POSITION, ACCENT TYPE and ACCENT VS. NO ACCENT in our study? It has to be emphasized again that these variables stem from an expert GToBI annotation of our stimulus sentences. They are thus linguistically informed categorisations of an utterance's intonation. The listener, of course, only perceives these categories via a continuous phonetic signal. However, to the extent that these three variables play a strong role in determining people's prominence judgments, this shows that the GToBI system captures important aspects of how prominence is perceived in German. Thus, our results provide independent vindication of the GToBI labelling system, since it shows that annotations performed under this scheme can capture listeners' prominence behaviors (see also Baumann & Röhr, 2015). The results furthermore suggest that it is possible to use non-expert judgments to get at linguistically relevant categorical information and, more specifically, that prominence marks by untrained listeners can be used as a proxy for GToBI labels.

Moreover, it has to be borne in mind that the GToBI categories we employed (particularly ACCENT TYPE) represent particular shapes of the pitch curve, for example whether a pitch curve is clearly rising onto or within an accented syllable, symbolized as L+H* in GToBI, or whether there is 'just' a pitch peak, symbolized as H*. The high importance of the ACCENT TYPE variable thus suggests that shape characteristics, even if only approximated via discrete categories, do play a particularly important role in prominence perception (Baumann & Röhr, 2015; Knight, 2008; Niebuhr, 2009).

In addition to the discrete prosodic variables, we found acoustic variables, such as DURATION, MAXIMUM F0, RMS AMPLITUDE or measures of SPECTRAL TILT, to be associated with prominence judgments in a statistically reliable fashion. From a purely confirmatory perspective, the present results replicate for German and for the precise task used (RPT) that not only pitch movement and height but also intensity and duration play a role (see Fry, 1955, 1958, 1965, and many studies since then, e.g., Kochanski et al., 2005). Moreover, we show that the distribution of energy across the spectrum matters, in particular spectral tilt (Sluijter et al., 1997). We also found that other, non-prosodic, variables are relevant for prominence perception, such as word frequency (Cole et al., 2010b) and semantic-syntactic factors such as part-of-speech, focus particles and argument position. However, our results also suggest that these variables may be less important when compared to ACCENT TYPE and ACCENT POSITION. Since GToBI labelers usually base their annotation on the pitch curve (see Grice et al., 2017), this result would seem to suggest that pitch is a more important aspect of the intonational grammar of German, even though additional prosodic and non-prosodic factors clearly also play a role.

The high importance of ACCENT POSITION confirms previous claims regarding the nuclear accent as central for the interpretation of utterances (see section 2.1). Our results show that untrained listeners pay most attention to nuclear accents, which have a high probability of attracting a prominence mark. The p-scores of the example in Figure 9a (see pitch contour in Fig.1a above) indicate that the nuclear accent on *Bachblütenkur* ('cure with Bach flowers') is perceived as most prominent – in particular since it occurs in a non-canonical position: the nuclear accent occurs early in the phrase, and not on the last argument *Bahber*, which is the default candidate for the nuclear accent in a broad focus

structure. Especially since we presented the sentence out of context, the position of the nuclear accent is highly marked. Calhoun (2010) claims that it is exactly this kind of mismatch of prosodic cues with structural expectations that attracts attention and consequently leads to the perception of prominence. Furthermore, *Bachblütenkur* is marked by the focus particle *auch* ('also'), enhancing its prominence by a non-prosodic means.

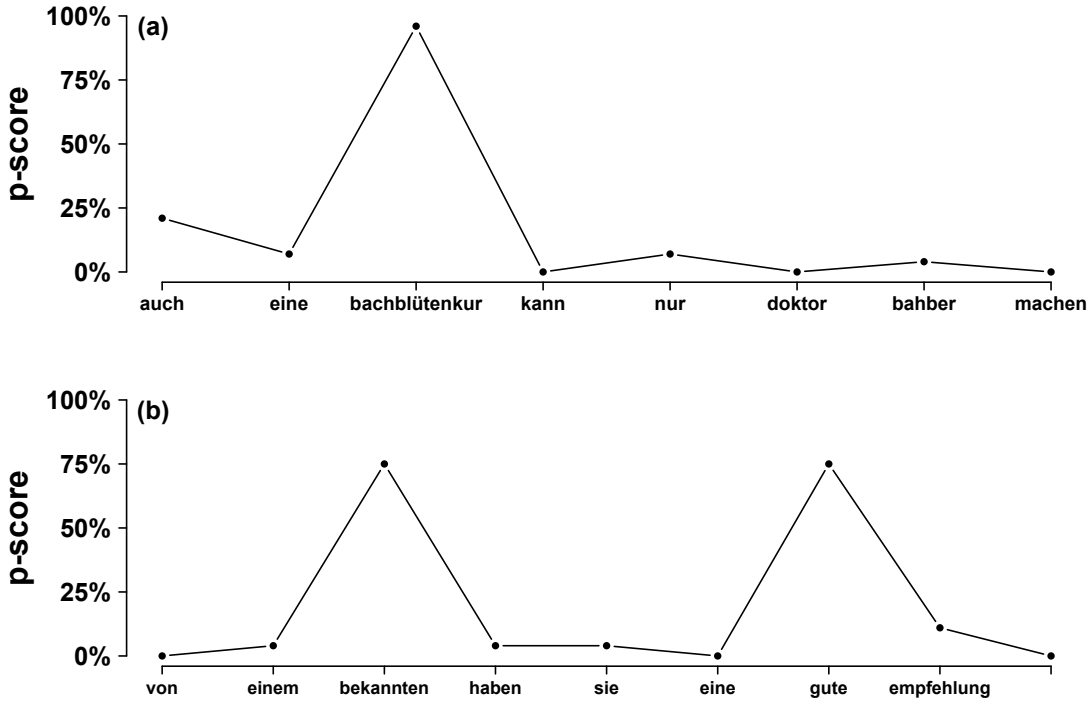


Figure 9: The utterances (a) *Auch eine Bachblütenkur kann nur Dr. Bahber machen* ('Also a cure with Bach flowers can only be done by Dr. Bahber') and (b) *Von einem Bekannten haben sie eine gute Empfehlung bekommen* ('From a friend they got a good recommendation') with their associated p-scores, corresponding to Figure 1. The first utterance has a rising nuclear accent on *Bachblütenkur* and a low phrase accent on *Dr.*

Bahber. The second utterance has a rising prenuclear accent on *gute* (‘good’) and a low nuclear accent on *Empfehlung* ('recommendation')

The high p-score on the word *Bachblütenkur* is not only due to the fact that it carries the nuclear accent but also that it is marked by a rise in pitch, which goes along with being accented. Our study once more confirms that degree and direction of a tonal movement (determining the ACCENT TYPE together with absolute pitch height) are crucial factors in prominence perception (see Baumann & Röhr, 2015). In contrast, the lack of tonal movement on phrase accents (= postnuclear prominences), such as the one on the second occurrence focus constituent *Bahber* (Figure 9a), is probably the main reason why they were only rarely judged as prominent by our listeners.

Another hint at the importance of accent types is shown in Figure 9b (see pitch contour of this example in Fig.1b above). Here, the F0 rise on *gute* (‘good’) leads to a much higher p-score than the low F0 target on *Empfehlung* (‘recommendation’) although the former accent is in prenuclear and the latter accent in nuclear position. In other words, ACCENT TYPE (rise) outweighs ACCENT POSITION (nuclear) in this utterance. Such cases are revealing because they show that words carrying nuclear accents are not necessarily the most prominent elements in an utterance, and that ACCENT POSITION is only one out of several important factors. This also serves to emphasize an important aspect of our random forest analysis: Our results are not to be interpreted in an absolute fashion, i.e., the highest ranking variable (in our case ACCENTED) will not *always* be the most important in determining prominence. However, across several different utterances, our results suggest

that those variables that we considered to be 'discrete prosodic' variables are going to be more important than other variables.

The surprisingly low p-scores for the longest words in the dataset (five words had six and eight syllables, respectively; see section 4.1.) may serve as yet another argument for the relevance of intonational aspects for prominence judgments: even though all of these words occurred as the last argument of the sentence, and the eight-syllable words always received a nuclear accent, none of them carried a prominent *type* of pitch accent. That is, the accents were either falling or high with a small pitch range.

The fact that all 17 variables tested were associated with perceived prominence suggests that prominence is indeed signaled by multiple factors simultaneously, a finding also supported by Watson (2010), Arnold et al. (2013), and Wagner et al. (2015) (see also Cole & Shattuck-Hufnagel, 2016). Of course, the different variables are correlated with each other in our stimuli. However, except for the high correlation between MEAN F0 and MAX F0 (with $r = 0.82$), most correlations between variables are relatively low. For example, Pearson's r was only 0.2 for the correlation between RMS AMPLITUDE and SPECTRAL EMPHASIS. Out of the 45 correlations between continuous variables, 39 had an r value smaller than absolute 0.5. This means that it happens quite frequently that the different prominence-related cues are not fully aligned. One possible interpretation of this result is that there are multiple notions of prominence, each with its own set of cues and with its own inherent 'prominence scale' (for a sketch of such a model, see Baumann & Cole, 2017).

From the perspective of optimal information transfer in communicative systems, it appears to be functionally relevant that there are multiple cues for the same linguistic

phenomenon, e.g., prominence marking. Winter (2014) discusses the different ways linguistic contrasts achieve robust transmission via speech (see also Mason et al., 2015) and argues that a key factor assuring robustness is "degeneracy" (what is called "functional redundancy" by Kitano 2004), a technical term used in systems science and computational biology to describe cases in which multiple "redundant" ("degenerate") system components achieve the same function. In contrast to redundancy, however, degeneracy entails that the system components are also characterized by diversity, i.e., they are not mere repetitions of the same component but *different* components achieving the same function. The present analyses reveal that prominence is characterized by many different structures, with multiple cues signaling prominent items. Crucially, these cues are linguistically diverse — some of them are syntactic or lexical, others are prosodic and discrete, yet others are prosodic and gradient. Within each of these groups of variables, we can make more fine-grained differentiations between cues: pitch, duration and intensity, for example, are all continuous-valued acoustic variables that signal prominence. Moreover, just as is the case with other domains of speech perception, some cues may be weighted more strongly in certain contexts. For example, if noise masks certain spectral cues, durational cues may still be perceivable. This highlights why it is not just important to have "redundant" cues, but also diverse ones.

The multiplicity of cues attains special significance in the light of listener variation. Our analyses of individual differences suggest that what ultimately matters for prominence perception is relative with respect to *who* is listening. Our moderate inter-rater reliability and the significance of the by-listener random slopes show that prominence perception is highly variable across individuals (see also Roy et al., 2017). From the perspective of

optimal information transfer, such listener differences could be seen as noise, as factors that potentially interfere with the perception of prominence. However, precisely because different listeners pay attention to different cues it is important that a multiplicity of cues is available: If one potential cue is not paid attention to by a particular listener, another cue can serve as a backup.

An exploratory study of the individual differences showed that they are not entirely random. We have seen two systematic patterns in an exploratory analysis of the listener behaviors. First, correlations show that discrete prosodic factors and continuous-valued prosodic factors pattern together, but they do not pattern together with non-prosodic factors such as lexical and syntactic variables. This observation was, at least in part, independently vindicated by an exploratory cluster analysis, which suggests that some listeners may pay more attention to structural, presumably more expectation-based, non-prosodic factors (semantic-syntactic, lexical) and less to prosodic factors (e.g., type of pitch accent), while others pay less attention to lexical-syntactic factors and more attention to prosodic, especially pitch-related, factors. While it is tempting to label the two groups "prosodic" listeners and "lexical-syntactic" listeners, we want to stress that at this point, larger studies with more participants need to confirm whether the listener groups that we found in this study are systematic sub-groupings of listeners that generalize to a larger population. Moreover, future research needs to establish whether the listener differences we found in this study are due to genuine perceptual differences or due to different interpretations of the task.

6. Conclusions

To sum up, the present study makes both empirical and methodological contributions. On the empirical side, the paper provides novel findings on German, demonstrating that prominence in this language is perceived in relation to multiple prosodic and non-prosodic factors, as has previously been shown for English. Further, the findings point to the pitch accent types captured by the GToBI system as a primary factor influencing listeners' prominence ratings. To the extent that these pitch accent types reflect generalizations over pitch contours, this demonstrates the importance of pitch contour as opposed to other prosodic variables (such as average pitch height, for instance). Moreover, this finding lends support for the pitch accent distinctions of the GToBI annotation system by showing that those categories determined by experts do in fact relate to the categories perceived by non-expert listeners. On the methodological side, this study extends prior work on Rapid Prosody Transcription (RPT) (Cole et al., 2010a, 2010b) by using random forests to analyze the relative contributions of prosodic and non-prosodic factors in prominence perception. A further extension is to utilize the random effects structure of mixed effects regression to analyze individual listener differences in the weighting of prosodic and non-prosodic factors in prominence ratings. Taken together, our results paint a complex picture of prominence perception that is characterized by both diversity of cues and diversity of listeners. Despite this diversity, however, prominence perception proves to be a communicatively robust system.

Acknowledgements

We thank Janina Kalbertodt for her invaluable help with the data collection and analysis as well as Simon Wehrle for creating some of the figures. Furthermore, we would like to thank Jennifer Cole, Petra Wagner and one anonymous reviewer for their extremely insightful comments on an earlier version of the paper. This work was funded by the German Research Foundation (DFG), grant BA 4734/1-2.

References

Al-Tamimi, J. (2017). Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: Implications for formal representations. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 8*, 1-40.

Arnold, D., & Wagner, P. (2008). The influence of top-down expectations on the perception of syllable prominence. *Proceedings of the ISCA Workshop on Experimental Linguistics*, 25-28.

Arnold D., Wagner P., & Baayen H. (2013). Using generalized additive models and random forests to model German prosodic prominence. *Proceedings of Interspeech 2013*, 272-276.

Arnold, D., Wagner, P., & Möbius, B. (2011). Evaluating different rating scales for obtaining judgments of syllable prominence from naïve listeners. In W.-S. Lee, & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Science* (pp. 252-255). Hong Kong: City University of Hong Kong.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica, 66* , 46-63.

Ayers, G. (1996). Nuclear accent types and prominence: Some psycholinguistic experiments. Ph.D. dissertation, The Ohio State University.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech, 47*, 31-56.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48. Version 1.1.11.

Baumann, S., & Cole, J. (2017). Accounting for context and variability in a prominence-based model of discourse meaning. Oral presentation at P&P 13, 29 September 2017, Berlin.

Baumann, S., Eckart, K., & Riester, A. (2016). Quantifying prosodic prominence for research in information structure. Poster presentation at Prosody and Information Structure in Stuttgart (PINS), 22 March 2016.

Baumann, S., Mücke, D., & Becker, J. (2010). Expression of second occurrence focus in German. *Linguistische Berichte*, *221*, 61-78.

Baumann, S., & Riester, A. (2013). Coreference, lexical givenness and prosody in German. *Lingua*, *136*, 16-37.

Baumann, S., & Röhr, C. (2015). The perceptual prominence of pitch accent types in German. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the*

18th International Congress of Phonetic Sciences (paper number 0298.1-5). Glasgow, UK: The University of Glasgow.

Beckman, M. (1986). *Stress and Non-Stress Accent*. Dordrecht: Foris.

Beckman, M., & Hirschberg J. (1994). The ToBI annotation conventions. Manuscript and accompanying speech material. Ohio State University.

Beckman, M., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S. Jun (Ed.), *Prosodic Typology - The Phonology of Intonation and Phrasing* (pp. 9-95). Oxford: Oxford University Press.

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3*, 255-309.

Bishop, J. (2012). Information structural expectations in the perception of prosodic prominence. In G. Elordieta & P. Prieto (Eds.), *Prosody and Meaning (Interface Explorations)* (pp. 239-270). Berlin: Mouton de Gruyter.

Boersma, P., & D. Weenink (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.80, retrieved from http://www.praat.org/

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Brown, L., Winter, B., Idemaru, K., & Grawunder, S. (2014). Phonetics and politeness: Perceiving Korean honorific and non-honorific speech through phonetic cues. *Journal of Pragmatics*, *66*, 45-60.

Brysbaert, M., Buchmeier, M., Conrad, M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*, 412-424.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977-990.

Büring, D. (2007). Intonation, semantics and information structure. In G. Ramchand, & C. Reiss (Eds.), *The Oxford Handbook of Linguistic Interfaces* (pp. 445-474). Oxford: Oxford University Press.

Büring, D. (2015). A theory of second occurrence focus. *Language, Cognition and Neuroscience, 30*, 73-87. DOI 10.1080/01690965.2013.835433.

Calhoun, S. (2010). The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language*, *86*, 1-42.

Cangemi, F., & Grice, M. (2016). The importance of a distributional approach to categoriality in autosegmental-metrical accounts of intonation. *Laboratory Phonology*, *7*, 1-20.

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, *30*, 1-31.

Cole, J., Mo, Y., & Baek, S. (2010a). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes*, *25*, 1141-1177.

Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010b). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, *1*, 425-452.

Cole, J., & Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology*, *7*, 1-29.

De Jong, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, *97*, 491-504.

Drager, K., & Hay, J. (2012). Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation and Change, 24*, 59-78.

Dragulescu, A. A. (2014). xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7.

El Zarka, D., Schuppler, B., Lozo, C., Eibler, W., & Wurzwallner, P. (2015). Acoustic correlates of stress and accent in Standard Austrian German. In *Phonetik in und über Österreich*. Vienna: ÖAW Austrian Academy of Sciences Press.

Eriksson, A., Thunberg, G. C., & Traunmüller, H. (2001). Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. *Proceedings of Interspeech 2001*, 399-402.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (2nd Ed.). New York: John Wiley.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America, 27*, 765-768.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech, 1*, 126-152.

Fry, D. B. (1965). The dependence of stress judgments on vowel formant structure. *Proceedings of the 5th International Congress on Phonetic Sciences*, 306-311.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). irr: Various coefficients of interrater reliability and agreement. R package version 0.84.

Grabe, E., Post, B., & Nolan, F. (2000). Modelling intonational variation in English: the IViE system. In S. Puppel, & G. Demenko (Eds.), *Prosody 2000. Speech Recognition and Synthesis* (pp. 51-57). Poznan: Adam Mickiewicz University, Faculty of Modern Languages and Literature.

Grice, M., & Baumann, S. (2016). Intonation in der Lautsprache: Tonale Analyse. In B. Primus, & U. Domahs (Eds.), *Handbuch Laut, Gebärde, Buchstabe* (pp. 84-105). Berlin: Mouton de Gruyter.

Grice, M., Baumann, S., & Benzmüller, R. (2005). German intonation in autosegmental-metrical phonology. In S. Jun (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 55-83). Oxford: Oxford University Press.

Grice, M., Baumann, S., & Jagdfeld, N. (2009). Tonal association and derived nuclear accents: The case of downstepping contours in German. *Lingua, 119*, 881-905.

Grice, M., Baumann, S., Ritter, S., & Röhr, C.T. (2017). GToBI. Übungsmaterialien zur deutschen Intonation. Available at www.gtobi.uni-koeln.de.

Grice, M., Savino, M,. Caffo, A., & Roettger, T. B. (2015) The tune drives the text - Schwa in consonant-final loanwords in Italian. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (paper number 0381.1-5). Glasgow, UK: The University of Glasgow.

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

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Halliday, M.A.K. (1967). *Intonation and Grammar in British English*. The Hague: Mouton.

Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of spectral emphasis in Swedish. *Journal of Phonetics, 31,* 39-62.

Heuft, B. (1996). *Eine prominenzbasierte Methode zur Prosodieanalyse und –synthese*. Frankfurt: Peter Lang.

Iseli, M., Shue, Y.L., & Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *Journal of the Acoustical Society of America*, *121*, 2283-95.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee and P. Hopper (Eds.). *Frequency and the Emergence of Linguistic Structure* (pp. 229-254). Amsterdam: John Benjamins.

Kakouros, S., & Räsänen, O. (2016). 3PRO - An unsupervised method for the automatic detection of sentence prominence in speech. *Speech Communication*, *82*, 67-84.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, *5*, 826-837.

Knight, R.A. (2008). The shape of nuclear falls and their effect on the perception of pitch and prominence: peaks vs. plateaux. *Language and Speech*, *51*, 223-244.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, *118*, 1038-1054.

Ladd, D.R. (2008). *Intonational Phonology* (2nd Ed.). Cambridge: Cambridge University Press.

Lam, T.Q., & Watson, D.G. (2010). Repetition is easy: Why repeated referents have reduced prominence. *Memory & Cognition*, *38*, 1137-1146.

Lancia, L., & Winter, B. (2013). The interaction between competition, learning and habituation dynamics in speech perception. *Laboratory Phonology*, *4*, 221-257.

Lea, W. A. (1977). Acoustic correlates of stress and juncture. *Studies in Stress and Accent*, *4*, 83-120.

Levshina, N. (2015). *How to do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins.

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182-195.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, *8*, 249-336.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2015). cluster: Cluster analysis basics and extensions. R package version 2.0.3.

Mason, P., Domínguez D., J. F., Winter, B., & Grignolio, A. (2015). Hidden in plain view: Degeneracy in complex systems. *BioSystems, 128*, 1-8.

Mücke, D. & Grice, M. (2014). The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation? *Journal of Phonetics*, *44*, 47-61.

Niebuhr, O. (2009). F0-based rhythm effects on the perception of local syllable prominence. *Phonetica*, *66*, 95-112.

R Core Team (2015). R: A language and environment for statistical computing. Vienna, Austria. Version 3.2.2.

Rietveld, A.C.M., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics*, *13*, 299-308.

Ritter, S., & Grice, M. (2015). The role of tonal onglides in German nuclear pitch accents. *Language and Speech*, *58*, 114-128.

Roettger, T. B. (2017). *Tonal Placement in Tashlhiyt: How an Intonation System Accommodates to Adverse Phonological Environments* (Vol. 3). Language Science Press.

Röhr, C. & Baumann, S. (2010). Prosodic marking of information status in German. *Proceedings of Speech Prosody 2010*, 100019:1-4.

Roy, J., Cole, J., & Mahrt, T. (2017). Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 8*, 1-36.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution, 1*, 103-113.

Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technischer Bericht, Universitäten Stuttgart und Tübingen. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

Schneider, P., & Wengenroth, M. (2009). The neural basis of individual holistic and spectral sound perception. *Contemporary Music Review*, *28*, 315-328.

Selkirk, E. (1984). *Phonology and Syntax. The Relation between Sound and Structure.* Cambridge, MA: MIT Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379-423 & 623-656.

Shattuck-Hufnagel, S., & Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*, 193-247.

Shport, I. (2015). Perception of acoustic cues to Tokyo Japanese pitch-accent contrasts in native Japanese and naive English listeners. *Journal of the Acoustical Society of America*, *138*, 307-318.

Sluijter, A. M. C., Shattuck-Hufnagel, S., Stevens, K., & Heuven, V. J. van (1995). Supralaryngeal resonance and glottal pulse shape as correlates of prosodic stress and accent in American English. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Conference of Phonetic Sciences* (pp. 630-633). Stockholm.

Sluijter, A. M., & Heuven, V. J. van (1995). Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. *Phonetica*, *52*, 71-89.

Sluijter, A. M., & Heuven, V. J. van (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, *100*, 2471-2485.

Sluijter, A. M., Heuven, V. J. van, & Pacilly, J. J. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, *101*, 503-513.

Streefkerk, B. (2002). *Prominence – Acoustical and Lexical/Syntactic Correlates*. Utrecht: LOT.

Streefkerk, B., Pols, L., & ten Bosch, L. (1999). Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's. *Proceedings of Eurospeech 1999*, Vol. 1, 551-554.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, *14*, 323-348.

Suzuki, R., & Shimodaira, H. (2014). pvclust: Hierarchical clustering with p-values via multiscale bootstrap resampling. R package version 1.3-2.

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, *24*, 135-178.

't Hart, J., Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation. An Experimental-Phonetic Approach to Speech Melody.* Cambridge: Cambridge University Press.

Traunmüller, H. (1997). Perception of speaker sex, age, and vocal effort. In R. Bannert, M. Heldner, K. Sullivan, & P. Wretling (Eds.), *PHONUM* 4 (pp. 183-186). Umea: Department of Phonetics.

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America*, *107*, 3438-3451.

Turco, G., Dimroth, C. & Braun, B. (2013). Intonational means to mark verum focus in German and French. *Language and Speech*, *56*, 460-490.

Turk, A. (2010). Does prosodic constituency signal relative predictability? A smooth signal redundancy hypothesis. *Laboratory Phonology, 1*, 227-262.

Turk, A., & Sawusch, J.R. (1996). The processing of duration and intensity cues to prominence. *Journal of the Acoustical Society of America, 99*, 3782-3790.

Uhmann, S. (1991). *Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie.* Tübingen: Niemeyer.

Veilleux, N., Shattuck-Hufnagel, S., & Brugos, A. (2006). *Transcribing Prosodic Structure of Spoken Utterances with ToBI.* January IAP 2006. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

Wagner, P. (2005). Great expectations - introspective vs. perceptual prominence ratings and their acoustic correlates. *Proceedings of Interspeech 2005*, 2381-2384.

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., Escudero Mancebo, D., Gili Fivela, B., Lacheret, A., Ludusan, B., Moniz, H., Chasaide, A., Niebuhr, O., Rousier-Vercruyssen, L., Simon, A.-C., Šimko, J., Tesser, F., & Vainio, M. (2015). Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (paper number 0202.1-5). Glasgow, UK: The University of Glasgow.

Wagner, P., Tamburini, F. & Windmann, A. (2012). Objective, subjective and linguistic roads to perceptual prominence. How are they compared and why? *Proceedings of Interspeech 2012*, 2386-2389.

Watson, D. G. (2010). The many roads to prominence: Understanding emphasis in conversation. *Psychology of Learning and Motivation*, *52*, 163-183.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*, 1-20.

Wickham, H., & Francois, R. (2015). dplyr: A grammar of data manipulation. R package version 0.4.2.

Widera, C., Portele, T., & Wolters, M. (1997). Prediction of word prominence. *Proceedings of Eurospeech 1997*, 999-1002.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*, 960-967.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society* (B), *73*, 3-36.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77*, 1-17.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*, 3-14.

Appendix A: Full list of stimuli (three blocks per 20 utterances)

Block 1
01 er steckt sich die banane ein
02 die beste klinik der stadt hat doktor bahber so gut ausgebildet
03 der mittelstand solle sich entscheiden ob er günstige teilzeitkräfte oder längere öffnungszeiten wolle
04 auf meinem bild hat der obdachlose das bier getrunken
05 und auch den zivildienst muss keiner mehr machen
06 sie sind schon sehr gespannt auf die ersten untersuchungsergebnisse
07 auch eine bachblütenkur kann nur doktor bahber machen
08 sie werfen die rosine weg
09 tom und isabel möchten an dem stand des frauenvereins ein bild kaufen
10 der oberarzt und seine kollegen möchten doktor bieber gerne als neuen arzt in ihrem krankenhaus einstellen
11 schon seit wochen freuen sie sich auf dieses thema
12 ich kann gar nicht glauben dass sogar doktor buhber so eins hat
13 herr müller ist der beliebteste lehrer an seiner schule
14 matthias hat mit doktor bahber geredet
15 der oberarzt und seine kollegen möchten doktor bieber gerne als neuen arzt in ihrem krankenhaus einstellen
16 schon seit wochen freuen sie sich auf dieses thema
17 sie liest sich die ballade durch
18 von einem bekannten haben sie eine gute empfehlung bekommen
19 freundlich sieht die janina aus
20 vorhin war er dafür extra noch auf dem markt beim obsthändler

Block 2
01 jetzt hat sogar doktor bahber so eins
02 sie werfen die rosine weg
03 es wird sehr schwer sie von einem günstigeren preis zu überzeugen
04 lecker sieht die banane aus
05 sie laden doktor bieber ein
06 carla muss für den deutsch-unterricht als hausaufgabe eine ballade auswendig lernen
07 sie sind schon sehr gespannt auf die ersten untersuchungsergebnisse
08 matthias hat mit doktor buhber geredet
09 tom und isabel möchten an dem stand des frauenvereins ein bild kaufen
10 es klingelt an der tür
11 sie liest sich die ballade durch
12 schon seit wochen freuen sie sich auf dieses thema
13 ich kann gar nicht glauben dass sogar doktor bahber so eins hat
14 herr müller ist der beliebteste lehrer an seiner schule
15 der oberarzt und seine kollegen möchten doktor bieber gerne als neuen arzt in ihrem krankenhaus einstellen
16 die eltern sind sich unsicher mit einem neuen medikament das sie vom arzt für ihr

kind bekommen haben
17 die mädchen werden sich mit ihrer neuen klassenkameradin sicher gut verstehen
18 von einem bekannten haben sie eine gute empfehlung bekommen
19 herr müller ist der beliebteste lehrer an seiner schule
20 eine akupunktur kann nur doktor bahber machen

Block 3
01 er schaut sich die nina an
02 sie wird den mädchen die sprache sicher sehr schnell beibringen können
03 die beste klinik der stadt hat doktor bahber so gut ausgebildet
04 sie rufen doktor bahber an
05 außerdem ist er dafür bekannt zu jedem neuen thema in der stunde einen kleinen film zu zeigen
06 sie sind schon sehr gespannt auf die ersten untersuchungsergebnisse
07 sie werfen die rosine weg
08 herr müller ist der beliebteste lehrer an seiner schule
09 die eltern sind sich unsicher mit einem neuen medikament das sie vom arzt für ihr kind bekommen haben
10 der oberarzt und seine kollegen möchten doktor bieber gerne als neuen arzt in ihrem krankenhaus einstellen
11 jetzt hat sogar doktor buhber so eins
12 schon seit wochen freuen sie sich auf dieses thema
13 tom und isabel möchten an dem stand des frauenvereins ein bild kaufen
14 freundlich sieht die janina aus
15 es klingelt an der tür
16 auch eine bachblütenkur kann nur doktor bahber machen
17 vorhin war er dafür extra noch auf dem markt beim obsthändler
18 die mädchen werden sich mit ihrer neuen klassenkameradin sicher gut verstehen
19 schon seit wochen freuen sie sich auf dieses thema
20 eine akupunktur kann nur doktor bahber machen

Appendix B: R packages used

Package	Version	Used for	Reference
`lme4`	1.1.11	mixed models	Bates, Maechler, Bolker, & Walker (2015)
`mgcv`	1.8.15	mixed models	Wood (2011)
`ranger`	0.8.0	random forests	Wright & Ziegler (2017)
`dplyr`	0.4.2	preprocessing	Wickham & Francois (2015)
`reshape2`	1.4.1	preprocessing	Wickham (2007)
`xlsx`	0.5.7	preprocessing	Dragulescu (2014)
`irr`	0.84	Fleiss' kappa	Gamer, Lemon, Fellows & Singh (2012)
`cluster`	2.0.3	silhouette values	Maechler, Rousseeuw, Struyf, Hubert & Hornik (2015)
`pvclust`	1.3.2	validating cluster solution	Suzuki & Shimodaira (2014)